

It's How You Say It: Identifying Appropriate Register for Chatbot Language Design

Ana Paula Chaves
anachaves@utfpr.edu.br

Northern Arizona University, USA
Fed. University of Technology–Paraná, Brazil

Eck Doerry, Jesse Egbert, Marco Gerosa
{Eck.Doerry,Jesse.Egbert,Marco.Gerosa}@nau.edu
Northern Arizona University, USA
Flagstaff, AZ, USA

ABSTRACT

Designing chatbots that produce language that is natural and appropriate to a given context is critical in satisfying user expectations. Currently, little is known about how a chatbot's linguistic choices should be designed to conform with the language humans produce in similar contexts. In this paper, we draw on existing sociolinguistic theory to adapt a technique called *register analysis* to (a) characterize the linguistic register used by humans in a specific conversational context; and (b) drive chatbot language design. Our exploratory study investigates the application of register analysis for tourist assistants chatbots and shows how the results could be used to develop them to adopt the appropriate register.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction design*.

KEYWORDS

chatbots; register; conversational style; language design

ACM Reference Format:

Ana Paula Chaves and Eck Doerry, Jesse Egbert, 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 (HAI '19)*, October 6–10, 2019, Kyoto, Japan. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3349537.3351901>

1 INTRODUCTION

Advances in conversational interfaces have helped to popularize chatbots [38], which are disembodied conversational interfaces that interact with end-users in natural language via text interfaces. The number of currently available chatbots attests to its growing popularity. In 2018, Facebook announced that its Messenger platform had been used to deploy 300,000 chatbots [9], ranging from general purpose chatbots such as Microsoft Zo¹ to customer services representatives and shopping advisors such as UPS² and Sephora³

¹<https://www.messenger.com/t/zo>

²<https://www.messenger.com/t/TheUPSBot>

³<https://www.messenger.com/t/sephora>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

HAI '19, October 6–10, 2019, Kyoto, Japan

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6922-0/19/10...\$15.00

<https://doi.org/10.1145/3349537.3351901>

bots. Some researchers claim that many websites and apps with conventional query-oriented graphical user interfaces will change to conversational interfaces in the near future [17].

In chatbots, natural language conversation is the primary resource for achieving interactional goals. In the current generation of chatbots, utterances are often pre-defined sentences with arbitrary style that is completely independent of any particular conversational context. For example, most Facebook Messenger chatbots, such as UPS and Sephora, use a simple decision tree mechanism combined with visual elements to produce pre-defined responses. Although this approach may ultimately provide users with the answers they are seeking, previous studies have shown that when a chatbot use inconsistent or unexpected patterns of language (e.g., excessive (in)formality or incoherent style), the conversation sounds unnatural, leading to frustration [15, 23, 28]. For example, Duijst [15] observed that users found it strange that the financial advisor used emojis combined with formal language in a situation of urgency (a stolen bank card). As humans, we have developed a sense of how to adapt the tone, idioms, and formulations we use in our communications to various conversational contexts. What is needed is a better understanding of how this mechanism works, and how we might use it to improve the quality of chatbot conversations.

In early chatbot design, the focus was on ensuring that chatbots produce coherent responses and grammatically correct sentences [34]. Modern chatbots have made strides in these foundational areas (see, e.g., [12]), and now the competitive focus in chatbot design must shift to consider not only what is said, but also how it is expressed linguistically. Currently, there is no formal techniques for addressing this challenge; chatbot designers develop chatbots based on their personal linguistic habits or ad hoc analyses of user characteristics. Formal linguistics research has established the importance of tone, word choice, and formulation in conversational contexts: sociolinguistic theory states that humans' linguistic choices are not arbitrary, but tailored to accomplish a particular function within the context [13], a concept called *register*. According to this theory, different individuals within the same situational contexts will make similar linguistic choices, thus adapting the register to the particular characteristics of context and conversational partner. Using the correct register is critical to the success of the communicative purpose of the statement; incorrect register can impact credibility, trust, and overall perception of quality in the conversation [24, 39]. Thus, designing conversational engines for the next generation of chatbots will require strong models of how humans adapt their conversational register to match varying contexts, as well as techniques for efficiently leveraging such models in chatbot design.

This paper introduces a strategy for identifying the register used by humans and current chatbots within a given context. Tourism

was selected as a testbed domain because chatbots have been widely adopted to support travel decisions [19, 22, 35]. Our exploratory study introduces the technique of register analysis, then uses it to characterize the language variation expressed when a tourism specialist provides information through text-based instant messenger tools. Using a comparative approach, we collected utterances from human tourist assistants and tourist info chatbots within similar conversational contexts and performed a register analysis to characterize the language variation across interlocutors. Our results show the prevailing characteristics of tourist assistants' discourses, and what linguistic features determine their typical register. Our contributions are twofold: we introduce register analysis as a strategy for identifying and characterizing the variations in the register, and we draw on these results to provide a set of preliminary guidelines for ensuring appropriate register in the design of conversational engines for chatbots in the domain of tourist information search.

