The impact of chatbot linguistic register on user perceptions: a replication study

Ana Paula Chaves[1] and Marco Aurelio Gerosa[2]

[1] Federal University of Technology - Parana, Campo Mourao–PR, Brazil
anachaves@utfpr.edu.br
[2] Northern Arizona University, Flagstaff AZ
marco.gerosa@nau.edu

Abstract. Chatbots often perform social roles associated with human interlocutors; hence, designing chatbot language to conform with the stereotypes of its social category is critical to the success of this technology. In a previous study, Chaves et al. performed a corpus analysis to evaluate how language variation in an interactional situation, namely the linguistic *register*, influences the user's perceptions of a chatbot. In this paper, we present a replication study with a different corpus to understand the effect of corpus selection in the original study's findings. Our results confirm the findings in the previous study and demonstrate the reproducibility of the research methodology; we also reveal new insights about language design for tourist assistant chatbots.

Keywords: Chatbot · Register · User perceptions.

1 Introduction

Many companies have adopted chatbots to offer 24/7 customer support [20] and to reduce the need for human support for several domains. Accordingly, chatbots often assume social roles traditionally associated with a human service provider, for example, a tutor [46], a healthcare provider [34], or a tourist assistant [42]. Human interlocutors usually have a mental model of the interaction with a representative of these roles. Grounded in the media equation theory [40, 18], which states that people respond to communication media and technologies as they do to other people, it is reasonable to assume that chatbot users project expectations in their interactions with chatbots.

One way to enhance a chatbot's impersonation of a social role is to carefully plan their use of language [33, 23]. When a chatbot uses unexpected levels of (in)formality or incoherent language style, the interaction may result in frustration or awkwardness [33]. Previous research has analyzed user preferences regarding chatbot language use [17, 1, 45, 43]; however, these studies focus on varying levels of formality or style, which is often disassociated from the particular interactional situation. Similar consideration can be made for some commercially available chatbots. For example, Golem[3], a chatbot designed to guide tourists

[3] Available at http://m.me/praguevisitor. Last accessed: November, 2021

through Prague (Czech Republic), utters sentences extracted from an online travel magazine[4] without any adaptation to the new interactional situation.

In sociolinguistics, the concept that defines how humans adapt their language use depending on the interactional situation is called "register" [8, 4]. Register consists of the relationship between the occurrences of *core linguistic features* in a conversation given the *context*. For example, the core linguistic features a person uses when writing an email to their supervisor are different from those used when texting a friend due to variation in the situational parameters (e.g., a supervisor vs. a friend; email vs. chat messaging tool). The *linguistic features* are the grammatical characteristics in the conversation (e.g., nouns, personal pronouns, or passive voices). The *context* is determined by a set of *situational parameters* that characterize the interactional situation (e.g., the participants and the relationship between them, channel, production circumstances, topic, and purpose) [8]. Despite being considered as one of the most important predictors of linguistic variation in human-human communication [5], register has not yet been widely explored as a theoretical basis for chatbot language design.

Previous studies [12, 11] explored language variation in the context of tourism-related interactions. The results confirmed that the core linguistic features vary as the situational parameters vary, resulting in different language patterns. A more recent study identifies whether register influences the user's perceptions of their interaction with chatbots [13], based on a corpus analysis approach. The findings show an association between linguistic features and user perceptions of appropriateness, credibility, and overall user experience, which point to the need to consider register for the design of chatbots.

The study presented by Chaves et al. [13] is grounded on corpus analysis. Although corpus analysis is a powerful approach to detect register characteristics [15], it may also bring a limitation: the nature of the corpus may influence how extreme the study's participants perceived the register differences. For example, if the register of the selected corpus is too far away from the expected language patterns, then the likelihood that participants perceive it as uncanny may increase. Therefore, the question remains whether the conclusion presented in that study is supported if a different corpus is selected.

In this paper, we present a methodological replication study of [13], as we applied the same methods as the original study but used a different corpus of conversations in the tourism domain. Our goal is to investigate the effect of corpus selection on the previous outcomes and demonstrate whether the relations between linguistic features and user perceptions still stand for the new setting. According to [16], replication is a valuable scientific resource as it allows methodological enhancement and improves confidence in the scientific findings.

2 Background

Chatbots are disembodied conversational interfaces that interact with users in natural language via a text-based messaging interface [14, 25]. Human-chatbot

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

interactions are built upon the use of language. Therefore, scholars have put effort into improving chatbot conversational skills on several fronts. Research on natural language generation has heavily focused on ensuring that chatbots produce coherent and grammatically correct responses and on improving functional performance and accuracy (see e.g. [31, 38]). Other studies have focused on comparing how humans adapt their language when interacting with chatbots, for example, by matching with the chatbot vocabulary [30] or with the language style [26]. However, the literature has overlooked how the chatbot should sound.

