Chatbots Language Design: The Influence of Language Variation on User Experience with Tourist Assistant Chatbots

ANA PAULA CHAVES, Northern Arizona University, and Federal University of Technology–Paraná
JESSE EGBERT, Department of English, Northern Arizona University
TOBY HOCKING, ECK DOERRY, and MARCO AURELIO GEROSA, School of Informatics, Computing, and Cyber Systems, Northern Arizona University

Chatbots are often designed to mimic social roles attributed to humans. However, little is known about the impact of using language that fails to conform to the associated social role. Our research draws on sociolinguistic to investigate how a chatbot's language choices can adhere to the expected social role the agent performs within a context. We seek to understand whether chatbots design should account for linguistic register. This research analyzes how register differences play a role in shaping the user's perception of the human-chatbot interaction. We produced parallel corpora of conversations in the tourism domain with similar content and varying register characteristics and evaluated users' preferences of chatbot's linguistic choices in terms of appropriateness, credibility, and user experience. Our results show that register characteristics are strong predictors of user's preferences, which points to the needs of designing chatbots with register-appropriate language to improve acceptance and users' perceptions of chatbot interactions.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Empirical studies in HCI**; *Natural language interfaces;*

Additional Key Words and Phrases: Chatbots, conversational agents, language design, register, user perceptions

ACM Reference format:
Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots Language Design: The Influence of Language Variation on User Experience with Tourist Assistant Chatbots. *ACM Trans. Comput.-Hum. Interact.* 29, 2, Article 13 (January 2022), 38 pages.
https://doi.org/10.1145/3487193

To Caitlin Abuel and Tyler Conger, NAU CS undergraduate students, for the contributions to the recruitment and qualitative data collection during the lab sessions. This work is partially supported by the National Science Foundation under Grants 1815503 and 1900903.
Authors' addresses: A. P. Chaves, Northern Arizona University, 1295 S Knoles Dr, Flagstaff, Arizona 86011, and Federal University of Technology—Paraná, R. Rosalina Maria dos Santos, 1233, Campo Mourão, PR 87301-006, Brazil; email: anachaves@utfpr.edu.br; J. Egbert, Department of English, Northern Arizona University, 705 S Beaver St, Flagstaff, Arizona 86011; email: Jesse.Egbert@nau.edu; T. Hocking, E. Doerry, and M. A. Gerosa, School of Informatics, Computing, and Cyber Systems, Northern Arizona University, 1295 S Knoles Dr, Flagstaff, Arizona 86011; emails: {Toby.Hocking, Eck.Doerry, Marco.Gerosa}@nau.edu.

1073-0516/2022/01-ART13 $15.00
https://doi.org/10.1145/3487193

1 INTRODUCTION

Recent advances in conversational technologies have promoted the increasing popularity of chatbots [63, 153], which are disembodied conversational interfaces that interact with users in natural language via a text-based messaging interface. A recent report on the chatbot market [62] attests to their increasing demand, predicting a global chatbot market of USD 1.25 billion by 2025. The skyrocketing interest in chatbot technologies has brought new challenges for the HCI field [25, 53, 63, 115] and, despite improvements in design, users often are not satisfied with their experiences [87, 103, 157], which may affect their attitudes toward the technology [86].

For chatbots, natural language conversation is the primary mechanism for achieving interactional goals. Therefore, developing a more comprehensive understanding of the linguistic design of chatbot conversations and its effects on users' perceptions is critical to the success of chatbot technologies. 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 [42, 85, 105]. 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., [80, 107, 108, 159]). Although current chatbots may, at some functional level, provide users with the answers they seek, the utterances portray arbitrary patterns of language that may not take into account the interactional situation. For instance, one would expect a chatbot representing a financial advisor to employ a different tone and linguistic features than that employed by a chatbot advising teens on current fashion choices. 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 users' perceptions, much less about how to tailor chatbot design to the particular situation of use. When exploring a list of key factors that could influence users' perceptions of chatbots, scholars even argued that using an appropriate language style is not relevant as long as the user can understand the chatbot's answer [9, 23]. In contrast, empirical studies have repeatedly demonstrated that chatbot's linguistic choices influence users' perceptions and behavior toward chatbots [5, 46, 127, 142, 146]. Using appropriate linguistic choices potentially increases human-likeness [59, 68, 79] and believability [79, 112, 113, 141], as well as enhance the overall perception of the quality of the interaction [78]. Resendez [127] showed that a chatbot's linguistic style evoke competence and trust as well as likeability and usefulness. 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 users' perceptions, and takes a first step toward developing a prescriptive basis for tailoring chatbot linguistic choices to specific interactional situations.

Humans have developed a sense of how to adapt their tone, idioms, and formulations to various conversational contexts. According to sociolinguists [16], human linguistic choices are not arbitrary but are closely tailored by speakers to convey not just the informational payload, but also a host of subtle but important social cues [82, 84]. This concept is called *register*, and it has emerged as one of the most important predictors of linguistic variation in human–human communication [13]. Register theory establishes, for example, that the patterns of language used to write this article differs from the language one would use to discuss the same topic for a lecture or while texting a friend. Although the relevance of register in human–human communication has been well-established in the sociolinguistic community, the potential applicability of these insights has not yet been explored in the context of chatbot language design.

Argamon [6] suggests that the sociolinguistic concept of register could be formalized 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, which is similar to the subconscious humans' language production process. The first step toward this goal is to understand whether register theory applies to chatbot interactions and how it might inform chatbots' designs. The cornerstone for this effort is an analysis to expose the perceived effect of chatbot language choices, while reducing the effect of variables other than language (e.g., variations in conversation context and content). This requires parallel conversations that are similar in content, but each represent a different linguistic register.

In our previous work [29, 30], we explored language variation in the context of tourism-related interactions, aiming to expose the relationships between register and the interactional situations within this domain. Our results conformed with established sociolinguistic theories: core linguistic features vary as the situational parameters vary, resulting in different language patterns. In this article, we move to the next step, analyzing the extent to which the register differences we previously identified play a role in shaping the user's perception of the human-chatbot interaction. The question guiding our investigation is: *Does the use of register-specific language influence the users' perceptions of tourist assistant chatbots?*

To answer this question, we analyzed the register of two corpora of conversations (*FLG* and *DailyDialog*) to characterize their register differences and used the outcomes to produce a parallel corpus (FLG_mod) in the tourism domain, which have similar informational content as *FLG*, but vary in language use patterns. Then, we performed a user study where we asked participants to compare answers from the two corpora and express their preferences in terms of appropriateness, credibility, and user experience (UX). Finally, we conducted a statistical analysis of our user study data using a Generalized Linear Model (GLM) [56] (binary response) to identify associations between the frequency of core linguistic features and user preferences with respect to language use.

Our results showed that there is an association between linguistic features and user's perceptions of appropriateness, credibility, and UX, and that register is a stronger predictor of this association than other variables of individual biases (participants, their social 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 users' perceptions of chatbot interactions.

2 BACKGROUND

In the following, we summarize the concept of sociolinguistic register and the rationale for applying this concept as the theoretical foundation for chatbot language design. We also discuss our choice of information search interactions in tourism contexts as a testbed for our research.

2.1 Register and Linguistic Variation

Linguists have long known that language varies according to the situational context in which it is used (see Malinowski [106]). Situationally defined varieties are commonly referred to as 'registers' within a language [125]. However, there are vastly different theories regarding how to best model register variation. In one theory, typically adopted in traditional sociolinguistics, researchers reduce the construct of register down to a simple continuum of formality that ranges from very formal to very casual (see, e.g., Joos [81], Trudgill [149]). In another prominent theoretical framework, known as Systemic Functional Linguistics (SFL), register is characterized according to three abstract parameters: field, tenor, and mode [65]. However, there is no agreement on how to opera-

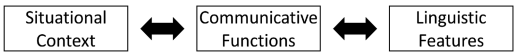


Fig. 1. Three-way relationship among situation, function, and linguistic form in the text-linguistic register framework.

tionalize these three parameters, and to date, nearly all of the research on register from Systematic Functional Linguistics scholars has been theoretical rather than empirical. As a result, Systematic Functional Linguistics has contributed little to our understanding of the influence of register on actual language use. A more recent theoretical framework for register addresses the limitations of these other theories: the text-linguistic register framework.

The text-linguistic register framework has been developed over the past 40 years by Douglas Biber and colleagues. In contrast with the simplistic idea of register as a formality scale, the text-linguistic framework addresses a comprehensive set of situational parameters that comprise the situation in which the conversation occurs, such as the participants, the channel, the production circumstances, and the purpose, among others. Moreover, the text-linguistic framework goes far beyond abstract theorizing; it is a fully developed methodology that has been applied in hundreds of empirical studies (see Biber [15] for overviews). According to the text-linguistic framework, "a register is a variety associated with a particular situation of use (including particular communicative purposes)" [15]. This methodology relies on three main components: qualitative description of the situational context, quantitative description of pervasive linguistic features, and interpretation of the relationship between situation and language in functional terms. Figure 1 displays the relationship among these three components of a text-linguistic register study (see also Biber [15], Biber and Conrad [16]).

Hence, according to the text-linguistic framework, register can be interpreted as *the distribution of the linguistic features in a situationally-defined language variety*; the linguistic features consist of the set of words or grammatical characteristics that occur in the conversation, and the situation is a set of parameters that characterize the context in which the conversation occurs, e.g., the participants, the channel, the production circumstances, among others.

Many text-linguistic studies have shown that register is crucial for linguistic research on language variation and use [13]: Most linguistic features are adopted in different ways and to varying extents across different registers. For example, some studies have focused on describing register variation in the use of a narrower set of features, such as grammatical complexity features [20, 21], lexical bundles [17, 73], and evaluation [71]. The *Longman Grammar of Spoken and Written English* [22] documents the systematic patterns of register variation for most grammatical features in English, exposing the power of register as a significant predictor of linguistic variation. Indeed, we draw on this grammar to support our discussion surrounding the use of particular linguistic features in the context of tourist assistant discourse (see Section 8).

Disregarding register in linguistic descriptions and computational language models can, and often does, result in incorrect conclusions about language use. Biber [13] offers many examples of incorrect conclusions regarding language use that have been made when disregarding registers. For instance, when a researcher states that a particular linguistic feature is more "frequent" in one text variety than in another (e.g., written vs. spoken), this conclusion intends to report that, when the feature does occur, it is preferred over the other. However, a common interpretation for that conclusion is that the feature is commonly found in that register, which is not necessarily accurate [13].

Given the crucial role register performs in human–human communication, we claim that register should be considered as a tool for chatbot language design. Although the relevance of register in human–human communication has been extensively underlined [13], 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 [9]. Still, there has been no systematic analysis of how users' 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 discuss the state of the art of chatbot language design and why register should be considered in chatbot interactions.