2 RELATED WORK

Although chatbot technology has only recently become robust enough to deploy in real-world contexts, there is already evidence that the register will be critical to chatbot acceptance; users expect that chatbots use language that is structurally correct and portrays a coherent style [15, 20, 23, 28]. In this section, we briefly review existing efforts to understand how user perceptions are influenced by the quality of discourse produced by chatbots.

2.1 Natural Language Interfaces for Chatbots

Recent advances in Natural Language Processing have contributed to the quality of chatbots' language [21]. Most current research on chatbot conversational skills focuses on improving intent recognition [1, 26], as well as grammar and sentence structure [34]. Only a few studies, however, address issues related to conversational style, investigating how patterns of language influence users' perceptions and behavior toward the chatbots [24, 31, 37].

The literature emphasizes the lack of consistency. Chatbots should display language that coheres with the expected identity [28, 29] and represented service [15, 20]. For example, in the customer services domain, [20] states that chatbots are expected to fulfil the role of a human representative and thus should produce language that conforms with this role. Mairesse and Walker [28] proposed a computational framework that demonstrates that personality can be expressed by language style. Although these studies show that style can communicate valuable information and has an influence on user perceptions, they do not provide frameworks for understanding how style influences perception or how one might choose an appropriate style for a particular context.

Other studies have explored conversational style and how it influences users' satisfaction [16, 37]. Elsholz et al. [16] compared interactions with chatbots that use modern English to those that use a Shakespearian language style. Users perceived the chatbot that used modern English style as efficient and easy to use, while the chatbots that used Shakespearian English were seen as fun to use. Although using a Shakespearian style might be appropriate for the use-case presented in the study (buying a ticket for a Shakespearian play), the communicative rationale for selecting such styles is less obvious; the Shakespearian style is unlikely to be appropriate for

general contexts. In sum, these studies show that style has an impact on user perceptions, but stop short of developing techniques for characterizing a particular style, or for matching particular styles to targeted contexts. In the next section, we formalize the notion of "conversational style" into a more concrete concept of register, introduce key elements of register theory, and motivate the use of register in the design of conversational engines for chatbots.

2.2 Register and language variation

In sociolinguistics, style is a set of linguistic variants that reflect aesthetic preferences, usually associated with particular authors or historical period [13] (e.g., Shakespearean vs modern English). Several studies have explored the challenges surrounding computational production of sentences that follow a particular style [24, 27, 31, 37], with explorations ranging from consistently mimicking the style of a particular character [27] to dynamically matching the style to the interlocutor [31]. Research on sociolinguistic emphasizes that "*core linguistic features like pronouns and verbs are functional, and, as a result, particular features are commonly used in association with the communicative purposes and situational context of texts*" [13]. Under this perspective, the linguistic choices that comprise the register of a statement serve not only the aesthetic purpose, but are systematically motivated by communicative functions that are associated with the situational contexts.

Register theory states that, for each interactional situation, there is a subset of norms and expectations for using language to accomplish various standard core communicative functions [13]. The literature in conversational agents shows clear evidence of the benefits of using register-appropriate language [15, 20], e.g., increasing the chatbots' believability [25, 30] and adaptability [20, 28]. Nevertheless, there has been no exploration to date of applying register theory in designing conversational engines for chatbots. In the remainder of this paper, we begin to address this shortcoming with a study that explores register analysis as a means of identifying and characterizing the conversational style of interactions within the domain of tourist information. This provides an important first step toward consciously choosing an optimal conversational style and tailoring the conversational engine of a chatbot to match the particular conversational contexts in which it is deployed.

3 METHOD

As our goal is to explore the register analysis as a means of identifying and characterizing the register of online tourist assistants, we derived the following research question to guide our analysis: *What linguistic features characterize the register of tourist assistants when providing tourist information online?* To answer this question, we applied register analysis to short question-and-answer conversations in the context of tourist information search. The register analysis consists of two main steps:

1. Situational analysis. The aim in this step is to place the register of the targeted situation within a broad taxonomy of situational features. We followed the situational analytical framework proposed by Biber and Conrad [13]. Then, we collected utterances from both humans and chatbots tourist assistants and compiled them into a corpus of utterances that we used as an input for the next step.

2. Register characterization. The aim here was to analyze the

linguistic features typically used in tourist assistants' discourse to characterize the register used in conversations with tourists. This included tagging and counting the linguistic features present in the utterances and interpreting them according to their function in the sentence. We detail these two steps in the following subsections.