In the chatbot field, language often complies with the individual characteristics of one intended persona [27, 24], which fits the definition of style [8]. Recent studies have focused on evaluating the effect of language style on the user's perceptions [17, 43, 37, 44, 37], many of them comparing different levels of formality. However, the formality of a chatbot should depend on how much the human interlocutors associate formality to the social category that the chatbot represents. The association between one's language use and the interactional situation goes beyond the scope of language style, inviting designers to account for *register*.

As introduced in Section 1, the linguistic register consists of the functional association between the occurrences of linguistic features and a given context. Register has not been widely investigated in the context of chatbots, with only a few studies pointing to its relevance [22, 35, 2]. As an effort to introduce the linguistic register as a theoretical basis for chatbot's language design, Chaves et al. [13] performed a study to evaluate how human interlocutors perceive register differences in a chatbot's discourse. The authors concluded that register could work as a tool to define what language pattern is appropriate for a given context, encapsulating the varying language styles that a persona-based chatbot might have. We are not aware of other studies that evaluate the influence of register on user experience with chatbots or the effect of corpus selection on the outcomes of register analysis. Thus, this paper presents a replication study to understand the influence of corpus selection on the study performed by Chaves et al. [13].

3 User perceptions of register: the original study

This section summarizes the method applied in [13], which we replicate in this study. In Sections 4 through 7 we detail these steps for the replication study.

Chaves et al. [13] explored the extent to which behavioral aspects of user experience (namely perceived appropriateness, credibility, and user experience) relate to the register a chatbot uses. The authors invited participants to compare excerpts of conversations expressed in two different registers. To isolate the effect of register, the content of the conversation needed to be equivalent, which means using parallel data–natural language texts with the same semantic content, but expressed in different forms [41]. Since this kind of data is rarely available, the study's approach includes the production of parallel corpora. Actual conversations were carefully manipulated to mimic the register characteristics of a different corpus. The research methodology followed the steps outlined below (see [13] for details):

Step 1–Data collection: the baseline corpus, named FLG, consists of text-based interactions between three human tourist assistants and tourists from Flagstaff, Arizona, USA. The corpus comprises 144 interactions with about 540 question-answer pairs. A second corpus is used for comparison purposes. In that study, the authors extracted this corpus from a larger one named *DailyDialog* [36].

Step 2–Register characterization: this step consists of the characterization of the registers present in each corpus through register analysis [4]. The authors first identified the situation in which the conversations occur and then defined the prevailing linguistic features in each corpus. The analysis relied on information from the Biber's grammatical tagger [6] and the linguistic features were analyzed both individually and aggregated into five dimensions according to the text-linguistic register framework [4].

Step 3–Text modification: this step aims at producing a new, parallel corpus, named FLG_{mod}. For every answer provided by a tourist assistant in the FLG corpus, the authors produced a corresponding answer that portrays the register characteristics of *DailyDialog*. FLG and FLG_{mod} fulfill the requirement of parallel data for user studies on register differences.

Step 4–The study: finally, the authors performed a user study to evaluate the impact of register on user experience. Participants were presented with individual questions and pairs of answers (from FLG and FLG_{mod}) and, for each, were asked to choose which answer they preferred based on three measures of quality: appropriateness, credibility, and user experience. The analysis included fitting a statistical learning model to identify the linguistic features that best predict the user choices.

All the research materials related to Chaves et al.'s study are available on GitHub [10]. The following sections present how we apply this methodological approach in our replication study.

4 Data collection

In Chaves et.al. [13], the situational parameters of *DailyDialog* have a large variability, particularly for the interlocutor's role and relationship among them. For example, *DailyDialog* includes conversations between travelers and immigration control personnel, guests and hotel concierge, tourists and tour guides, among many others. For the replication, we want to select a corpus with less variability of situational parameters within the corpus. We chose the *Frames* [3] dataset, which is a corpus of 1349 human-human, text-based interactions[5] within the context of booking travel packages. We followed the situational analytical framework [8] to identify the situational parameters in comparison to FLG (Table 1). The purpose of this *situational analysis* is to characterize the interactions using a conversational taxonomy based around seven parameters (Table 1). The variation in the situational parameters between *Frames* and FLG is mainly

[5] Download and more details at https://www.microsoft.com/en-us/research/project/frames-dataset/

in the purpose and topic parameters; *Frames* focuses on pre-travel decision-making, while *FLG* focuses on en-route information search.

5 Register characterization

We characterized the register of the *Frames* corpus as presented in [13]. In a nutshell, we submitted the conversations from *Frames* to the Biber's grammatical tagger [6], which tags and counts linguistic features. The features are also aggregated into *dimension scores* using a factor analysis algorithm [6] to reveal the prevailing characteristics of the register (i.e., the levels of personal involvement, narrative flow, contextual references, persuasion, and formality) [4].