2.2 Chatbot Language Design

Chatbots are typically designed to mimic the social roles usually associated with a human conversational partner, for example, a buddy [144], a tutor [44, 143], healthcare provider [50, 110], a salesperson [59, 161], a hotel concierge [92], or, as in this research, a tourist assistant [30, 31]. Research on mind perception theory [67, 83, 93] 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" [93]. 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 at representing. According to [45], it is critical that the research community engages with social constructs to understand user behavior and facilitate adoption of conversational technologies.

A variety of factors influence how people perceive chatbot communication skills [32, 48, 77, 136] and, as user expectations of proficiency increase, one important way to enhance chatbot interactions is by carefully planning their use of language to mimic how humans would use language in that particular situation [60, 85].

Most linguists agree that the language choices made by humans are systematic [84], 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., [7, 96]), dialect (e.g., [90, 140]), genre (e.g., [82, 119]), and register (e.g., [11, 16]).

Style has particularly captured the attention of researchers on conversational agents [48, 78, 99, 117, 146], with explorations ranging from consistently mimicking the style of a particular character [99, 139] to dynamically matching the style to the conversational partner [69, 117, 129]. For example, Svikhnushina and Pu [137] claim that the chatbot's alignment with the user style and language is essential to shape the UX, and that style matching may ensure higher levels of personalization and engagement.

Although some scholars define style as "the meaningful deployment of language variation in a message" [48], sociolinguistics define style as a set of linguistic variants that reflect aesthetic preferences, usually associated with particular speakers or historical periods [16] (e.g., Shakespearean vs. modern English). The popularity of style relates to the common use of persona-based design, where chatbots language are planned to comply with the individual characteristics of the intended persona [61, 72]. In particular, a number of studies have sought to empirically evaluate the influence of conversational style on UXs with chatbots [5, 46]. For example, Elsholz et al. [46] 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. Cowan et al. [39] demonstrated that the conversational partner accent influence the human interlocutor lexical choices. In a study to identify migrants expectations regarding an information-search chatbots [34], participants reported that a too "casual" and "loose" way of speaking did not appropriately suit the domain and the topics. Based on an exploratory analysis, Tariverdiyeva [142] concluded that "appropriate degrees

of formality" (renamed "appropriate language style" in a subsequent work [9]) 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 [78, 89]. For example, when assessing the effects of language style on brand trust in online interactions with customers, Jakic et al. [78] concluded that the perceived language fit between the brand and the product/service category increases the quality of interaction.

Sociolinguistic studies also emphasize that the *"core linguistic features like pronouns and verbs are functional"* rather than aesthetic [16], which points to register.[1] To some extent, register and style incorporate similar kinds of linguistic analysis, which comprises understanding the use of core lexical and grammatical features [16]. However, style refers to the language variation expressed in a text **within** a single register, since style focuses on the aesthetic and personal choices while register focuses on the association between the patterns of language and the interactional context, which is determined by situationally defined parameters.

Register theory states that for each interactional situation there is a subset of norms and expectations for using language to accomplish communicative functions [16]. In a conversation, every utterance is influenced by the social atmosphere [8, 76], 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, among others [16, 82]. This results in the emergence of situationally defined language varieties, which ultimately determine the interlocutor's linguistic choices [13, 16]. In this sense, we are not suggesting to disregard style in chatbot language design; instead, we want to demonstrate that the register must encapsulate the possible variability in which the linguistic style is defined.

2.2.1 Register for Chatbot Language Design. This article aims at identifying associations between the frequency of core linguistic features (e.g., nouns, pronouns, verb tenses, adverbs[2]) and the users' preferences about a chatbot's language use. To identify the typical patterns of language present in the interactions, we adopted the text-linguistic framework, as laid out by Biber and Conrad [16] and introduced in Section 2.1. Thus, our register analysis consists of three main steps: situational description, linguistic analysis, and functional interpretation. The situational analysis aims at placing the interactions within a broad taxonomy of situational parameters, following the situational analytical framework proposed by Biber and Conrad [16]. In the following, we argue on the rationality behind choosing tourism as the interactional domain.

2.3 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 [152]. As the sector grows, the demand for timely and accurate information on destinations also increases; a recent survey [102] 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 [26, 145, 155].

[1]In some instances, linguists have used the terms 'genre' or 'style' to refer to situationally defined language varieties. However, both of these terms are more commonly used to refer to other constructs in linguistics. For example, style is typically used to refer to author- or speaker-specific patterns of language use (i.e., idiolect), as studied in the field of stylistics. The term genre is typically used to refer to conventional structures used to create complete texts, in contrast with the pervasive and functional features typically associated with register (see Biber [15]).

[2]For a glossary of the linguistic features considered in this study, see Section 2 of the supplementary materials.

However, conducting en-route travel information searches using small-screen mobile devices can be an overwhelming experience [91, 120], due to the information overload compounded by a lack of reliable mechanisms for finding accurate, trustworthy, and relevant information [91].

Radlinski and Craswell [123] 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 [2, 75, 94, 100]. Loo [102] 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 [62] 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[3] and Marina Alterra[4] virtual assistants.

Travel planning is also a domain in which perceived competence and trustworthiness is central to UX; 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 this domain exists and can serve as the seed for our 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 UX 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.

3 RESEARCH METHOD

As aforementioned, this research aimed at exploring the extent to which UX 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 UX, we compared conversations that are equivalent in content but vary in language patterns.

Finding such parallel data—natural language texts that have the same semantic content, but are expressed in different forms [116]—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 [148]). Although perhaps useful for 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, much less in the design of tourist assistants.

Our approach, 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 [46, 69, 142] (i.e., preferences associated with authors or historical period), we relied on the 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 affects user perceptions of conversational quality.

[3] https://botlist.co/bots/expedia.
[4] https://botlist.co/bots/marina-alterra.

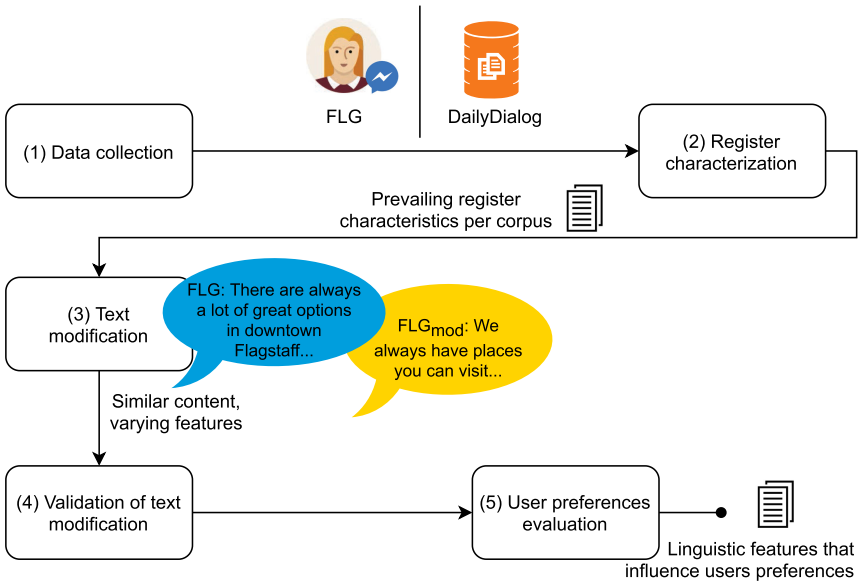


Fig. 2. Overview of the research method. The method consists of five main steps, and the outcomes of one step is seeded into the next step.

Our approach comprised five steps, as illustrated in Figure 2. We collected two corpora of conversations in the tourism domain and used them to develop parallel corpora that, while equivalent in content, differ across varying dimensions of conversational register. These conversations were then presented to users to generate a multi-faceted evaluation of subjective conversational quality. Finally, the user perceptions were analyzed with respect to the variations in register to expose a relationship between language patterns and user perceptions. These steps are further described below.

(1) **Data collection:** To provide a foundation for our analysis, we collected conversations of human domain experts (tourist assistants) interacting with tourists in a text-based tourist information search scenario; we refer this as the *FLG* corpus. Because 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* [97]. 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 Section 4.

(2) **Register characterization:** Our next aim was to characterize the conversational registers present in our two corpora. Based on a broad set of linguistic features that have been previously identified as relevant for characterizing conversational register [11], 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. [29]. The register characterization step is detailed in Section 5.

(3) **Text modification:** Having identified discrete register variations present in the two corpora, our 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}. For example, consider the sentence "There are always a lot of great options in downtown Flagstaff..." (*FLG*); the text modification for increasing the frequency of *first- and second-person pronouns*, *present tense verbs* ("visit"), and *possibility modals* ("can") as well as reducing the frequency of *amplifiers* ("great"), *nouns* ("downtown Flagstaff"), and *prepositions* ("of", "in") resulted in "We always have places you can visit" (FLG_{mod}) (see Figure 2). Section 6 details the text modification.

(4) **Validation of text modification:** 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, which is presented in Section 6.2. We invited participants to compare the parallel answers in terms of naturalness and content preservation. After performing these foundational steps, we ended up with two parallel corpora (*FLG* and FLG_{mod}).

(5) **User preferences evaluation:** After developing and validating 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 have the greatest impact on aspects of UX. 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: appropriateness, credibility, and UX. Then, we fitted a statistical learning model to identify the linguistic features that best predict the users' choices. This study is detailed in Section 7.

Having established an overview of our study method, the following sections detail each of the steps outlined above.

4 DATA COLLECTION

The data collection comprises two sub-steps: (i) to collect a baseline corpus of human–human conversations between tourist assistants and tourists; and (ii) to select a corpus of conversations in the tourism domain. In the following, we introduce the collected corpora.

4.1 FLG

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 [51], including in-state, out-of-state, and international visitors. According to the 2017–2018 Flagstaff Visitor Survey [147], 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. Tourists posed questions using a Facebook Messenger account provided by the recruiter and waited for the responses (synchronous interaction). 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 posted these questions to the tourist assistants using the same Messenger account. Questions from websites were posted in between one interaction with real tourists and the other, which also helped to keep the tourist assistant consistently busy. 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 [47] over the summer of 2018. The human tourist assistants participated in the study from our 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.

4.2 DailyDialog

The second corpus we selected is *DailyDialog* [97], 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 crawled 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 show customer-service provider interactions, e.g., hotel guest-concierge, business person-receptionist, tourist-tour guide, and so on. We chose *DailyDialog* because it contains a large set of conversations in the tourism domain and it is likely be used as a baseline model for chatbot conversations [57].

We downloaded the *DailyDialog* corpus from its website.[5] After filtering to focus on tourism-related interactions, the subset of *DailyDialog* used in this research comprises 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.