3.1 Situational analysis

Our first step was to characterize the target situational context, presented in Table 1, according to the framework for situational analysis [13]. We collected two datasets that fit the situational parameters (Table 1). The first one consisted of a corpus of 155 conversations from human tourist assistants when providing tourism information through Facebook Messenger. The corpus was collected in Flagstaff, Arizona, USA, a highly touristic region that is near the Grand Canyon and other parks, monuments, and natural and recreational areas. The tourist assistants were three experienced professionals who worked in the Flagstaff Visitors Center. All were native English speakers and 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 and the other one was 35-44. Although they had more than four years of experience in providing touristic information, they had never professionally provided information through an online platform. Questions were collected from tourists visiting downtown Flagstaff and supplemented with questions about Flagstaff placed by tourists in websites such as Quora, Google Maps, and TripAdvisor.

The second dataset consisted of a corpus of 89 interactions with chatbots designed to perform the role of tourist assistants. We searched online catalogues such as the BotList (<https://botlist.co>), Slack App Directory (<https://slack.com/apps/category/At0MQP5BEF-bots>), and Skype Bots (<https://www.skype.com/en/features/bots>) to find travel-related chatbots. We selected only chatbots that fit the situational parameters in Table 1 and that were capable of responding in natural language rather than using visual elements and links to external pages. We found three chatbots: Octa (<https://www.octa.ai>), Prago (<https://pragobot.com>), and Golem (<https://www.praguevisitor.eu>). Octa is a chatbot intended for kids, providing kids-friendly information on a destination. For instance, Octa provides information about activities and restaurants popular with kids. A typical interaction with Octa would include choosing a city, selecting an attraction, and getting information about description of the attraction, costs, and hours of business. Both Prago and Golem are designed specifically to guide tourists through Prague (Czech Republic). Golem is a chatbot for an online travel magazine while Prago is presented in the character of a local who guides the tourists through the local's favorites activities and culture. The three chatbots' natural language was scripted, hard-coded rather than dynamically generated. For the sake of convenience, we identify interlocutors from the first and second datasets as humans (Guide1, Guide2, and Guide3) and chatbots (Prago, Golem, and Octa).

Table 1: Situational parameters description

Participants	Tourists and tour guides
Relationship among participants	Tourist and tour guide, where tour guides own the knowledge
Channel	Written, instant messaging tool
Production	Quasi-real-time vs. scripted; human-written
Setting	Private, shared time, virtually shared place
Communicative purpose	Information search
Topic	Tourism, local information

3.2 Register characterization

After collecting the raw conversations from both our human and chatbots tourist assistants, we filtered the conversations to keep the tourist assistants' sentences only, discarding the responses from tourists. Then, we tagged the remaining utterances using the Biber grammatical tagger [6] for 125+ linguistic features. This tagger has been used for many large-scale corpus investigations, including studies of register variation [4, 7, 14] and the Longman grammar of spoken and written English [8]. The tagger analyzes individual features (normalized per 1,000 words), as well as the aggregation of these features into a set of factors called "dimension scores" (see [3]). Ultimately, the tagger scores the linguistic features into five dimensions, which are described as follows [3]:

Dimension 1–Involvement is associated with the oral vs literate opposition, where high/positive scores indicate personal involvement, interactional and generalized content, while low/negative scores indicate informational density and exact informational content. Examples of features with positive weights in this factor are private verbs (e.g., "*think*", "*believe*"), that-deletion, first- and second-person pronouns, be as a main verb, WH-questions (e.g., "*where*", "*who*", "*why*") and clauses, final prepositions (e.g., "*what can I help you **with**?*"), amplifiers (e.g., "*very*", "*absolutely*"), emphatics (e.g., "*just*", "*most*"), among other. Features with negative weights include nouns, prepositions (e.g., "*at*", "*across*"), and attributive adjectives (e.g., "*parking lot*", "*playground equipment*").

Dimension 2–Narrative flow distinguishes narrative from non-narrative discourses, where high/positive scores indicate narrative and reconstruction of events while low/negative scores indicate descriptive or expository discourse. Linguistic features in this dimension are past tense verbs, third-person pronoun, perfect aspect verbs (e.g., "*have/has/had visited*"), public verbs (e.g., "*say*", "*assert*"), and present participial clauses (participial clauses where the subject is shared with the main clause, e.g., "*visiting the city, you see...*"), all of which have positive weight in the dimension score.

Dimension 3–Contextual reference is associated with the explicit vs situation-dependent reference opposition, where high/positive scores indicate a discourse that presents highly explicit and elaborated, endophoric reference, where utterances use precise references to previous ones, and common ground among interlocutors is not assumed. For example, the sentence "*The Grand Canyon, which is one of the seven natural wonders, ...*" implies that the tourist does not necessarily have previous knowledge about the Grand Canyon. Low/negative scores indicate exophoric, situation-dependent reference, which implies common ground among interlocutors. For example, the sentence "*The Grand Canyon is a must-see*" implies that the interlocutors share knowledge about the Grand Canyon. Examples of linguistic features with positive weight are WH relative clauses (e.g., "*which*", "*what*", "*who*", etc.), phrasal coordination (clauses combined by coordinating conjunctions such as "*and*", "*or*", "*but*"), and nominalizations (use of words that are not nouns as nouns, e.g., "*brewery*", "*equipment*"). The features with negative weights are adverbs (e.g., "*about*", "*probably*"), and time (e.g., "*afterwards*", "*again*") and place (e.g., "*abroad*", "*far*") adverbials.