We applied a one-way MANOVA to generate a statistical comparison of the dimension scores across corpora (*Frames* and *FLG*), where the dependent variables are the values of the five dimension scores. The independent variables are the *Frames* (control group) and the three tourist assistants from the *FLG* corpus, namely $TA1$, $TA2$, and $TA3$ (experimental groups). We considered each tourist assistant as a group to account for stylistic variation among them. The MANOVA revealed that the dimension scores for *FLG*'s tourist assistants significantly differ from the dimension scores for *Frames* ($Wilks = 0.95, F = 5.71, p < 0.0001$). We also performed a one-way univariate analysis ($df = 3, 1489$) for each of the five dimensions to identify the individual dimensions that influence the prevailing register characteristics (Table 2).

The dimension score reveals that the purpose (decision-making vs. information search) likely impacted the narrative flow (dimension 2) and the persuasion (dimension 4) since the tourist assistant in *Frames* focused on describing the options rather than arguing on what were the best deals. The variation within the TAs suggests the influence of stylistic preferences [11].

Table 1: Situational analysis (*Frames* vs. *FLG*).

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

Table 2: Univariate analysis of dimension scores ($df = 3, 1489$). For each dimension, the table shows the estimated dimension score ± the standard error per group (*Frames*, $TA1$, $TA2$, $TA3$), and the corresponding F- and p-values.

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

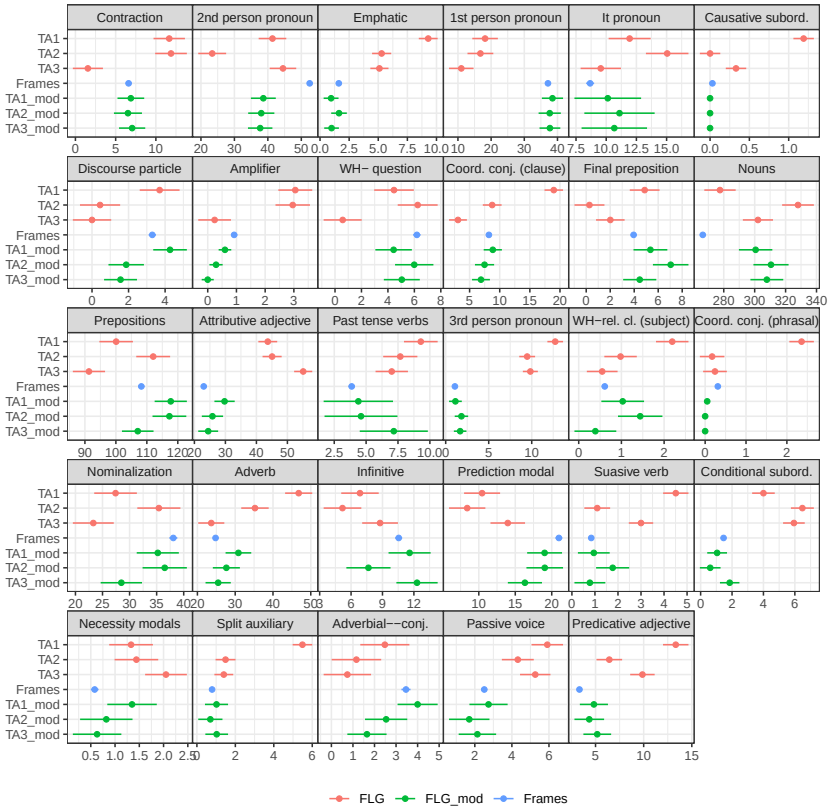


Fig. 1: Visualization of ANOVA results for individual features

Figure 1 depicts the linguistic features that vary significantly between the two corpora (FLG and $Frames$), as revealed by the ANOVA analysis per feature. The figure shows the estimates for $Frames$ (control group, in blue) and each tourist assistant in FLG ($TA1$, $TA2$, and $TA3$, in red). The horizontal line represents the standard error. We found that 29 out of 48 linguistic features were significantly different across corpora. The next step was to modify the conversations to produce another parallel corpus.

6 Text modification

Text modification process followed the same procedures as described in [13]. Firstly, we cloned the FLG corpus to create the initial FLG_{mod_2}. Then, we inspected the $Frames$ corpus to understand how one particular feature is used in $Frames$. Then, we reproduced the use in FLG_{mod2} using a Python script (see [10]). Figure 1 shows the comparison between $Frames$ and FLG_{mod_2} after performing a sufficient number of modifications to substantially reduce the F-values. The figure shows the estimates for $Frames$ (control group, in blue) and

Table 3: Example of a modified answer (FLG vs. FLG_{mod_2}).