4.3 Corpora Characteristics: The Situational Parameters

We followed the situational analytical framework proposed by Biber and Conrad [16] to identify the situational parameters in which the conversations took place. The main outcome of the situational analysis is presented in Chaves et al. [30] and summarized in Table 1. In a nutshell, the situational analysis consisted of interpreting the context of the conversations to attribute value to each parameter in the situational analytical framework. For example, for the *Production* parameter in Table 1, we defined that *DailyDialog* is *planned* because conversations in *DailyDialog* were

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

Table 1. Situational Analysis

Situational parameter	DailyDialog	FLG
Participants	Customer and service providers	Tourists and tourist assistants
Relationship	Role, power, and knowledge relations vary	Service provider, tourist assistants 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)

Situational parameters are extracted from the situational analytical framework [16].

written and carefully planned to accomplish educational purposes. In contrast, *FLG* is *quasi-real-time* because, although the language was written, it was produced in an instant messaging tool, where tourist assistants had limited time to plan the answer to the tourists' questions.

According to register theory [16], differences in situational parameters result in 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*. DailyDialog includes several kinds of face-to-face interactions with service providers, for example, hotel check-in or airport pickup, while the topic and purpose in *FLG* is narrowed to local information search for a tourist destination. Regarding the channel and production, people in *FLG* produced language through an instant messenger tool in (quasi-) real time. In contrast, *DailyDialog*'s utterances are human-written, but carefully planned to mimic face-to-face conversations. Finally, *DailyDialog* portrays significant variation in the Participants and their Relationship, including the role the participants perform in the conversations, the power relationship, and knowledge (e.g., hotel concierge and tourist vs. border control officer and tourist).

Given differences in the situational parameters, we expect that the language characteristics in *FLG* differ from the language characteristics in *DailyDialog*. In the next section, we discuss our 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.

5 VARIATIONS OF CONVERSATIONAL REGISTER WITHIN TOURISM-RELATED INTERACTIONS

To characterize the varying conversational registers used in our two corpora conversations, we performed a register analysis [11]. 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 [11, 16]. 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 subsections present this analysis and its outcomes in detail.

5.1 Register Analysis

5.1.1 Procedures. Our register analysis relied on information from the Biber grammatical tagger [14] to identify the linguistic variation present in each corpus. Details about the tagger can be found elsewhere [11, 14]. In summary, given a set of texts, this tool tags and counts the linguistic features present in the text, and return the counts normalized per 10,000 words. The tagger also calculates *dimension scores* for each text, which are based on aggregations of subsets of features derived using a multidimensional analysis algorithm. The dimension scores reveal the prevailing characteristics of the register, i.e., the levels of each of the following characteristics:

- **Dimension 1–Personal Involvement:** indicates oral vs. literate opposition; high/positive scores indicate involved, interactional and generalized content, while low/negative scores indicate informational density and exact informational content.
- **Dimension 2–Narrative Flow:** distinguishes narrative from non-narrative discourses; high/positive scores indicate narrative and reconstruction of events, while low/negative scores indicate descriptive or expository discourse.
- **Dimension 3–Contextual Reference:** indicates explicit vs. situation-dependent reference opposition; high/positive scores indicate a highly explicit and elaborated discourse, where utterances use precise references to previous ones, and common ground among interlocutors is not assumed. Low/negative scores indicate exophoric, situation-dependent reference, which implies common ground among interlocutors.
- **Dimension 4–Persuasiveness:** indicates the overt expression of persuasion; high/positive scores indicate that the discourse expresses the speaker's point of view or assess advisability, while negative scores indicate the absence of opinions or arguments.
- **Dimension 5–Formality:** distinguishes abstract from non-abstract information; high/positive scores indicate a technical and formal discourse, while low/negative scores indicate a non-technical, informal discourse.

The supplementary materials contain a glossary of the linguistic features per dimension, including definitions and examples of use.

We first analyzed the dimension scores to understand the linguistic characteristics and varieties of the discourse in each corpus. Following Biber [11], 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 our 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 individual features used to calculate the dimension scores, we performed an ANOVA statistical test ($df = 3{,}1139$) where the dependent variables are the frequency of occurrences of a feature normalized per 10,000 words, whereas 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 ($\alpha = 0.05$).

5.1.2 Results. The MANOVA revealed that our three tourist assistants' dimension scores are significantly different from the average *DailyDialog* discourse ($Wilks = 0.92, F = 6.23, p < 0.0001$). Table 2 summarizes the univariate analysis per dimension.

The dimensional analysis revealed that *DailyDialog* portrays an oral discourse while tourist assistants in *FLG* are more literal 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 participants' role and power. *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 a descriptive rather than narrative language (negative estimates for Dimension 2) and slightly persuasive language (positive estimates for Dimension 4).

Table 2. Univariate Analysis of Dimension Scores ($df = 3{,}1139$)

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

For each dimension, the table shows the estimated dimension score ±, the standard error per group ($DailyDialog$, $TA1$, $TA2$, $TA3$), and the corresponding *F*- and *p*-values.

Since our ultimate goal is to reproduce the patterns of language (i.e., register) of *DailyDialog* within the conversations of the *FLG* corpus to produce our 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 prevailing linguistic features that contribute to the dimension score varied across corpora. Thus, we statistically compared the occurrences of every linguistic feature per dimension. The left side of Table 3 lists the linguistic features that vary significantly between the two original corpora (*FLG* and *DailyDialog*), as revealed by the ANOVA analysis per feature. The table includes the estimates for *DailyDialog* (control group) and each tourist assistant in *FLG* (TA1, TA2, TA3) as well as the *F*-values. A more complete table that includes the non-significant linguistic features and the *p*-values per feature is presented in the supplementary materials, which also include a glossary with examples of the features.

As indicated in Table 3, the register analysis reveals 22 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, with the typical language presented in *DailyDialog* varying significantly from that presented in *FLG* for a core set of linguistic features.

6 TEXT MODIFICATION

Having characterized the differences in register between the *FLG* and *DailyDialog* corpora, our next step was to clone the *FLG* corpus and then use the register characterization to modify its utterances, mimicking 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 [24, 101, 130]. We applied this technique to alter the linguistic features in *FLG* to approximate its language to the patterns presented in *DailyDialog*. 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. One important aspect of text modification is that both the informational content and basic linguistic integrity of the text should be preserved [24, 101, 130], otherwise, the quality of the produced sentences can be compromised. 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. The validation of text modification is presented in Section 6.2.

Table 3. ANOVA Results for Individual Features Comparison between *DailyDialog* and Both Original and Modified Corpora

Features	Estimates ± Std.Err. (FLG)					Estimates ± Std.Err. (FLG_{mod})			
	$DailyDialog$	TA1	TA2	TA3	F	$TA1_{mod}$	$TA2_{mod}$	$TA3_{mod}$	F
Dimension 1: personal involvement									
Private verb	14.7 ± 0.8	6.7 ± 3.5	5.9 ± 3.5	5.7 ± 3.3	5.6	14.5 ± 3.5	16.8 ± 3.5	14.7 ± 3.2	0.1
That-deletion	4.8 ± 0.4	0.7 ± 1.8	0.4 ± 1.8	0.6 ± 1.7	5.0	4.1 ± 1.9	4.3 ± 1.9	7.3 ± 1.8	0.8
Contraction	26.7 ± 1.1	11.7 ± 5.1	11.9 ± 5.1	1.6 ± 4.8	13.0	30.7 ± 5.1	26.5 ± 5.1	22.4 ± 4.7	0.5
Present verb	159.2 ± 1.7	125.9 ± 7.6	119.1 ± 7.7	120.8 ± 7.3	21.6	157.1 ± 7.6	153.2 ± 7.6	151.4 ± 7.0	0.6
2nd person pronoun	81.8 ± 1.52	41.4 ± 7.0	23.4 ± 7.1	44.5 ± 6.7	38.5	68.5 ± 7.1	59.4 ± 7.0	66.4 ± 6.5	5.6
1st person pronoun	55.1 ± 1.3	18.2 ± 5.9	16.8 ± 6.0	11.0 ± 5.7	40.6	37.6 ± 6.0	39.6 ± 6.0	41.9 ± 5.5	6.1
Causative subord.	0.20 ± 0.06	1.19 ± 0.29	0.00 ± 0.29	0.33 ± 0.28	3.93	0.39 ± 0.28	0.00 ± 0.28	0.32 ± 0.26	0.39
Discourse particle	6.3 ± 0.5	3.7 ± 2.3	0.4 ± 2.3	0.0 ± 2.2	4.9	4.3 ± 2.3	3.5 ± 2.3	3.5 ± 2.1	1.2
Indefinite pronoun	7.57 ± 0.53	4.79 ± 2.43	1.60 ± 2.46	2.23 ± 2.33	3.68	5.08 ± 2.47	3.46 ± 2.45	3.36 ± 2.27	2.13
Coord. conj. (clause)	6.6 ± 0.4	19.0 ± 1.9	8.8 ± 2.0	3.1 ± 1.9	15.0	6.4 ± 1.9	6.7 ± 1.9	6.2 ± 1.8	0.0
Final preposition	1.50 ± 0.21	4.9 ± 0.98	0.3 ± 1.0	2.0 ± 1.0	4.5	1.4 ± 0.9	0.3 ± 0.9	0.8 ± 0.9	0.8
Nouns	225.6 ± 2.5	277.7 ± 11.6	328.0 ± 11.8	302.1 ± 11.2	41.9	258.1 ± 11.6	280.5 ± 11.4	275.6 ± 10.6	15.5
Prepositions	70.6 ± 1.4	100.1 ± 6.6	112.1 ± 6.7	91.3 ± 6.4	20.2	80.0 ± 6.6	77.2 ± 6.5	69.9 ± 6.0	1.0
Attributive adjective	19.5 ± 0.9	43.7 ± 4.2	45.1 ± 4.2	55.1 ± 4.0	43.6	26.5 ± 4.0	23.0 ± 4.0	26.3 ± 3.7	2.1
Dimension 2: narrative flow									
3rd person pronoun	4.0 ± 0.5	12.7 ± 2.1	9.5 ± 2.1	9.8 ± 2.0	9.7	5.0 ± 2.0	5.1 ± 2.0	3.8 ± 1.8	0.2
Dimension 3: contextual reference									
WH-rel. cl. (subject)	0.2 ± 0.1	2.2 ± 0.3	0.1 ± 0.3	0.6 ± 0.3	13.2	0.8 ± 0.3	0.0 ± 0.3	0.4 ± 0.3	1.8
Nominalization	21.0 ± 0.9	27.4 ± 4.2	35.4 ± 4.2	23.3 ± 4.0	4.4	23.0 ± 4.1	26.2 ± 4.1	21.3 ± 3.8	0.6
Time adverbial	7.5 ± 0.5	3.5 ± 2.2	1.6 ± 2.3	2.0 ± 2.2	4.8	4.5 ± 2.3	4.7 ± 2.3	9.1 ± 2.1	1.2
Adverb	36.3 ± 1.2	46.8 ± 5.3	35.2 ± 5.4	23.7 ± 5.1	3.4	40.8 ± 5.3	40.9 ± 5.2	33.3 ± 4.9	0.6
Dimension 4: persuasiveness									
Prediction modal	21.8 ± 0.8	10.5 ± 3.7	8.4 ± 3.7	14.0 ± 3.6	7.9	22.9 ± 3.8	18.1 ± 3.8	19.0 ± 3.5	0.6
Suasive verb	0.6 ± 0.1	4.5 ± 0.7	1.1 ± 0.7	3.0 ± 0.6	16.2	1.1 ± 0.6	0.8 ± 0.6	0.3 ± 0.5	0.4
Conditional subord.	2.5 ± 0.3	4.0 ± 1.2	6.5 ± 1.3	5.9 ± 1.2	6.0	1.7 ± 1.2	2.9 ± 1.2	2.8 ± 1.1	0.2
Split auxiliary	1.4 ± 0.2	5.5 ± 0.9	1.5 ± 0.9	1.4 ± 0.9	6.3	1.9 ± 0.9	1.4 ± 0.9	1.9 ± 0.8	0.2
Dimension 5: formality									
Adverbial–conj.	5.2 ± 0.4	2.5 ± 1.8	1.2 ± 1.9	0.7 ± 1.8	4.0	3.1 ± 1.9	4.6 ± 1.8	4.4 ± 1.7	0.5

The left side of the table presents the estimates and standard error for each independent variable ($DailyDialog$, $TA1$, $TA2$, $TA3$) and the F-value. The right side of the table shows the estimates, standard error, and F-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 $df = 3{,}1139$. Linguistic features are grouped according to the Dimension to which they belong.