Dimension 4–Persuasiveness focuses on the overt expression of persuasion, where positive scores indicate that persuasion is overtly marked, either for expressing the speaker's point of view

or assessing advisability, while negative scores indicate discourses with no opinions or arguments. The linguistic features associated with this dimension all have positive weight and include infinitives, predictive modals (e.g., “*would*,” “*will*”), and conditional subordination (“*if*,” “*unless*”).

Dimension 5–Formality distinguishes abstract from non-abstract information, where high/positive scores indicate informational discourse that is technical and formal while low/negative scores indicate non-technical, informal discourse. Key linguistic features for this dimension include conjuncts (two or more clauses linked together by connectors such as “*and*,” “*but*,” “*thus*,” etc.), agentless passives (e.g., “*the city was found...*”), past participial (e.g., “*have seen*,” “*has visited*”), and by-passives (e.g., “*built by Native Americans*”), all of which have positive weight in the dimension score.

After generating the dimensional scores, we applied a one-way permutation multivariate analysis method (perMANOVA) to generate a statistical comparison of dimensional scores spanning each of the human and chatbot interlocutors; the values of the dimension scores were the dependent variables, and the six interlocutors were the independent variable. Each *text* corresponds to one observation in our permutation model, where a text is a set of one or more contiguous sentences produced by a tourist assistant (either human or chatbot). We used the perMANOVA to benefit from the distribution-free inferences achieved by permutation to account for assumptions violations [2]. The (pseudo) F-test statistic tests the location effect among groups, which is the equivalence of the positions of the group centroids in the space. All the F-statistics reported were calculated using $df_1 = 5$ and $df_2 = 288$ degrees of freedom. The specific hypotheses were:

H_0 : The position of the group centroids for the dimension scores do not differ per interlocutor vs

H_a : The position of the group centroids for dimension scores differ for one or more interlocutors.

Since our main goal was to identify the register of tourist assistants, we analyzed the commonalities among interlocutors, which is determined by the non-significant dimensions and associated features. Given a significant overall perMANOVA test, we performed pairwise comparisons (with False Discovery Rate–FDR’s correction for multiple comparisons) to evaluate the contrasts between pairs of interlocutors. Both model and contrasts were calculated using 10K permutations. Based on this analysis, we then qualitatively interpreted the function of individual linguistic features to identify how they determine or contribute to the style used by the interlocutors.

4 RESULTS

We begin our results with an overview of our multivariate analysis of the dimension scores. We then focus on each dimension, highlighting in which aspects the register of the interlocutors is either consistent or varying. For each dimension, we discuss the prevailing register characteristics and point out the linguistic features that determine the prevailing characteristics for each dimension.

4.1 Multidimensional analysis: an overview

Table 2 shows the number of texts, the mean and standard deviation for each dimension score as well as the label of the prevailing register characteristic per interlocutor. The perMANOVA analysis

Table 2: Descriptive statistics and the label for the prevailing register characteristic per interlocutor

	#texts		Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Guide1	53	Mean (μ): Std dev (σ): Prevailing reg. charac:	12.75 18.75 Involved	-4.42 1.5 Non-narrative	-3.73 5.6 Situation-dependent	1.76 7.08 Moderately Persuasive	-0.72 3.33 Moderately concrete
Guide2	45	Mean (μ): Std dev (σ): Prevailing reg. charac:	3.12 21.19 Moderately involved	-4.13 1.84 Non-narrative	-2.89 7.06 Situation-dependent	-0.09 7.5 Moderately non-persuasive	-1.09 5.27 Moderately concrete
Guide3	50	Mean (μ): Std dev (σ): Prevailing reg. charac:	-4.93 15.57 Informational	-4.55 1.86 Non-narrative	-3.36 6.94 Situation-dependent	2.18 10.2 Persuasive	-2.03 3.18 Concrete
Octa	37	Mean (μ): Std dev (σ): Prevailing reg. charac:	-3.45 21.35 Informational	-5.05 1.22 Non-narrative	-1.72 5.65 Situation-dependent	-2.94 5.81 Non-persuasive	0.18 6.38 Moderately abstract
Prago	28	Mean (μ): Std dev (σ): Prevailing reg. charac:	12.23 15.28 Involved	-4.54 1.55 Non-narrative	-3.65 5.19 Situation-dependent	-1.3 4.24 Moderately non-persuasive	1.36 6.1 Moderately abstract
Golem	24	Mean (μ): Std dev (σ): Prevailing reg. charac:	12.13 23.95 Involved	-4.7 1.44 Non-narrative	-11.52 13.79 Situation-dependent	-3.27 4.05 non-persuasive	-1.91 4.07 Moderately concrete

showed that the register of tourist assistants differs significantly ($F=5.83$, $p<0.00001$) across interlocutors.