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

each tourist assistant in FLG ($TA1_mod$, $TA2_mod$, and $TA3_mod$, in green). The horizontal line represents the standard error. Table 3 shows an example of a modified answer, where modified words are highlighted in bold, and the tags attributed to the words are between square brackets.

Once the FLG_{mod_2} parallel corpus was developed, we perform the user study on the impact of linguistic register on user perceptions.

7 User perceptions study

Following the procedures described in [13], we ranked the question-answer pairs in our parallel corpora based on the Levenshtein distance [32] between the pairs of original (FLG) and modified answers (FLG_{mod2}), and selected the top 10% (54 question-answer pairs) for evaluation.

We recruited participants to answer an online questionnaire, where they chose, given a question, which answer would better represent a tourist assistant chatbot. For each tourist's question, presented on the screen one at a time, participants could choose one out of three options: the original answer (from FLG), the modified version (from FLG_{mod2}), or "I don't know" (see an example in Figure 2). Original and modified answers were presented in a randomized order. In total, participants answered 27 questions, nine for each of the constructs (perceived language appropriateness, perceived credibility, and user experience).

7.1 Participants

Participants were recruited through Prolific[6] in September 2020. We received a total of 174 submissions, 29 of which were discarded due to either technical issues in the data collection or failure to answer the attention checks ($N = 145$). All the participants claimed English as their first language and were located in the USA. Additionally, we configured Prolific to recruit only participants who did not participate in the previous study, to avoid biases in the data collection. Most participants had a four-year bachelor's degree (49) or some college, but no degree (42). Twenty participants graduated from high school, and 17 had

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

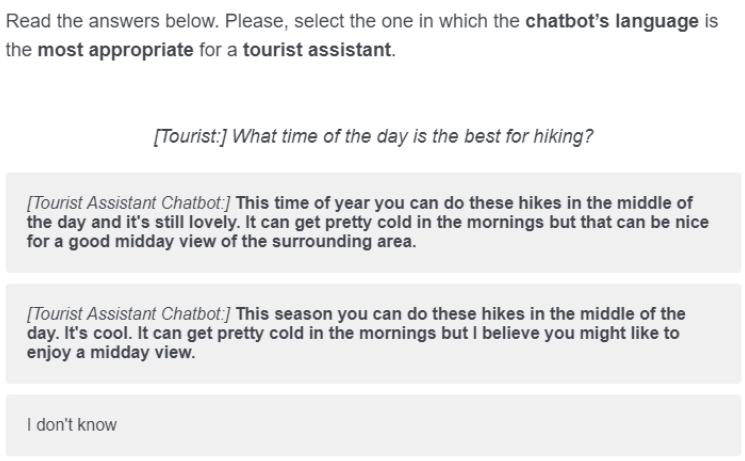


Fig. 2: Example of a question. The participant was invited to select the answer that portrays the most appropriate language.

Master's degrees. Common educational backgrounds were STEM (34), Arts and Humanities (28), and others (29). Three participants had non-binary gender; 70 declared themselves as female, and 72 as male. The age range is 18-60 ($\mu = 30.77$ years-old, $\sigma = 10.32$).

7.2 Analysis of the linguistic features

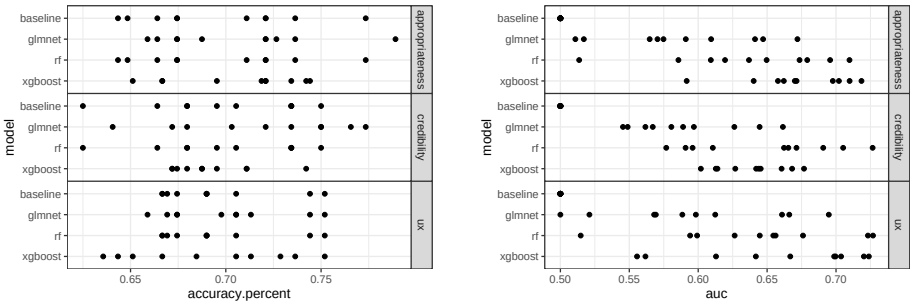
We fitted the generalized linear model (GLM), using the *glmnet* package in R [21], using the same cross-validation algorithm presented by [13][7]. For comparison purposes (to determine an upper bound on prediction accuracy), we also fitted two non-linear learning models: random forest and gradient boosting, as performed in [13]. The prediction variables include (i) the difference between original and modified counts per linguistic feature; (ii) variables representing the participant who answered the question; (iii) the participants' self-assessed social orientation; and (iV) the author of the answer in FLG ($TA1$, $TA2$, $TA3$).