6.1 Modification Process

We manipulated the answers in *FLG* to approximate the estimate values (presented in Table 3) of a particular feature to the corresponding estimate in *DailyDialog*. For example, the estimate for *private verbs* in *DailyDialog* is 14.70 occurrences per 10,000 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.

6.1.1 Procedures. Inspired by previous studies on chatbot language use (see e.g., [46]), modifications were performed semi manually, using the AntConc tool [3] and a Python script as support tools. The Python script took in 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 then searched for all the occurrences of the first parameter in the *FLG* corpus and replaced it with the second. This process is repeated until there are no more entries in the list. The AntConc tool is used to help identify the targets to be added to the Python input list. The script then saves the modified interactions and continues to generate the complete FLG_{mod} corpus. The new texts were then submitted to the Biber tagger and subjected to renewed register analysis, as described in Section 5. If the features of interest in FLG_{mod} were still statistically different from the estimates 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*.

6.1.2 Results. The right side of Table 3 shows the comparison between *DailyDialog* and FLG_{mod}. The table shows the estimates for each tourist assistant in FLG_{mod} and the F-value

Table 4. Example of a Modified Answer

Original answer from *FLG* corpora	Modified answer (*FLG* content, *DailyDialog* form)
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.

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.

compared to the *DailyDialog* (control group). 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. 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 non-significant differences. Table 4 shows an example of a modified answer. 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}. The statistical results for the non-significant features can be found in the supplementary materials.

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.

6.2 Validation of Modifications

We performed a validation study to verify whether the modifications preserved the content and the naturalness of the original text. Details of this modification process are presented in the following subsections.

6.2.1 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 one original answer from *FLG* and the corresponding modified version from FLG_{mod}, in no particular order. 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"

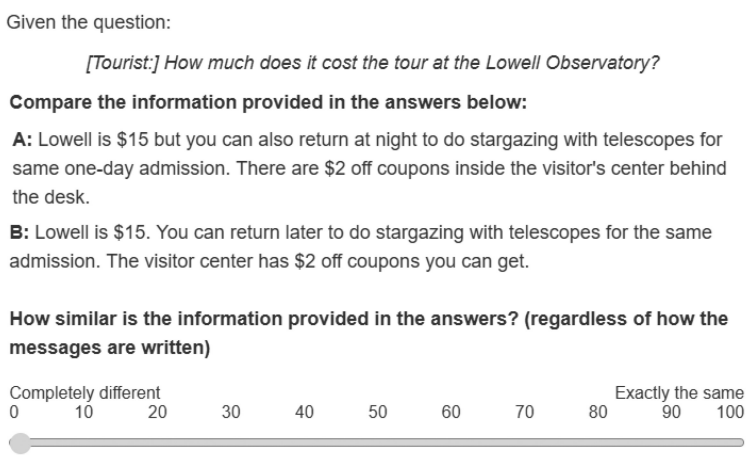


Fig. 3. Example of a content preservation question. Participants selected their answer to the question using the slider, where 0 represents that the content in the answers is completely different, while 100 represents that the content in the answers A and B is exactly the same.

(100). One example of a content preservation question is presented in Figure 3. 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. 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. The items were inspired in previous studies on conversation [111, 133, 156]. 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 [10] with the score for content similarity as the dependent variable. The random effects are the questions and the participants' 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 [35]. 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.

6.2.2 Participants. Participants were recruited through Prolific,[6] 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 [118]). 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

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

Table 5. Random Effects

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

The variance explained by the participants is larger than the residual variance, which shows that a significant portion of the variation is influenced by participants' biases.

Table 6. Number of Times the Original and Modified Answers Received a Negative, Neutral, and Positive Scores

Group	Negative (1–3)	Neutral (4)	Positive (5–7)
Original (*FLG*)	180	84	1,317
Modified (FLG_{mod})	255	105	1,210

Both original and modified answers consistently received more positive than negative scores.

Table 7. CLMM Results per Evaluated 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

The table shows the estimate, standard error, Z-values, and p-values for each item.

claimed English as their first language and were located in USA territory. Each participant received $5 (US dollars) as a compensation for their participation.

6.2.3 Results. Our 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, 89.5), which is reasonably above the upper quartile (75). Both random effects are significant, and the effect of participants has the largest variance (see Table 5). Hence, we conclude that the linguistic modifications made in producing the FLG_{mod} corpus reasonably preserved the content of answers in *FLG*.

Regarding naturalness, the number of positive ratings is consistently higher than negative ratings for both corpora (see Table 6). The CLMM models show that the scores for every item do not significantly vary, as presented in Table 7. Additionally, the participants' biases account for a lot of variance (see supplementary materials 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, our text modification process produced a parallel corpus, FLG_{mod}, with equivalent informational content as the *FLG* corpus, but with the register characteristics of the *DailtDialog*

conversations. Our validation study indicates that the modifications introduced to generate the FLG_{mod} corpus preserved the content and the naturalness expressed in *FLG*.

7 REGISTER-SPECIFIC LANGUAGE FOR CHATBOTS

In our last research step, we finally address the research question driving our effort: Does the use of register-specific language influence the users' perceptions of tourist assistant chatbots? To answer this question, we compared original and modified corpora in a study to identify which answers the participants perceive as more appropriate, credible, and providing the best UX. 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*, even though the answers sounded natural and equivalent in terms of content (as discussed in the previous subsection). Our 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.

7.1 Measurements

We compared sentences from *FLG* and FLG_{mod} under the lens of attitudinal metrics that are potentially influenced by the user's expectations and perceptions of the chatbot's language use, namely appropriateness of language, credibility, and UX.

Two important characteristics that influence the register are the identity of the participants (e.g., human vs. chatbot) and the relationship among the participants [16], which is influenced by the role they represent (e.g., tourist vs. tourist assistant). As we discussed in Section 2, users may have expectations concerning the assistant's communicative behavior and the stereotypes of its social category [78, 89]. Hence, we evaluated whether the register variation influence the *appropriateness* of the language, given the chatbot's social role. The measure of appropriateness speaks to the fit between what the user expects in that conversational context in terms of linguistic form and content and what they actually encounter in the answer.

The perceived communicative behavior might also influence how the users perceive the chatbot's *credibility*. Credibility is presented as a rating of confidence in the accuracy of the information contained in the answer; the failure to convey expertise through language compromises credibility [162] and trustworthiness [104, 138]. According to Corritore et al. [38], 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.

Finally, UX refers to the overall experience of a person using a software product, which includes their perceptions and attitudes such as emotions, beliefs, behaviors, and accomplishments [74]. Because UX is a very broad concept, it is crucial to delimit the scope of the term for this particular study. According to Peras [121], four perspectives compose the concept of UX: usability, performance, affect, and satisfaction. UX is often measured under the perspective of usability and performance [49, 109, 124], through attributes such as effectiveness, efficiency, task success, and robustness [121]. For this study, since *what* is said is similar between *FLG* and FLG_{mod}, UX should be mostly influenced by *how* it is said; hence, usability and performance should be similar across treatments. As a result, this research addresses the perspectives of affect and satisfaction, focusing mainly on the user perceptions and behavior resulting from the interaction [150]. We evaluate UX in terms of the user's general satisfaction with the chatbot language use and, consequently, the

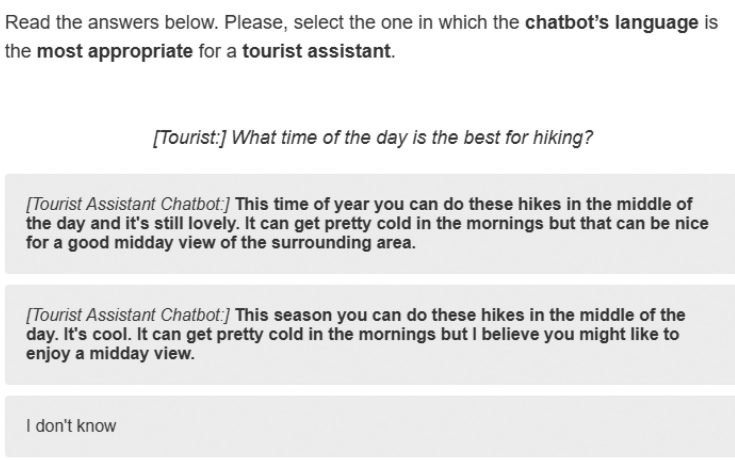


Fig. 4. Example of a question from the study. In this example, the participant was invited to select the answer that portrays the most appropriate language. Participants selected their responses by clicking on their preferred answer or on the "I don't know" option.

interaction's outcome. Supplementary materials contain the initial instruction to the constructs as presented to the participants. In the following subsections, we present the method used to evaluate user perceptions of the two parallel corpora and its outcomes.

7.2 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 our register-shifting process. To focus our comparison on answers with distinct differences in register, we selected answers that had been substantially modified: we calculated the Levenshtein distance [41, 158] between the pairs of original and modified answers, and selected the 54 question–answer pairs with the highest values of distance. The Levenshtein distance corresponds to the minimum cost (i.e., the number of insertions, deletions, or substitutions) of transforming one text into the other [41]. See the supplementary materials 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 4). Original and modified answers were presented in a randomized order, whereas "I don't know" was always the last option on the list. Supplementary materials contain samples of this instrument for the other constructs as well.

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 10 question–answer pairs to be evaluated for that construct, which were 9 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. [98]; the social orientation toward chatbots determines the participants' preferences regarding human-like social interactions with chatbots [98].