The interlocutors’ style is descriptive and situational-dependent, as might be expected since the linguistic function is to present information using references to the physical and temporal situation of the discourse. Human tourist assistants show a trend toward being opinionated, while chatbots use less persuasive language. Most interlocutors tend to use more concrete discourse and only two tend to use more technical language. Although the overall purpose was to provide information, there is a lack of pattern regarding the personal involvement expressed in the interlocutors discourse. In the next sections, we discuss each dimension, highlighting the linguistic features that determine the register adopted by the interlocutors, along with excerpts to illustrate these observations.

4.2 Dimension 1: involvement vs informational

Our analysis revealed no patterns regarding the personal involvement in the tourist assistants’ discourse. For example, in Listings 1 and 2, the tourist asks for information about things to do. In response, Guide2 adopts a non-opinionated strategy and provided a highly objective response: she only lists possible toddler friendly activities. In contrast, Guide1 expresses personal (“*I just double checked*”) and active language since “*in between seasons*” suggests that the exact information is that no activities were available at that time. In that case, purely informational content could discourage the tourist from visiting the attraction in the near future.

Listing 1: Guide2, conversation3.15

[Tourist:] What are the best toddler friendly things to do in Flagstaff?
[Guide2:] There is good playground equipment at Thorpe Park and Bushmaster Park. A short easy hike at Buffalo Park also has activity centers. There are kids’ activities at Heritage Square on Fridays and Saturdays starting at 4:30 (movies on Saturdays after the sun sets).

Listing 2: Guide1, conversation3.i41

[Tourist:] Hi
[Guide1:] Hello. What can I help you with?
[Tourist:] What activities are available at Snowbowl?
[Guide1:] Right now, they’re in between seasons at the Snowbowl, but I just double checked their website and the lifts for skiing and snowboarding will open this coming Friday.

Two main aspects may explain the different levels of personal involvement: the sub-types of questions, and the adopted strategies to convey the information. The influence of the sub-types of

questions can be observed in the Listings 3 and 5. In Listing 3, the tourist asks for exact information (opening hours and tickets cost), which resulted in a highly informational response. In contrast, the tourist in Listing 5 asks for a recommendation (“*should I go*,” “*my best choice*”), which resulted in an involved response (“*I would say*”).

Listing 3: Octa, conversation11

```
[Tourist:] NEXT! >>
[Octa:] $image Ferry Building
One San Francisco Bay Trail, San Francisco, CA 94111, USA
[Tourist:] Why is this epic?
[Octa:] One of the symbols of the city is the iconic Ferry Building. $emoji
While the building was once the gateway to San Francisco, today it is a
bustling market filled with foodie-favorite stalls. A fun farmers' market
takes place out back on Tuesdays, Thursdays and Saturdays. $emoji $emoji
[Tourist:] Tickets
[Octa:] Free to visit
[Tourist:] Opening hours
[Octa:] These are the opening hours for the week: [...]
```

The chatbots produced nearly all of their utterances as either recommendation (Prago and Golem) or information (Octa), regardless of the request. For example, Prago typically answers tourists' requests with a personal tone (e.g., “*my tips*” in Listing 4). Human interlocutors, in contrast, tended to use personal tone more consistently when the tourist asked for recommendations. For example, although Guide 2 is the least personally involved human interlocutor, she used personal tone in Listing 5 (“*I would say*”), due to the nature of the request (“*my best choice*”). As a consequence, the level of involvement vs informational density significantly varies across interlocutors (see 1a), as the humans adapted their register to the context. The difference in the mean scores is significant at any reasonable significance level ($F=7.27$, $p<0.00001$). In Section 5, we discuss the implications of this variability.

To gain deeper insight into how the scores on Dimension1 are driven by specific phrasing decisions, we performed a detailed analysis of individual linguistic features. The analysis showed that utterances with greater informational density are marked with frequent occurrences of context-specific nouns (e.g., “*symbols*,” “*city*,” “*gateway*,” “*market*,” etc.) and attributive adjectives (“*iconic ferry building*,” “*bustling market*,” “*short easy hike*,” etc.), and infrequent occurrences of verbs, first and second person pronouns, contractions, and WH-questions. Prepositional phrases are also common to indicate accurate information, such as place (“*at Buffalo Park*”) and date/time (“*at 4:30*,” “*for the week*”). In contrast, utterances with

greater personal involvement are marked by features that characterize verbal discourse, namely the occurrence of contraction, first- and second-person pronoun, final preposition, private verbs (“*I think*”), demonstrative pronouns, and emphatics (e.g., “*What can I help you with?*,” “*I just double checked*,” Listing 2). Interestingly, chatbots used significantly more private verbs than human interlocutors. Additionally, amplifiers are not a common feature, with only a few occurrences for the most involved interlocutors (Guide1, Golem, and Prago); this contradicted our expectation that amplifiers would appear as a means to advertise the attractions.