The evaluation dataset started with 3,915 observations (145 participants, 27 evaluations per participant). We discarded blank answers and the "I don't know" option, resulting in a dataset with 3,858 observations. Each question-answer pair was evaluated from 22 to 26 times per construct. As in the original study [13], participants overall preferred the answers from the original corpus, although the modified version was sometimes chosen.

Figure 3 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). The prediction accuracy of `glmnet` and `xgboost` are only slightly

[7] The R code and datasets are available on GitHub [10].

better than the baseline. Nevertheless, the AUC values are consistently better than the baseline. As in the original study [13], the non-linear models are not considerably more accurate than the linear model, which justify the use of the *glmnet* model.



(a) Accuracy (b) AUC

Fig. 3: Accuracy (a) and AUC (b) results per model for each construct (appropriateness, credibility, and user experience). The baseline represents a model that always predicts the most frequent class (original).

Table 4 presents the coefficients of the linguistic features selected in six or more folds. The first and second columns indicate, respectively, the linguistic feature of interest and the sign of original − modified calculation, which indicates whether one particular feature was increased or decreased in the text modification process. A positive sign (+) for a feature f_i indicates that $\text{count}_{\text{original}}(f_i) > \text{count}_{\text{modified}}(f_i)$, while a negative sign (-) indicates the opposite ($\text{count}_{\text{original}}(f_i) < \text{count}_{\text{modified}}(f_i)$). 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, while features with positive coefficients increase the likelihood of the model predicting the modified class.

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

Original answers also have significantly more *agentless passives* and *third-person pronouns*. These features have a positive coefficient for all three constructs, indicating that frequent occurrences of these features increase the likelihood of modified answers being chosen. This outcome suggests that participants

Table 4: Coefficients and standard deviation of the non-zero variables per construct. The dots indicate that the corresponding feature was not selected for that particular construct.

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

are more likely to prefer answers in which these features are less frequent. The same conclusion applies to *"it" pronouns*, *conditional subordinations*, *attributive adjectives*, and *emphatics*, but these features are not relevant for all constructs. *"It" pronouns* did not show up as relevant for user experience, *conditional subordinations* and *attributive adjectives* did not influence appropriateness, and *emphatics* show up as relevant only for user experience.

Modifications have significantly more *prepositions* than the original answers. This feature has a negative coefficient for all three constructs, which indicates that frequent occurrences of *prepositions* increase the likelihood of original answers being chosen. This outcome suggests that participants are more likely to prefer answers in which these features are less frequent. The same inference applies to *infinitives* and *nominalizations*, although *infinitives* did not show up as a relevant feature for user experience, and *nominalizations* did not predict the appropriateness. Modifications also have a larger number of *prediction modals*. This feature has a positive coefficient for all three constructs, which indicates that increasing their occurrences increased the likelihood of modified answers being chosen. This outcome highlights the preferences for answers in which these features are more consistently present. The same conclusion applies to *adverbial–conjuncts*, although this feature shows up as relevant only for credibility. Noticeably, when we aggregate the estimate to the standard deviation for this feature, it sums up to zero, suggesting that this outcome may be noise.

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

7.3 Discussion

Chaves et al.from [13] showed that using language that fits to a particular context, i.e, that is register-specific, has a significant impact on user perceptions of their interaction with chatbots. In this study, we replicate that study's methodology to investigate whether this conclusion is supported if FLG is compared to a different corpus. Our results supported the insights from the original study. The study strengthens the conclusion that the chatbot's language–characterized under the lens of register–can impact the user perceptions, which, ultimately, may result in increased quality, acceptance, and adoption [29, 39, 19].

Secondly, the analysis of individual linguistic features supports the previous inferences for at least four linguistic features, namely *preposition*, *causative subordination*, *third-person pronoun*, and *conditional subordination*. In this replication study, *preposition* was selected for all the constructs, and participants preferred answers in which this feature is less frequent. Prepositions (e.g., *at, in, of*, etc) in our study were often used to provide an extra piece of information (e.g., *in Flagstaff, at 4am*). Many occurrences of this feature may reduce the efficiency, violating the maxim of quantity that states that a sentence in a conversation should have just the right amount of content [28].

Conditional subordination is negatively associated with credibility in both studies, which supports the inference that when the chatbot gives options, it sounds as it is not confident about the information provided [13]. Using less conditional subordination requires more personalization so that the chatbot is assertive when offering a suggestion or recommendation, instead of offering a list of options (i.e., ***if*** *you like hiking, then check [...]; otherwise, check [...]*).