The outcome of the questionnaire consists of the users' 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 UX.

7.3 Participants

Participants were recruited through Prolific,[7] 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 US territory. Most participants had either a four-year bachelor's degree (59) or some college, but no degree (49). A total of 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, $\sigma = 11.08$).

A total of 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. The distributions of participants are detailed in the supplementary materials. Participants could provide comments and feedback on the studies through a message resource provided by the platform. Only four participants used that resource: to report a problem when loading the study page (1 participant); to report a mistake and ask to change their answers for a particular question (2 participants); and to provide positive feedback on the completed study (1 participant).

In order to augment and clarify our quantitative findings, we invited 12 participants from the Northern Arizona University community (students and staff) to perform the study in the lab and share the reasoning about their choices. At least two researchers sat with participants in the lab during these sessions and independently took notes on participants' comments. After every session, the researchers debriefed about their notes and merged the outcomes. Since the number of lab participants is limited, we did not use the qualitative data to draw conclusions regarding the participant's preferences. Instead, we used these notes only to support the discussion presented in Section 8, pointing out examples of quotes from participants that align with our quantitative results. Quotes are identified with the notation [LabPn], where "LabP" stands for "Lab Participant" and "n" is a number between 1 and 12 that represents the participant. The 12 answers to the questionnaire were not included in the statistical analysis to avoid data source bias. Each participant (for both Prolific and lab studies) received $5 (US dollars) as a compensation for their participation.

7.4 Analysis of the Linguistic Features

To model the users' preferences between original and modified versions of *FLG*, we fitted a GLM for two-class logistic regression, using the *glmnet* package in R [56]. Because we are interested in the difference between original and modified versions of *FLG* for each linguistic feature of interest (listed in Table 3), 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 10K words), and that after linguistic modification, the corresponding answer has 62.5 *private verbs*. Thus, for this particular answer, the

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

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.

7.4.1 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 [56]. The prediction function is defined as

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

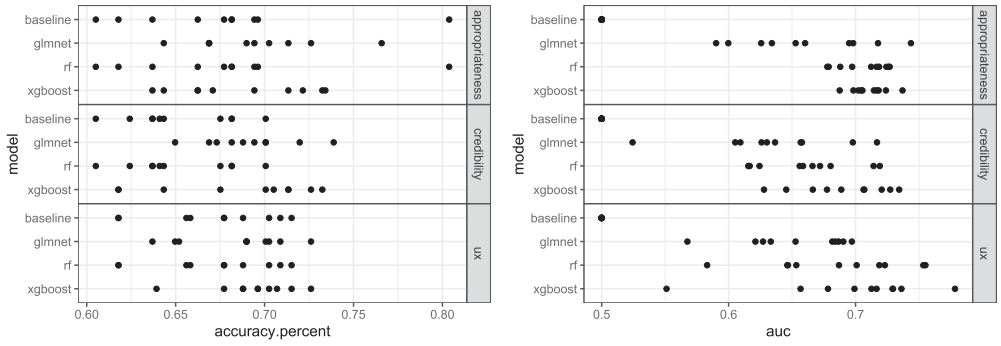
where x is a feature vector of p real numbers representing (i) the difference between original and modified counts per linguistic feature; (ii) variables representing the participant who answered the question (1 if the observation was answered by that participant, 0 otherwise), the participants' self-assessed 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 w of 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 [70, 134, 135]. For the gradient boosting, we used the `xgboost` package in R [33], 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 the supplementary material, and the R code and datasets are available on [28].

7.5 Results

Our 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 the supplementary materials.



(a) Accuracy percentage (b) AUC

Fig. 5. Accuracy (a) and AUC (b) results per model for each construct (appropriateness, credibility, and UX). 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.

Figure 5 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 plots are available in the supplementary materials), as the AUC values are consistently better than the baseline. Additionally, the `glmnet` model consistently selects the same variables to the k-folds.

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.

7.5.1 Coefficients. Table 8 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. The supplementary materials 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 UX construct.

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

Table 8. Coefficients and Standard Deviation of the Non-Zero Variables Per Construct

		Mean of coefficients ± Std. Deviation		
Linguistic features	orig. – mod.	Appropriateness	Credibility	User Experience
Dim 1: Coordinating conj. clause	(+)	−0.032 ± 0.003	−0.031 ± 0.005	−0.038 ± 0.005
Dim 1: Attributive adjective	(+)	−0.006 ± 0.001	−0.006 ± 0.001	−0.007 ± 0.002
Dim 4: Conditional subordination	(+)	−0.002 ± 0.002	0.005 ± 0.001	·
Dim 1: Nouns	(+)	0.002 ± 0.001	0.004 ± 0.001	0.003 ± 0.001
Dim 1: Prepositions	(+)	0.002 ± 0.001	0.001 ± 0.001	−0.001 ± 0.001
Dim 3: WH-relative clause (subj. pos.)	(+)	0.024 ± 0.004	0.034 ± 0.004	0.024 ± 0.004
Dim 1: Final preposition	(+)	0.035 ± 0.007	0.038 ± 0.010	0.017 ± 0.012
Dim 4: Suasive verbs	(+)	·	·	−0.015 ± 0.003
Dim 1: Causative subord.	(+)	·	·	−0.007 ± 0.005
Dim 2: Third person pronoun	(+)	·	·	0.012 ± 0.002
Dim 5: Adverbial–conjuncts	(−)	−0.029 ± 0.007	−0.017 ± 0.004	−0.056 ± 0.003
Dim 4: Prediction modals	(−)	−0.008 ± 0.001	−0.006 ± 0.002	−0.011 ± 0.001
Dim 1: Contractions	(−)	−0.002 ± 0.001	−0.008 ± 0.001	·
Dim 1: Second-person pronoun	(−)	−0.002 ± 0.001	−0.002 ± 0.001	·
Dim 1: First-person pronoun	(−)	0.007 ± 0.001	0.006 ± 0.002	0.009 ± 0.004
Dim 1: Private verbs	(−)	0.006 ± 0.001	·	·
Dim 1: Present verbs	(−)	0.003 ± 0.001	0.005 ± 0.001	0.003 ± 0.001
Dim 1: That-deletion	(−)	0.008 ± 0.003	·	0.004 ± 0.003
Dim 3: Time adverbials	(−)	0.018 ± 0.005	0.005 ± 0.003	0.007 ± 0.005
Dim 1: Indefinite pronoun	(−)	·	0.006 ± 0.004	·

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.

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 UX 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 UX, 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 UX.

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 participants' 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 UX only; and *indefinite pronoun* is relevant for credibility only.

In conclusion, our 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 UX 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.

8 DISCUSSION

The results of the study presented here show that users are sensitive to conversational register and, specifically, that register has a significant impact on user perceptions of conversational quality; a chatbot that adopts an unusual patterns of language risks losing credibility and acceptance by users. In this section, we discuss our findings and their implications for the design of register-specific language engines for chatbots.

User perceptions of conversational quality are important in chatbot design because the central interactional goal of chatbots is 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 human likeness has little relevance when compared to the chatbot's ability to handle the conversation [54]. In this sense, Balaji [9] argues that appropriate language style is not relevant for determining user satisfaction as long as the user can understand the chatbot's answer, advising only that the chatbot's language style should be "mildly appropriate to the service the chatbot provides". We suggest, however, that the narrow focus on traditional usability metrics, such as effectiveness and efficiency, fail to adequately capture the broader context of UX, which goes beyond simply comprehending a chatbot's utterance to, more importantly, whether the user trusts and ultimately uses the chatbot's advice. Specifically, UX also involves the "user's perceptions and responses that result from the use of the system" [74], including emotions, beliefs, preferences, and perceptions, among others. The results presented here suggest that these crucial user perceptions are specifically shaped by *how* information is conveyed–as characterized by the conversational register–rather than by the explicit information content of conversational exchanges. In this sense, our results support earlier behavioral observations that found that language fit can impact the quality of the interactions as well as the users' perceptions and behaviors toward the chatbots [78]. Importantly, the work presented here refines these observations by adapting register theory to 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 factors critical for the UX with chatbots, like appropriateness and credibility. Using register analysis to characterize different situations and how they map to the most appropriate register profiles also 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. As a start in this direction, in

the following we summarize some of the key insights regarding the complex relationship between conversational register and UX.

8.1 Key Findings

In this section, we begin with several broad observations drawn from our findings before moving on to discuss specific linguistic features that have the greatest impact on user perceptions.

8.1.1 Chatbot Language Design should Account for Register. Register is an established theory in the sociolinguistics domain (see e.g., [6, 11, 22]), and has been shown as a reliable predictor of language variation across conversational contexts [13]. We aimed at exploring the applicability of these results to human-chatbot interactions, develop a strong rationale for accounting for register in chatbot design, and provide a concrete mechanism for implementing theory into design practice. Our results show that, in terms of perceived appropriateness, credibility, and UX, register characteristics are more relevant than individual preferences or personal habits, i.e., none of the covariates representing individual biases (participants, their social orientation, and answers' authors) were identified as predictors of participants' choices (see Section 7.5.1). Thus, register characteristics can be seen as primary drivers for these perceptions, and designers should certainly consider register compliance to ensure chatbot acceptance and success.

8.1.2 Certain Linguistic Features Are Preferred when Efficiency Matters. Like other information search interactions, the interactions in our corpus 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 [145, 151, 155]. Efficiency is also a priority in other task-oriented domains where chatbots operate, such as customer services [25]. Our 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 our 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 [22], 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 [11]. 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 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" [22], 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 [22]. 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) [22]. 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 [11], 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" [101], since conversations favor the omission of unnecessary words to accommodate real-time production constraints. Hence, user preferences for *that-deletion* in our data could be associated with the preference for efficiency in conversations.

8.1.3 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 one hand, scholars grounded in media equation theory [52, 114] have shown that people prefer agents who reflect human social and conversational protocols, e.g., conform to gender stereotypes associated with tasks [55]; self-disclose and show reciprocity when recommending [95]; demonstrate a positive mood [66], and so on. On the other hand, overly humanized agents can create inaccurate expectations in users [59] and result in the "uncanny effect" [4, 36], which eventually leads to more frustration when the chatbot fails to live up to these increased expectations [59]. 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. This study brought to light a crucial aspect of human-chatbot interaction, namely the need for balancing the chatbot's anthropomorphic clues–several specific findings in our 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 [126]. According to Reiter and Dale [126], 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. Our 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 [22]. 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, our evidence 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 users' 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].

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. As these findings reveal, chatbot language should conform to the expectations of its social category, as previous literature has suggested [59]. Still, the register of chatbot conversation must also consider its artificial nature, particularly positioning the agent as a representative of a broader entity. As register theory suggests [16], 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.