Listing 4: Prago, conversation7 (excerpt)

```
[Tourist:] Where else to go
[Prago:] Here are my tips.
[Tourist:] Stromovka
[Prago:] This used to be a closed reserve but now it is a park where you can go
and have a barbecue.
Well, you need to bring your own stuff, you can't hunt there anymore. And I
think just rabbits live there, no deers anymore.
But you can find there also garden pubs, no worries. $link
```

Table 3 summarizes the linguistic features that characterize the register of tourist assistants regarding the level of involvement vs informational density. The column labeled as “Consistently present” lists linguistic features that appear in more than 75% of the texts; “Consistently absent” lists features that are absent in more than 75% of the texts; and “No clear tendency” lists features neither present nor absent (the counts are equal to 0.0 for 25–75% of the data). For Dimension 1, these features determined the internal variability.

Listing 5: Guide3, conversation2.10 (excerpt)

```
[Tourist:] Should I go to the Grand Canyon or Sedona? What is my best choice?
[Guide2:] I would say the Grand Canyon, it's iconic and on many `bucket lists
`. Sedona is closer however (30 minutes vs. 1.5 hours), so if you are
pressed for time, I would say that the red rocks of Sedona are definitely
worth seeing.
```

Listing 6: Golem, conversation24(excerpt)

```
[Tourist:] Czech restaurants
[Golem:] Here are my favorite Czech restaurants $emoji
[Tourist:] More...
[Golem:] Sure, what would you like?
[Tourist:] Tell me more
[Golem:] When people arrive to Prague, this is where we take them for their
first Czech food experience
Directly across from St Nicholas, the view is as amazing as the goulash, potato
dumplings, and sweet red cabbage
Great service and value, the perfect pit-stop on your way to or from the castle.
```

Table 3: Linguistic features for Dimension 1

Dimension 1		
Consistently present	No clear tendency	Consistently absent
Present verbs Second-person pronoun Nouns (greater for Octa) Prepositions (greater for Guide 2) Attributive adjective	Private verbs (greater for chatbots) Contractions (greater for chatbots-Golem) Emphatics (greater for the most involved: Guide1, Golem, Prago) First-person pronoun (greater for the most involved: Guide1, Golem, Prago) Pronoun “it” (greater for Prago) Possibility modal (lower for Guide 3 and Octa) Coordinating conjunction (greater for Guide1, Guide2, Prago)	That-deletions (greater for Prago) Do as pro-verb Demonstrative pronoun Be as main verb Causative subordination Discourse particle (greater for Guide1) Nominal pronoun (greater for Guide1, Octa, Prago) Hedge Amplifiers (greater for the most involved: Guide1, Golem, Prago) WH questions WH clause Final preposition (greater for Guide 1)

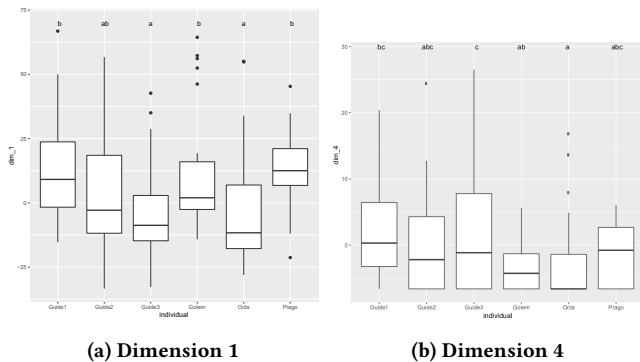


Figure 1: Mean scores per interlocutor. Contrasts significant at $p<0.05$

4.3 Dimension 2 and Dimension 3: narrative and references

In Dimension 2 (narrative vs non-narrative) and Dimension 3 (explicit vs situational dependent references), interlocutors consistently present language that is both expository and situation-dependent.

Table 4: Linguistic features for Dimensions 2 and 3

Dimension 2		
Consistently present	No clear tendency	Consistently absent
-	Third-person pronoun (greater for humans) Past tense verbs	Perfect aspect verbs Public verbs
Dimension 3		
Consistently present	No clear tendency	Consistently absent
Nominalizations Adverbs (greater for Guide1, Guide2, Prago)	Place adverbials (greater for Golem)	WH relative clause–object WH relative clause–pred piping WH relative clause–subject Phrasal coordination–subject Time adverbials

The expository register shows that the exchanged information was focused on immediate actions rather than past events, while the situation-dependent references highlight the shared context among tourists and assistants and their common ground with respect to the information being provided. All interlocutors have a low mean score (approx. -4.5) for Dimension 2, and there is no significant difference among them ($F=1.37$, $p=0.23$). For Dimension 3, only Golem is significantly different from other interlocutors ($F=6.04$, $p<0.0001$), but the mean score is even more extreme to the left with greater standard deviation, as Table 2 indicates.