Regarding *causative subordination* (e.g., *because*), participants are more likely to prefer answers in which this feature occurs. Noticeably, this feature is a fairly uncommon feature (mean of 1 per 1,000 words [4]) when compared to others such as *nouns* (mean of 180 per 1,000 words [4]). Our result indicates that, when this feature occurs, it is preferred over receiving the information without the subordination. This preference can be associated with human-likeness since the absence of the *causative subordination* results in a broken discourse, which is associated with robotic sounding (e.g., *[...] it was used as a major trading hub,* ***because*** *there are artifacts that were found there [...]* vs. *[...] it was used as a major trading hub. There are artifacts that were found there [...]*).

Participants preferred less frequent occurrences of *third-person pronouns* in all three constructs in the current study. *Third-person pronouns* are used in FLG to add details about business or attractions (e.g., "*they* have public restrooms"), which may reduce efficiency. Additionally, by using *third-person pronouns*, the assistant provides the information from an external standpoint, which results in a more impersonal tone [9] (e.g., "**they** *[the museum community]* have musical performances" vs. "**we** *[the assistant as part of the museum community]* have musical performances"). These results suggest that reinforcing the tourist assistant chatbot as a representative of its social category by using (plural) first-person pronouns would be preferable over the impersonal third-person pronoun.

On the other hand, this study shows conflicting outcomes regarding the levels of *prediction modals* and *contractions*. In [13], participants preferred lower levels of these features, whereas the current study indicates a preference for higher levels of these features in all the three constructs. This dissonance can be explained by the differences in how *Frames* and *DailyDialog* use these features. In *DailyDialog*, *would* is the most common *prediction modal*, which mostly co-occurred with first-person pronouns (i.e., *"I would", "I'd"*). As discussed in [13], the co-occurrences with first-person pronouns may have influenced the outcomes for prediction modals and contractions, as these co-occurrences cause uncanny effects due to excessive personification. In contrast, *Frames* has more frequent occurrences of other forms of *prediction modals*, such as *shall* and *will*. *Will* is the most frequent modal and it often co-occurs with *nouns* (e.g., *"campgrounds will open"* or *"downtown will have vegetarian options"*) instead of personal pronouns. As a consequence, the negative effect was flattened and the frequent occurrences of *prediction modals* and *contractions* resulted in a positive effect.

This study allowed us to observe inferences about features not manipulated in the original study. *Split auxiliaries* are likely influenced by the preferences for frequent occurrences of *adverbs*, as split auxiliaries occur when adverbs are placed between auxiliaries and their main verb [4] (e.g., *"will* ***obviously*** *limit"*). Many *adverbs* are used in *FLG* to indicate the tourist assistant's stance (e.g., *"Absolutely!"* and *"there are definitely some,"* which indicate assurance, or *"you'd probably be fine,"* which indicates uncertainty). These features emphasize the level of confidence that the assistant has about the information, which positively affects the user perceptions of all the evaluated constructs. The same conclusions apply to *predicative adjectives*, which is also frequently used for marking stance [4]. This feature was selected as a predictor of credibility, which clearly relates to the ability to express opinion (e.g., *"That would* ***be difficult.*** *" "the sandwiches* ***are delicious,*** *" "the restaurant* ***is good****"*).

In contrast, results show that participants are more likely to prefer answers in which the occurrences of *agentless passives*, *infinitives*, *nominalizations* are less frequent. These features are, in general, uncommon in conversations [9, 4], which may justify the user's negative impressions about higher frequencies of these features. *Emphatics* and *"it" pronouns* are rather frequent in conversations. However, *emphatics* are characteristic of informal, colloquial discourse, whereas "it" pronouns indicate limited informational content [4]. In this research, the tourist assistant chatbots are representatives of a professional category specialized in providing information, which may increase the user's expectations for formal and specialized discourse. This results contradicts the results presented by [37], which showed that customers may not assign different roles to chatbots as communication partner in a human-chatbot customer service setting.

Contributions and implications The results presented in this paper, combined with the findings from [13], emphasize that there is more to chatbot language variation than the dichotomy of formal vs. informal language. Previous literature that focuses on the spectrum of formality has found inconsistent results about whether chatbot language should sound (in)formal (see, e.g., [37]).

[37] even suggested that different perceptions of formality may have been influenced by the function which some linguistic features perform in the utterances. Our results reinforced that disregarding register may overshadow language variation's nuances and result in an overly simplified model for chatbot language. Research on sociolinguistics has long stressed the role of register in predicting language variation in human-human conversations [7]. Our findings demonstrate that this role can be stretched to include human-chatbot conversations and that register analysis is a powerful tool for characterizing the patterns of language that a chatbot would be expected to use within a particular domain.