8.1.4 Certain Linguistic Features Impact the Perceived Level of Personalization. Previous literature has shown the benefits of personalized interactions with chatbots [43, 132, 144], particularly in domains where the chatbot needs to build rapport and trustful relationships with the user, such as in financial services [42], companion (or buddy) chatbots [132, 144], and recommendation systems [27]. At the same time, users might feel uncomfortable with some aspects of personalized content [144]. In our 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, our 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 [22]. 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 [22]. 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 [22], 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. [29]) rather than information. Recommendations are more inherently personalized than information search results, and consequently require more personalization in their expression (see e.g., [58, 88, 128]). 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.

8.1.5 Avenues for Future Investigation. The study we presented in this article accomplished its primary goal and, at the same time, 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 sessions' participants. 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 UX). 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" [LabP2]
> "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.

Our results also showed a positive association between *attributive adjectives* and all the three evaluated metrics (appropriateness, credibility, and UX). 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 [22], 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 [22]. 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." Our

analysis showed that *suasive verbs* influence the UX, but were not shown to make the language more credible or appropriate. Closer analysis shows that this association, too, deserves further investigation. In our data, *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.

8.1.6 Applicability of Results to Other Domains. The register theory aims at linking the appearance of certain linguistic features in utterances to the situational parameters of the conversation [16]. 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 services 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 [1] 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 our analysis as predictors of appropriateness, credibility, and UX (see Section 7.5.1). In contrast, we expect that a sales chatbot would require more persuasion (features in Dimension 4, see supplementary materials for the full list of linguistic features), and recommendation-based chatbots would require more personalization, as we discussed previously in this section. In a proof-of-concept conversational interface for the purpose of sexual harassment prevention training [40], participants reported that using a *first-person* narrative made the vignette feel realistic and relatable, which diverges from our findings for the information-search domain. In any case, a register analysis, as presented in this article, could be used as a tool to analyze the conversation register used by expert humans in such conversational scenarios. For example, [122] claims that some domains require varying levels of language precision, and there is a need to determine what people and businesses consider the adequate language in each context. The register analysis could guide the identification of specific linguistic features of the target register that are relevant to producing the desired impact on user perceptions.

8.2 Implications for Chatbot Design

This research 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 not uncommon to find chatbots that extract and present information directly from websites, books, or manuals without any adaptation. For example, Golem[8] is a chatbot designed to guide tourists through Prague (Czech Republic); its utterances are extracted from an online travel magazine[9] without any adaptation to the new interactional situation (which differs in production, channel, and setting). Moreover, new generations of chatbots will be expected to dynamically generate their own custom-constructed utterances. They will need sophisticated language engines that are able to dynamically adapt their conversational register to changing situational parameters. In this context, corpora such as *DailyDialog* are likely to become a baseline for training the conversation models [57] at heart of such language engines; our study

[8]Available in Facebook Messenger at http://m.me/praguevisitor. Last accessed: June, 2020.
[9]https://www.praguevisitor.eu.

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 article provides a list of linguistic features that conform with user expectations about chatbots' 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. Section 8 discusses the relation of these outcomes with other task-oriented domains, such as information searches and customer services. Additionally, the methodology presented in this article 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. We also show the importance of using questions from real users to build corpora of conversations and to design valuable chatbots. As our study shows, the analysis does not necessarily require a large number of individuals to identify the linguistic features that characterize the register of a particular situation or domain.

Finally, the parallel corpora *FLG* and FLG_{mod} are available for researchers and practitioners who are interested in (i) developing chatbots for tourism information searches, as they comprise a set of frequently 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 [28].

9 LIMITATIONS AND FUTURE WORK

Register characterization relies on the multidimensional approach proposed by Biber [11, 12], which is the main theoretically motivated approach taken within register analysis [6]. Other approaches, such as register classification [6], could be explored in the context of human-chatbot interactions. Additionally, the register analysis also relies on Biber's grammatical tagger 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., [11, 12, 37]), The Longman Grammar of Spoken and Written English [22], and a study of register variation on the Internet [18, 19]. Although this tagger achieves accuracy levels comparable to other existing taggers [11], mis-taggings are possible. To mitigate this effect, we manually inspected a small subset of tagged files to search for mis-tags that could potentially impact the outcomes.

We performed manual text modifications to produce the FLG_{mod}, which inherently introduced a subjective element in the exact choice of changes applied to shift the register. We mitigated this threat by manually inspecting *DailyDialog* corpora for every feature we modified, to understand the function of the feature in the corpus, and produce modifications using similar patterns. We also performed a validation with human participants for content preservation and quality of modifications, as presented in Section 6.2.

We included in the cross-validation model only features that vary between *FLG* and *DailyDialog*, 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 users' preferences. We claim that *DailyDialog* is an appropriate dataset to be used in this study since it has been widely used in natural language and dialogue generation research (see, e.g., [64, 131, 160]), and might eventually become a baseline for learning conversation models [57]. However, it is also important to compare *FLG* against corpora produced in other interactional situations to evaluate varying sets of features and confirm the inferences presented in this article. Moreover, future research should consider applying the

register characteristics we found in this article to utterances from other tourist assistant chatbots currently available (i.e., chatbots that operate under similar situational parameters, such as Golem) to confirm our findings in other contexts.

UX is a complex concept, being potentially influenced by usability, performance, affect, and satisfaction [121]. Our study focuses on the user's expectations about the chatbot's language and the social relations suggested by the role the chatbot represents. Thus, we defined positive UX as the sentences that suggested the most satisfactory interaction, resulting in the best experience. We acknowledge, however, that other variables may influence the UX such as personal experiences with chatbots or tourist assistants, propensity to anthropomorphization, among others. It is necessary to define UX models that encapsulate these variables and establish metrics that goes beyond the assessment of usability and user satisfaction to better understand the nuances of UXs with chatbots.

The register analysis presented in this article was based on counts of the occurrences of features, normalized per 10,000 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. Future research is needed to expand the analysis to incorporate these aspects.

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 only. To increase the diversity of tourists' questions, we mined questions from websites, as discussed in Section 4. Concerning the tourist assistants, in our previous study [29], we highlighted that the tourist assistants, although experienced in their field, were not experienced in online interactions and ended up adopting different strategies to answer the tourists. This resulted in a stylistic variation between tourist assistants within the register. Additionally, subtle differences in the communicative purpose (mainly information vs. recommendation requests) also resulted in variability in the tourist assistant's utterances. Thus, an interesting extension of this study would include hiring tourist assistants with a more diverse profile and including questions about other touristic cities. Also, the influence of sub-registers must be further investigated in order to account for subtle changes in the situational parameters. It is important to note that, even with only three tourist assistants, we were able to identify the impact of certain linguistic features on the three metrics (appropriateness, credibility, and UX) used to represent perceived conversational quality; this suggests that a very large corpus is not necessary to identify the core linguistic features of chatbot dialogues that influence user perceptions. Moreover, ground-breaking work has addressed the issue of catching the nuances of user intents [154]. These advances can potentially contribute to the development of robust algorithms for communicative purpose detection in the near future.

Finally, with limited qualitative observations from our in-lab sessions to support the quantitative findings, our ability to draw conclusions based on participants' statements is incomplete. The purpose of adding this qualitative element was to augment and clarify our quantitative findings, by identifying participants' impressions aligned with our quantitative analysis and comparing the impressions of our participants to the interpretations of linguistic features we find in the register literature, such as in the Longman Grammar [22]. A deeper qualitative investigation is needed to draw stronger conclusions about motivations behind participants' choices.

10 CONCLUSIONS

This article focuses on investigating the impact of chatbot language use on user perceptions of language appropriateness, credibility, and UX. We collected two corpora of conversations between tourists and tourist assistants in different interactional situations and compared the language variation in these corpora by adapting techniques associated with register analysis, which are well-established by sociolinguists. Based on this analysis, we produced two parallel corpora of conversational exchanges that were equivalent in informational content but differed in linguistic form. We then used the parallel corpora to perform a study of the impact of register on user preferences, asking participants to rate parallel responses draw from the corpora on three metrics of perceived conversational quality: appropriateness, credibility, and UX. Participants were told that a tourist assistant chatbot had generated all responses.

Our results revealed statistically relevant associations between certain linguistic features present in utterances within the two corpora. Additionally, the results also showed that the linguistic features are a stronger predictor of this association than the variables representing individual biases (participants, their social orientation, and answers' authors). This outcome strongly suggests that attention to an appropriate conversational register is an important factor for the perceived quality of chatbot conversations and, therefore, critical to future chatbots' success. Although our study focused on the tourism domain, we expect these outcomes to be applicable to other interactions that share similar situational parameters (e.g., different information search scenarios). More generally, this study demonstrates that the theoretical foundation of register analysis introduced in this article 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 conversational quality. This makes register a promising cornerstone for rationalizing the design of future generations of chatbots, by reproducing this study for other domains, as well as developing empirical studies that extend this study to other aspects of register theory, such as the impact of sentence structure. Specific future directions for our research will focus on evaluating the effect of register-specific language within dynamically generated interactions with chatbots on user perceptions of conversational quality.

REFERENCES

[1] B. Abu Shawar and E. Atwell. 2004. Evaluation of chatbot information system. In *Proceedings of the 8th Maghrebian Conference on Software Engineering and Artificial Intelligence*. Centre de Publication Universitaire, Tunis, 12.

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

[3] Laurence Anthony. 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of the International Professional Communication Conference, 2005.* IEEE, 729–737.

[4] Markus Appel, David Izydorczyk, Silvana Weber, Martina Mara, and Tanja Lischetzke. 2020. The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior* 102 (2020), 274–286.

[5] Theo Araujo. 2018. 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 (2018), 183–189.

[6] Shlomo Argamon. 2019. Register in computational language research. *Register Studies* 1, 1 (2019), 100–135.

[7] Shlomo Argamon, Moshe Koppel, and Galit Avneri. 1998. Routing documents according to style. In *Proceedings of the 1st International Workshop on Innovative Information Systems*. Citeseer, 85–92.

[8] Mikhail Mikhaĭlovich Bakhtin. 2010. *Speech Genres and Other Late Essays*. University of Texas Press, Austin, TX.

[9] Divyaa Balaji. 2019. *Assessing User Satisfaction with Information Chatbots: A Preliminary Investigation.* Master's thesis. University of Twente. Retrieved from https://essay.utwente.nl/79785/1/Balaji_MA_BMS.pdf.

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

[11] Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

[12] Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge University Press, New York, NY.