In analyzing the detailed linguistic features, we found that discourse is characterized by few occurrences of third-person pronouns and public verbs, as well as few occurrences of verbs overall. When verbs occur, they are mostly in the infinitive or present tense, e.g., Octa’s answer to “*Why it is epic?*” in Listing 3 involves only five verbs out of 46 words, three of which are in present tense. The discourse thus shows no narrative features (Dimension 2). Human interlocutors tended to use third-person pronouns more often than chatbots, although the difference is only marginally significant. In Dimension 3, nominalization does occur (e.g., “*centers*,” “*farmers*”), but other features with positive weight for this dimension (WH-clauses and phrasal coordination) are consistently absent. Adverbs, on the other hand, are very frequent (e.g., “*also*,” “*often*,” “*definitely*”) as well as place adverbials (e.g., “*across*,” “*inside*,” “*away*”), and these both have a negative weight for Dimension 3. Table 4 summarizes the linguistic features observed in Dimensions 2 and 3.

4.4 Dimension 4: persuasiveness

Although persuasion is not strongly marked for any interlocutor (highest mean score=2.18⁴), humans tended to use persuasive and opinionated language slightly more often than chatbots, as Figure 1b suggests. As pointed out earlier, human interlocutors tend to use involved, opinion-oriented discourse when tourists ask for explicit recommendations. For example, Guide2 is the least persuasive among human interlocutors, but Listing 5 shows the use of opinionated language, i.e., first-person pronoun along with a predictive modal to indicate the personal preference (“*I would say the Grand Canyon*”), as well as a conditional clause along with predictive modal to convey an alternative (“*if you are pressed for time, I would say...*”). In contrast, the utterance about the attraction in Listing 4 is not marked with persuasive features, although it is also highly involved. The linguistic feature that most characterizes persuasion is prediction modals, which co-occur with first-person pronouns (“*I would*”)–found mostly in humans’ utterances–and second-person pronouns (“*you would*,” “*you will*”). Conditional subordination also co-occurs with second-person pronouns (“*if you*”). Verbs in the

⁴In Biber [3], the mean scores for Dimension 4 for the registers with an overt expression of persuasion are above 3.

Table 5: Linguistic features for Dimension 4

Dimension 4		
Consistently present	No clear tendency	Consistently absent
-	Infinitives (greater for Octa and Prago) Prediction modals (greater for humans)	Suasive verbs Split auxiliaries Necessity modals Conditional subordination

Table 6: Linguistic features for Dimension 5

Dimension 5		
Consistently present	No clear tendency	Consistently absent
-	-	Conjuncts, Agentless passive By-passive Passive postnominal modifier Subordinating conjunction (e.g., <i>as</i> , <i>except</i> , <i>until</i>)

infinitive form are also common, but unlike the other features, they are generally used across interlocutors; the most common infinitive verbs are “*to do*,” “*to see*,” “*to visit*,” “*to go*,” and “*to get*.” Table 5 summarizes the linguistic features that characterize the register with respect to Dimension 4.

4.5 Dimension 5: Formality

We found that discourse showed no clear tendency with respect to formality in the discourse (Dimension 5). Although providing informational content, interlocutors did not focus heavily on providing exact, technical information, thereby toning down the formality in their language. Nonetheless, the register remains more formal than casual daily conversations, for example, which would display extreme negative scores for this dimension. Interlocutors used varying vocabulary as well as active verbs that emphasize the subject, which is often a noun or a third-person pronoun. The topic of the conversations varies from the abstract (e.g., addresses, costs, views, etc.) to concrete references (e.g., activities, places, food, etc.).

Overall, the linguistic features that define Dimension 5 (e.g., passives, conjunctions, past participial clauses) are generally rare, and the difference among interlocutors is only marginally significant ($F=2.47$, $p=0.03$). The differences that do exist are driven by the agentless passive verb and the conjunctive adverbs, which are slightly more frequent for particular interlocutors. Chatbots tend to use more conjunctive adverbs than the human interlocutors; Prago shows the largest difference, mainly using “*Okay*” and “*Well*” to add a casual verbal tone. Guide1 and Octa used agentless passive verb (e.g., “*are/is located on*,” “*is considered*”) more often than other interlocutors, although the difference is not significant. When the agentless passive and past participle occur, the two most common functions are adding information regarding a place/attraction/activity of interest, and tailoring the recommendation for the tourist’s possible preference. For example, in Listing 5, Guide2 uses past participle along with conditional clause (“*so if you are pressed for time*”) to give options to the tourist regarding the best place to go. Table 6 summarizes the linguistic features that characterized the register with respect to Dimension 5.