Additionally, this paper brings insights on the user expectations about chatbots' language use in the context of tourism information search. Firstly, it reinforces the need for language efficiency; as the authors point out in [13], efficiency is crucial in information search scenarios, which makes detailed descriptions unnecessary in many cases. Thus, filling words or expressions, and additional pieces of information should be avoided. Secondly, it provides understanding on possible linguistic features that (overly) convey human-likeness. The study showed that users prefer fluid, connected sentences ("and", "or", "because") rather than several simple utterances; and that positioning the chatbot as part of a social group (e.g. using "*we*", but not overusing "I") is preferred over impersonal tone and external standpoint ("they", "there are"). Finally, it is important that the chatbot has the ability to express opinion by using predicative adjectives (e.g. "is good") and make it clear the level of confidence about the information (even when the assistant is not so sure), which can be expressed with specific adverbs (e.g. "absolutely", "probably").

8 Conclusions

This paper presented a methodological replication study that aims at demonstrating the reproducibility of the methodology proposed by Chaves et al. [13] to characterize the register of chatbot discourse. Additionally, this replication helps understand the effect of corpus selection in previous findings. Our results confirmed the influence of using register-specific language in user perceptions of their interactions with chatbots and reinforced the need to consider register when designing chatbot language. Since we performed a methodological replication, the limitations discussed in [13] also applies to this study, including the subjectivity introduced by the semi-manual text modification and the choice of the linguistic features submitted to the *glmnet* model for feature selection. Additionally, we emphasize that the conversations in the *Frames* corpus were collected in a lab setting, so the language expressed in the corpus reproduces the expectations of the users regarding the tourist assistant's patterns of language.

Our findings open a wide range of new research opportunities in chatbot language design. Given the identified relevance of linguistic register for user experience, the next natural step is to walk toward building computational models to perform register adaptation, enabling the future development of chatbot conversational engines both tailored to the target context and able to adapt register

to the specific communicative purpose dynamically. Additionally, we need to develop efficient frameworks to automate the steps of this methodology to facilitate the register characterization and linguistic feature selection for particular contexts. We invite the research community to embrace these challenges to increase chatbots' acceptance as online service providers and improve user experiences with this technology.

References

1. Araujo, T.: Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in Human Behavior **85**, 183–189 (2018)
2. Argamon, S.: Register in computational language research. Register Studies **1**(1), 100–135 (2019)
3. Asri, L.E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., Suleman, K.: Frames: A corpus for adding memory to goal-oriented dialogue systems. In: Proceedings of the SIGDIAL 2017 Conference. pp. 207–219. Association for Computational Linguistics, Saarbrucken, Germany (2017)
4. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge, UK (1988)
5. Biber, D.: Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory **8**(1), 9–37 (2012)
6. Biber, D.: Mat–multidimensional analysis tagger. available at: https://goo.gl/u7h9gb (2017)
7. Biber, D.: Text-linguistic approaches to register variation. Register Studies **1**(1), 42–75 (2019)
8. Biber, D., Conrad, S.: Register, genre, and style. Cambridge University Press, New York, NY, USA, 2 edn. (2019)
9. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., Quirk, R.: Longman grammar of spoken and written English, vol. 2. Pearson Longman, London, UK (1999)
10. Chaves, A.P.: Github repository. available at: https://github.com/chavesana/chatbots-register (2020)
11. Chaves, A.P., Doerry, E., Egbert, J., Gerosa, M.: It's how you say it: Identifying appropriate register for chatbot language design. In: Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19). p. 8. ACM, New York, NY, USA (Oct 2019). https://doi.org/10.1145/3349537.3351901
12. Chaves, A.P., Egbert, J., Gerosa, M.A.: Chatting like a robot: the relationship between linguistic choices and users' experiences. In: ACM CHI 2019 Workshop on Conversational Agents: Acting on the Wave of Research and Development. p. 8. https://convagents.org/, Glasgow, UK (2019)
13. Chaves, A.P., Egbert, J., Hocking, T., Doerry, E., Gerosa, M.A.: Chatbots language design: the influence of language variation on user experience. ACM Transactions on Computer-Human Interaction (*to appear* Author's version at arXiv:210111089)
14. Chaves, A.P., Gerosa, M.A.: How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. International Journal of Human–Computer Interaction **0**(0), 1–30 (2020). https://doi.org/10.1080/10447318.2020.1841438, https://doi.org/10.1080/10447318.2020.1841438