[13] Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8, 1 (2012), 9–37.
[14] Douglas Biber. 2017. MAT–Multidimensional Analysis Tagger. Retrieved 27 August, 2021 from https://goo.gl/u7h9gb.
[15] Douglas Biber. 2019. Text-linguistic approaches to register variation. *Register Studies* 1, 1 (2019), 42–75.
[16] Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style* (2nd ed.). Cambridge University Press, New York, NY.
[17] Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25, 3 (2004), 371–405.
[18] Douglas Biber and Jesse Egbert. 2016. Using multi-dimensional analysis to study register variation on the searchable web. *Corpus Linguistics Research* 2, 21 (2016), 1–23.
[19] D. Biber and J. Egbert. 2018. *Register Variation Online*. Cambridge University Academic Press, Cambridge.
[20] Douglas Biber and Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9, 1 (2010), 2–20.
[21] Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *Tesol Quarterly* 45, 1 (2011), 5–35.
[22] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman Grammar of Spoken and Written English*. Vol. 2, Pearson Longman, London.
[23] Nina Böcker. 2019. *Usability of Information-Retrieval Chatbots and the Effects of Avatars on Trust*. B.S. thesis. University of Twente.
[24] Susan Bosher and Melissa Bowles. 2008. 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 (2008), 174.
[25] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: Changing user needs and motivations. *Interactions* 25, 5 (2018), 38–43.
[26] Dimitrios Buhalis and Soo Hyun Jun. 2011. E-tourism. In *Contemporary Tourism Reviews*. Chris Cooper (Ed.), Goodfellow Publishers Limited, Woodeaton, Oxford, 1–38.
[27] Jhonny Cerezo, Juraj Kubelka, Romain Robbes, and Alexandre Bergel. 2019. Building an expert recommender chatbot. In *Proceedings of the 1st International Workshop on Bots in Software Engineering*. IEEE, New York, NY, 59–63.
[28] Ana Paula Chaves. 2020. GitHub Repository. Retrieved 27 August, 2021 from https://github.com/chavesana/chatbots-register.
[29] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 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* . ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/3349537.3351901
[30] Ana Paula Chaves, Jesse Egbert, and Marco Aurelio Gerosa. 2019. Chatting like a robot: The relationship between linguistic choices and users' experiences. In *Proceedings of the ACM CHI 2019 Workshop on Conversational Agents: Acting on the Wave of Research and Development*. Retrieved from https://convagents.org/.
[31] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or multiple conversational agents? An interactional coherence comparison. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 191:1–191:13.
[32] Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 0, 0 (2020), 1–30. DOI: https://doi.org/10.1080/10447318.2020.1841438
[33] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* . ACM, New York, NY, 785–794. DOI: https://doi.org/10.1145/2939672.2939785
[34] Zhifa Chen, Yichen Lu, Mika P. Nieminen, and Andrés Lucero. 2020. Creating a chatbot for and with migrants: Chatbot personality drives co-design activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, New York, NY, 219–230.
[35] R. H. B. Christensen. 2019. ordinal—Regression models for ordinal data. R package version 2019.12-10. Retrieved 27 August, 2021 from https://CRAN.R-project.org/package=ordinal.
[36] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2018. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2018), 539–548.
[37] Susan Conrad and Douglas Biber. 2014. *Multi-Dimensional Studies of Register Variation in English*. Routledge, New York, NY.
[38] Cynthia L. Corritore, Robert P. Marble, Susan Wiedenbeck, Beverly Kracher, and Ashwin Chandran. 2005. Measuring online trust of websites: Credibility, perceived ease of use, and risk. In *Proceedings of the Americas Conference on Information Systems*. Association for Information Systems, Atlanta, GA, 370.

[39] Benjamin R. Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P. Branigan, João Cabral, and Leigh Clark. 2019. What's in an accent? The impact of accented synthetic speech on lexical choice in human-machine dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM, New York, NY, 1–8.

[40] Hyo Jin Do, Seon Hye Yang, Boo-Gyoung Choi, Wayne T. Fu, and Brian P. Bailey. 2021. Do you have time for a quick chat? Designing a conversational interface for sexual harassment prevention training. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 542–552.

[41] AnHai Doan, Alon Halevy, and Zachary Ives. 2012. String matching. In *Principles of Data Integration*. AnHai Doan, Alon Halevy, and Zachary Ives (Eds.), Morgan Kaufmann, Boston, Chapter 4, 95–119. DOI: https://doi.org/10.1016/B978-0-12-416044-6.00004-1

[42] Daniëlle Duijst. 2017. *Can We Improve the User Experience of Chatbots with Personalisation*. Master's thesis. University of Amsterdam.

[43] Willem Duijvelshoff. 2017. Use-cases and ethics of chatbots on Plek: A social intranet for organizations. In *Proceedings of the Workshop on Chatbots and Artificial Intelligence*.

[44] Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. 2013. Towards academically productive talk supported by conversational agents. In *Intelligent Tutoring Systems*. Stefano A. Cerri, William J. Clancey, Giorgos Papadourakis, and Kitty Panourgia (Eds.), Springer, Berlin, Heidelberg, 531–540.

[45] Justin Edwards and Elaheh Sanoubari. 2019. A need for trust in conversational interface research. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM, New York, NY, 1–3.

[46] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. 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*. ACM, New York, NY, 301–305.

[47] Facebook. 2018. Tech for Tourism-Rresearch Page. Retrieved 27 August, 2021 from https://www.facebook.com/VisitFlagstaff/.

[48] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[49] Kraig Finstad. 2010. The usability metric for user experience. *Interacting with Computers* 22, 5 (2010), 323–327.

[50] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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 (2017), e7785.

[51] Flagstaff. 2019. The City of Flagstaff–Arizona. Retrieved 27 August, 2021 from http://www.flagstaff.az.gov/2/Community-Profile.

[52] B. J. Fogg. 2003. Computers as persuasive social actors. In *Persuasive Technology*. B. J. Fogg (Ed.), Morgan Kaufmann, San Francisco, Chapter 5, 89–120.

[53] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *Interactions* 24, 4 (2017), 38–42.

[54] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: User experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM, New York, NY, 1–9.

[55] Jodi Forlizzi, John Zimmerman, Vince Mancuso, and Sonya Kwak. 2007. How interface agents affect interaction between humans and computers. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*. ACM, New York, NY, 209–221.

[56] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, 1 (2010), 1–22. DOI: https://doi.org/10.18637/jss.v033.i01

[57] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. 2019. On a chatbot providing virtual dialogues. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* . INCOMA Ltd., 382–387.

[58] Damianos Gavalas and Michael Kenteris. 2011. A web-based pervasive recommendation system for mobile tourist guides. *Personal and Ubiquitous Computing* 15, 7 (2011), 759–770.

[59] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems 2017*. Association for Information Systems, 13.

[60] Eun Go and S. Shyam Sundar. 2019. Humanizing Chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (Aug. 2019), 304–316.

[61] Google Developers. 2021. Conversation Design by Google. Retrieved March 3, 2021 from https://developers.google.com/assistant/conversation-design/what-is-conversation-design.

[62] Grand View Research. 2017. Chatbot market size to reach $1.25 Billion by 2025. CAGR: 24.3%. Retrieved 27 August, 2021 from https://www.grandviewresearch.com/press-release/global-chatbot-market.

[63] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–11.

[64] Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. In *International Conference on Learning Representations*. OpenReview. https://openreview.net/forum?id=BkgBvsC9FQ.
[65] Michael A. K. Halliday and Ruqaiya Hasan. 1992. *Cohesion in English*. Penguin, London.
[66] Yugo Hayashi. 2016. 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*. The Cognitive Science Society, 818–823.
[67] E. S. Heyselaar and T. Bosse. 2019. Using theory of mind to assess users' sense of agency in social chatbots. In *Proceedings of the Conversations 2019: 3rd International Workshop on Chatbot Research*. Springer, Cham, 1–13.
[68] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49, C (2015), 245–250.
[69] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* . ACM, New York, NY, 111–118. DOI: https://doi.org/10.1145/3308532.3329473
[70] Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7, 3 (2006), 355–373.
[71] Susan Hunston. 2010. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. Routledge, New York, NY.
[72] Shinhee Hwang, Beomjun Kim, and Keeheon Lee. 2019. A data-driven design framework for customer service chatbot. In *Proceedings of the International Conference on Human–Computer Interaction*. Springer, Cham, 222–236.
[73] Ken Hyland. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32, 1 (2012), 150–169.
[74] ISO 9241-11. 2018. *Ergonomics of Human-System Interaction: Part 11: Usability: Definitions and Concepts*. International Organization for Standardization, ISO.
[75] S. Ivanov and C. Webster. 2017. Adoption of robots, artificial intelligence and service automation by travel, tourism and hospitality companies–a cost-benefit analysis. In *Proceedings of the International Scientific Conference "Contemporary Tourism–Traditions and Innovations*. Sofia University, 19–21.
[76] Muayyad Jabri, Allyson D. Adrian, and David Boje. 2008. Reconsidering the role of conversations in change communication: A contribution based on Bakhtin. *Journal of Organizational Change Management* 21, 6 (2008), 667–685.
[77] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, New York, NY, 895–906.
[78] Ana Jakic, Maximilian Oskar Wagner, and Anton Meyer. 2017. The impact of language style accommodation during social media interactions on brand trust. *Journal of Service Management* 28, 3 (2017), 418–441.
[79] Marie-Claire Jenkins, Richard Churchill, Stephen Cox, and Dan Smith. 2007. Analysis of user interaction with service oriented chatbot systems. In *Human–Computer Interaction. HCI Intelligent Multimodal Interaction Environments*. Julie A. Jacko (Ed.), Springer, Berlin, 76–83.
[80] Ridong Jiang and Rafael E. Banchs. 2017. Towards improving the performance of chat oriented dialogue system. In *Proceedings of the 2017 International Conference on Asian Language Processing*. IEEE, New York, NY, 23–26.
[81] Martin Joos. 1967. *The Five Clocks*. Vol. 58, Harcourt, Brace & World, New York, NY.
[82] George Kamberelis. 1995. Genre as institutionally informed social practice. *Journal of Contemporary Legal Issues* 6 (1995), 115.
[83] Merel Keijsers and Christoph Bartneck. 2018. Mindless Robots get bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, 205–214.
[84] Adam Kilgarriff. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1, 2 (2005), 263–276.
[85] Jurek Kirakowski, Anthony Yiu, and P. O'Donnell. 2009. Establishing the hallmarks of a convincing chatbot-human dialogue. In *Human-Computer Interaction*, Inaki Maurtua. InTech, London.
[86] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 45–54.
[87] Takanori Komatsu, Rie Kurosawa, and Seiji Yamada. 2012. How does the difference between users' expectations and perceptions about a robotic agent affect their behavior? *International Journal of Social Robotics* 4, 2 (2012), 109–116.
[88] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly* 30, 4 (2006), 941–960.
[89] Robert M. Krauss and Chi-Yue Chiu. 1998. Language and social behavior. In *The Handbook of Social Psychology*. D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), McGraw-Hill, New York, NY, 41–88.