15. Conrad, S., Biber, D.: Multi-dimensional Studies of Register Variation in English. Routledge, New York, NY, USA (2014)
16. Dennis, A.R., Valacich, J.S.: A replication manifesto. AIS Transactions on Replication Research **1**(1), 1 (2015)
17. Elsholz, E., Chamberlain, J., Kruschwitz, U.: Exploring language style in chatbots to increase perceived product value and user engagement. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. pp. 301–305. ACM, New York, NY, USA (2019)
18. Fogg, B.: Computers as persuasive social actors. In: Fogg, B. (ed.) Persuasive Technology, chap. 5, pp. 89 – 120. Interactive Technologies, Morgan Kaufmann, San Francisco (2003)
19. Følstad, A., Brandtzaeg, P.B.: Users' experiences with chatbots: findings from a questionnaire study. Quality and User Experience **5**, 1–14 (2020)
20. Følstad, A., Nordheim, C.B., Bjørkli, C.A.: What makes users trust a chatbot for customer service? an exploratory interview study. In: International Conference on Internet Science. pp. 194–208. Springer (2018)
21. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, Articles **33**(1), 1–22 (2010). https://doi.org/10.18637/jss.v033.i01, https://www.jstatsoft.org/v033/i01
22. Gnewuch, U., Morana, S., Maedche, A.: Towards designing cooperative and social conversational agents for customer service. In: International Conference on Information Systems 2017, Proceedings 1. Association for Information Systems, South Korea (2017)
23. Go, E., Sundar, S.S.: Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Computers in Human Behavior **97**, 304–316 (Aug 2019)
24. Google Developers: Conversation design by google. https://developers.google.com/assistant/conversation-design/what-is-conversation-design. Last access in March, 3, 2021 (2021), https://developers.google.com/assistant/conversation-design/what-is-conversation-design
25. Grudin, J., Jacques, R.: Chatbots, humbots, and the quest for artificial general intelligence. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–11. ACM, New York, NY (2019)
26. Hill, J., Ford, W.R., Farreras, I.G.: Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. Computers in Human Behavior **49**, 245–250 (2015)
27. Hwang, S., Kim, B., Lee, K.: A data-driven design framework for customer service chatbot. In: International Conference on Human-Computer Interaction. pp. 222–236. Springer (2019)
28. Jacquet, B., Hullin, A., Baratgin, J., Jamet, F.: The impact of the gricean maxims of quality, quantity and manner in chatbots. In: 2019 International Conference on Information and Digital Technologies (IDT). pp. 180–189. IEEE (2019)
29. Jakic, A., Wagner, M.O., Meyer, A.: The impact of language style accommodation during social media interactions on brand trust. Journal of Service Management **28**(3), 418–441 (2017)
30. Jenkins, M.C., Churchill, R., Cox, S., Smith, D.: Analysis of user interaction with service oriented chatbot systems. In: Jacko, J.A. (ed.) Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. pp. 76–83. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)

31. Jiang, R., E Banchs, R.: Towards improving the performance of chat oriented dialogue system. In: 2017 International Conference on Asian Language Processing (IALP). pp. 23–26. IEEE, New York, NY, USA (2017)
32. Kessler, B.: Computational dialectology in irish gaelic. In: Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics. p. 60–66. EACL '95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995). https://doi.org/10.3115/976973.976983, https://doi.org/10.3115/976973.976983
33. Kirakowski, J., Yiu, A., et al.: Establishing the hallmarks of a convincing chatbot-human dialogue. In: Human-Computer Interaction. InTech, London, UK (2009)
34. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y., et al.: Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association **25**(9), 1248–1258 (2018)
35. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, B.: A persona-based neural conversation model. arXiv preprint arXiv:1603.06155 (2016)
36. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. In: International Joint Conference on Natural Language Processing (IJCNLP). pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (2017)
37. Liebrecht, C., Sander, L., van Hooijdonk, C.: Too informal? how a chatbot's communication style affects brand attitude and quality of interaction. In: Conversations 2020: 4th international workshop on chatbot research (2020)
38. Maslowski, I., Lagarde, D., Clavel, C.: In-the-wild chatbot corpus: from opinion analysis to interaction problem detection. In: International Conference on Natural Language, Signal and Speech Processing. pp. 115–120. International Science and General Applications, Marocco (2017)
39. Morrissey, K., Kirakowski, J.: 'realness' in chatbots: Establishing quantifiable criteria. In: International Conference on Human-Computer Interaction. pp. 87–96. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
40. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 72–78. ACM, New York, NY, USA (1994)
41. Nevill, C., Bell, T.: Compression of parallel texts. Information Processing & Management **28**(6), 781–793 (1992)
42. Pillai, R., Sivathanu, B.: Adoption of ai-based chatbots for hospitality and tourism. International Journal of Contemporary Hospitality Management (2020)
43. Resendez, V.: 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 (2020)
44. Svenningsson, N., Faraon, M.: Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In: Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference on ZZZ. pp. 151–161 (2019)
45. Tariverdiyeva, G.: Chatbots' Perceived Usability in Information Retrieval Tasks: An Exploratory Analysis. Master's thesis, University of Twente (2019)
46. Tegos, S., Demetriadis, S., Tsiatsos, T.: An investigation of conversational agent interventions supporting historical reasoning in primary education. In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) International Conference on Intelligent Tutoring Systems. pp. 260–266. Springer International Publishing, Cham (2016)