[90] William Labov, Sharon Ash, and Charles Boberg. 2005. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Walter de Gruyter, Boston, MA.
[91] Tania C. Lang. 2000. The effect of the Internet on travel consumer purchasing behaviour and implications for travel agencies. *Journal of Vacation Marketing* 6, 4 (2000), 368–385.
[92] Mirosława Lasek and Szymon Jessa. 2013. Chatbots for customer service on Hotels' websites. *Information Systems in Management* 2, 2 (2013), 146–158.
[93] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on your virtual mind?: Mind perception in human-agent negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* . ACM, New York, NY, 38–45. DOI: https://doi.org/10.1145/3308532.3329465
[94] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: How do people talk with a robot? In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, 31–40.
[95] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103, C (2017), 95–105.
[96] Geoffrey N. Leech and Mick Short. 2007. *Style in Fiction: A Linguistic Introduction to English Fictional Prose.*. Pearson Education, London.
[97] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 986–995.
[98] Vera Q. Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. 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* . ACM, New York, NY, 264–275.
[99] Grace I. Lin and Marilyn A. Walker. 2017. Stylistic variation in television dialogue for natural language generation. In *Proceedings of the Workshop on Stylistic Variation*. Association for Computational Linguistics, 85–93.
[100] Greg Linden, Steve Hanks, and Neal Lesh. 1997. Interactive assessment of user preference models: The automated travel assistant. In *User Modeling*. A. Jameson, C. Paris, and C. Tasso (Eds.), Springer, Vienna, 67–78.
[101] Michael H. Long and Steven Ross. 1993. *Modifications That Preserve Language and Content*. Technical Report. ERIC.
[102] Jaclyn Loo. 2017. The future of travel: New consumer behavior and the technology giving it flight. Google/Phocuswright Travel Study 2017. Retrieved on 27 August 2021 from https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/new-consumer-travel-assistance/?utm_source=HospitalityTrends.
[103] Ewa Luger and Abigail Sellen. 2016. 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*. ACM, New York, NY, 5286–5297.
[104] Rhonda W. Mack, Julia E. Blose, and Bing Pan. 2008. Believe it or not: Credibility of blogs in tourism. *Journal of Vacation Marketing* 14, 2 (2008), 133–144.
[105] François Mairesse and Marilyn A. Walker. 2009. *Can Conversational Agents Express Big Five Personality Traits through Language?: Evaluating a Psychologically-Informed Language Generator*. Cambridge & Sheffield, Cambridge University Engineering Department & Department of Computer Science, University of Sheffield.
[106] Bronislaw Malinowski. 1923. *The Problem of Meaning in Primitive Languages*. Harcourt, Brace & World, Inc, New York, Chapter Supplement to C.K., 296–336.
[107] Irina Maslowski, Delphine Lagarde, and Chloé Clavel. 2017. In-the-wild chatbot corpus: From opinion analysis to interaction problem detection. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing*. International Science and General Applications, 115–120.
[108] Dominic W. Massaro, Michael M. Cohen, Sharon Daniel, and Ronald A. Cole. 1999. Developing and evaluating conversational agents. In *Human Performance and Ergonomics* (2nd ed.). P. A. Hancock (Ed.), Academic Press, Cambridge, Chapter 7, 173–194.
[109] Niamh McNamara and Jurek Kirakowski. 2006. Functionality, usability, and user experience. *Interactions* 13, 6 (2006), 26–28.
[110] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications* 129 (2019), 56–67.
[111] Junko Mori. 1992. *About What and What about: The Distinction between Topic and Issue in Conversations*. Ph.D. Dissertation. University of Delaware.
[112] Thomas William Morris. 2002. Conversational agents for game-like virtual environments. In *Artificial Intelligence and Interactive Entertainment*, Ken Forbus and Magy Seif El-Nasr. AAAI Press, Palo Alto, CA, 82–86.
[113] Kellie Morrissey and Jurek Kirakowski. 2013. 'Realness' in Chatbots: Establishing quantifiable criteria. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, Berlin, Heidelberg, 87–96.

[114] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 72–78.
[115] Mario Neururer, Stephan Schlögl, Luisa Brinkschulte, and Aleksander Groth. 2018. Perceptions on authenticity in chat bots. *Multimodal Technologies and Interaction* 2, 3 (2018), 60.
[116] Craig Nevill and Timothy Bell. 1992. Compression of parallel texts. *Information Processing & Management* 28, 6 (1992), 781–793.
[117] Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.
[118] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17, C (2018), 22–27.
[119] Brian Paltridge. 1994. Genre analysis and the identification of textual boundaries. *Applied Linguistics* 15, 3 (1994), 288–299.
[120] Aneta Pawłowska. 2016. 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 (2016), 106–115.
[121] Dijana Peras. 2018. Chatbot evaluation metrics: Review paper. In *Proceedings of the 36th International Scientific Conference on Economic and Social Development*. Varazdin Development and Entrepreneurship Agency, Varazdin, 89.
[122] Claudio S. Pinhanez. 2020. HCI research challenges for the next generation of conversational systems. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. ACM, New York, NY, 1–4.
[123] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*.
[124] Nicole M. Radziwill and Morgan C. Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional* 19, 3 (2017), 25–36.
[125] Thomas B. W. Reid. 1956. Linguistics, structuralism and philology. *Archivum Linguisticum* 8, 1 (1956), 28–37.
[126] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY.
[127] Valeria Resendez. 2020. *A Very Formal Agent: How Culture, Mode of Dressing and Linguistic Style Influence the Perceptions Toward an Embodied Conversational Agent?* Master's thesis. University of Twente.
[128] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender Systems Handbook*. F. Ricci, L. Rokach, B. Shapira, and P. Kantor (Eds.), Springer, Boston, MA, 1–35.
[129] David A. Robb, José Lopes, Stefano Padilla, Atanas Laskov, Francisco J. Chiyah Garcia, Xingkun Liu, Jonatan Scharff Willners, Nicolas Valeyrie, Katrin Lohan, David Lane, Pedro Patron, Yvan Petillot, Mike J. Chantler, and Helen Hastie. 2019. Exploring interaction with remote autonomous systems using conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. ACM, New York, NY, 1543–1556.
[130] E. Sato. 2007. A guide to linguistic modification: Strategies for increasing English language learner access to academic content. LEP Partnership. Retrieved on 27 August 2021 from https://ncela.ed.gov/files/uploads/11/abedi_sato.pdf.
[131] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *32nd AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California, 5456–5463.
[132] Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.
[133] Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2014. Towards natural clarification questions in dialogue systems. In *Proceedings of the AISB Symposium on Questions, Discourse and Dialogue*. Vol. 20, Society for the Study of Artificial Intelligence and Simulation of Behaviour, Brighton, 8.
[134] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307 (2008), 11. Retrieved from http://www.biomedcentral.com/1471-2105/9/307.
[135] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8, 25 (2007), 21. Retrieved from http://www.biomedcentral.com/1471-2105/8/25.
[136] Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. User expectations of conversational chatbots based on online reviews. In *Proceedings of the Designing Interactive Systems Conference 2021*. ACM, New York, NY, 1481–1491.
[137] Ekaterina Svikhnushina and Pearl Pu. 2021. Key qualities of conversational chatbots–the PEACE model. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 520–530.
[138] Jill Sweeney and Joffre Swait. 2008. The effects of brand credibility on customer loyalty. *Journal of Retailing and Consumer Services* 15, 3 (2008), 179–193.
[139] Bakhtiyar Hussain Syed. 2020. *Adapting Language Models for Style Transfer*. Master's thesis. International Institute of Information Technology Hyderabad.

[140] Benedikt Szmrecsanyi. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6, 1 (2011), 45–76.
[141] Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. 2018. The ethnobot: Gathering ethnographies in the age of IoT. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 604.
[142] Gunay Tariverdiyeva. 2019. *Chatbots' Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis*. Master's thesis. University of Twente.
[143] Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. 2016. An Investigation of conversational agent interventions supporting historical reasoning in primary education. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. Alessandro Micarelli, John Stamper, and Kitty Panourgia (Eds.), Springer International Publishing, Cham, 260–266.
[144] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'neill. 2017. How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians. In *Proceedings of the Human–Computer Interaction–INTERACT* . Bernhaupt R., Dalvi G., Joshi A., K. Balkrishan D., O'Neill J., and Winckler M. (Eds.). Lecture Notes in Computer Science, Vol. 10513. Springer, Cham, 441–459.
[145] Think with Google. 2016. How Micro-Moments are Reshaping the Travel Customer Journey. Retrieved 27 August, 2021 from https://www.thinkwithgoogle.com/marketing-resources/micro-moments/micro-moments-travel-customer-journey/.
[146] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*. ACM, New York, NY, 42–51.
[147] Rebecca Ruiz Thomas Combrink, Melinda Bradford. 2018. *2017-2018 Flagstaff Visitor Survey*. Technical Report. 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.
[148] Alexey Tikhonov and Ivan P. Yamshchikov. 2018. What is wrong with style transfer for texts? CoRR
[149] Peter Trudgill. 2000. *Sociolinguistics: An Introduction to Language and Society*. Penguin Books Limited, London.
[150] Thomas Tullis and William Albert. 2013. *Measuring the User Experience, Second Edition: Collecting, Analyzing, and Presenting Usability Metrics* (2nd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA.
[151] Iis Tussyadiah. 2020. 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, C (2020), 102883.
[152] UNWTO. 2017. UNWTO Tourism Highlights, 2017 Edition. World Tourism Organization, Madrid.
[153] M. S. Walgama and B. Hettige. 2017. Chatbots: The next generation in computer interfacing–A review. In *Proceedings of the KDU International Research Conference* . General Sir John Kotelawala Defence University, 7.
[154] Peter Wallis and Bruce Edmonds. 2019. How language works & what machines can do about it. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM, New York, NY, 1–3.
[155] Dan Wang, Zheng Xiang, and Daniel R. Fesenmaier. 2016. Smartphone use in everyday life and travel. *Journal of Travel Research* 55, 1 (2016), 52–63.
[156] Leo Wanner, Elisabeth André, Josep Blat, Stamatia Dasiopoulou, Mireia Farrús, Thiago Fraga, Eleni Kamateri, Florian Lingenfelser, Gerard Llorach, Oriol Martínez, and G. Meditskos. 2017. Design of a knowledge-based agent as a social companion. *Procedia Computer Science* 121, C (2017), 920–926.
[157] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching strategies for successful human-agent interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 448–459.
[158] Wikibooks. 2020. Algorithm Implementation/Strings/Levenshtein distance—Wikibooks, The Free Textbook Project. Retrieved May 8, 2020 from https://en.wikibooks.org/w/index.php?title=Algorithm_Implementation/Strings/Levenshtein_distance&oldid=3678144.
[159] Katerina Zdravkova. 2000. Conceptual framework for an intelligent chatterbot. In *Proceedings of the 22nd International Conference on Information Technology Interfaces*. IEEE, New York, NY, 189–194.
[160] Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. 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. Association for Computational Linguistics, 1098–1107. DOI: https://doi.org/10.18653/v1/P18-1101
[161] Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. 2018. Lingke: A fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 108–112.
[162] Darius Zumstein and Sophie Hundertmark. 2017. Chatbots-An interactive technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet* 15, 1 (2017), 96–109.

Received April 2021; revised July 2021; accepted September 2021