In summary, the register of online tourist assistants is characterized by non-narrative, situation-dependent, moderately-concrete discourse. It varies according to the sub-registers in terms of personal involvement and persuasion. The sub-registers that stood out from our corpus were recommendation and information requests.

5 DISCUSSION

In this section, we discuss our findings and their implications for the design of chatbots that have their register consciously tailored to specific contexts. To show the practical value of register analysis

in shaping the design of future generations of chatbots, we also translate these observations into some preliminary guidelines for designing tourist assistants chatbots. The idea of generating situationally appropriate language does not imply deceiving people into thinking the software is human. Overly humanizing artificial agents can create the wrong expectation about the agents' responsiveness [18, 20]. Nevertheless, as different speaker roles can prompt notable variance in speaking habits [11], users unconsciously establish a coherence between chatbots' utterances and a plausible utterance that the people whom the agent represents might form in that context [20]. By approximating chatbots' language to the register established by the context and expected by the users, we aim at reducing expectation breakdowns.

The impact of sub-registers: the internal variability in Dimensions 1 and 4 may be explained by the existence of sub-registers in the corpus. Sub-registers express the variation in language use to match specific communicative purpose. For example, the questions "*When was this city was founded?*" and "*Which trail provides the best scenic view?*" calls for different levels of involvement since the former implicitly asks for a narrative story, while the latter asks for a recommendation. We also noticed that the interpretation of the purpose of tourists' questions influences the language variation. Prago chatbot, for example, consistently uses the sentence "*Here is my list of recommendations,*" which suggest that tourists' questions are rigidly interpreted as desiring a personalized recommendation rather than an information request. In traditional register analysis, each text is interpreted as belonging to a single register [33, 36]. However, our results showed that interactive conversations in tourism domain might involve several purposes, which impact the patterns of language. Hence, additional investigation will be needed to understand the impact of sub-registers on register variability and to develop techniques for adapting register to dynamic variations of communicative purpose within conversations.

Recommendation vs information: when observing the register variability, the sub-registers that stood out from our data were recommendation vs information requests. Interestingly, the difference was primarily evident for the human interlocutors, who could fluidly adapt their responses to the evolving perceived purpose of the conversation. For chatbots that are built around pre-recorded responses, the designer is required to make a static *a priori* decision as to whether the chatbot's utterances should sound like a recommendation or a neutral information. Humans, on the other hand, build their utterances on the fly, allowing them to continually judge whether to use informational tone or an opinionated response. Even for chatbots that apply generative models [32] to dynamically construct responses, implementing this decision may be challenging; at very least, more research will be needed to identify subtle cues that humans use to determine when shifts in the register. While current chatbot design may not allow adapting utterances to the proper register, guidelines G1 and G2 in Table 7 show how utterances might be designed to convey a recommendation or informational tone. For informational utterances, the language should emphasize nouns, adverbs, and prepositions while avoiding contractions, final prepositions, passives, conjunctive adverbials and past participial clauses. In case of recommendation, the language should emphasize first- and second-person pronouns as well as prediction modals.

Table 7: Preliminary guidelines for designing tourist assistants chatbots

G1	prefer literal language for information search only when the goal is information search only, the resulting language should be literal, unambiguous and accurate although formality should be smoothed
G2	prefer involved language for recommendations design the chatbot to convey personal tone and involvement (e.g., " <i>I would visit,</i> " " <i>you will enjoy</i> ")
G3	provide options when users' preferences are unknown design the chatbot to use conditional clauses as well as involved language to provide options to the users (e.g., " <i>If you..., you can/will/would</i> "). When interactivity is appropriate, then WH-questions may be used as well.
G4	be polite prefer active, emphatic, and involved language when face-threatening utterances arise

On the difficulty of dynamically tailoring recommendations: when human tourist assistants were asked about how natural they thought it was to provide online information compared to their daily work, they unanimously stated that they missed visual cues to tailor their responses. Curiously, none of them asked "compensatory" profiling questions that might fill some gaps; the human interlocutors only asked the preferred type of food or activities. In Table 7, guideline G3 describes how chatbot designers might provide flexible, acceptable answers to the tourists without asking distracting and potentially overly personal profiling questions.

Politeness: politeness is the ability to use communicative strategies to maintain social harmony [10]. According to the politeness theory [10], interlocutors use politeness to avoid Face-Threatening Acts (FTA), which, on its turn, is a speech act that threaten, either positive or negatively, the "face", or self-image, of an interlocutor [10]. In our study, interlocutors tend to use involved and active language when the content may be face-threatening. For example, in Listing 2, admitting that there were no ongoing activities in the attraction might negatively threaten the tourist's face. To counteract, Guide1 proactively checked when activities would be available. The resulting utterance was active and personal ("*I just double checked their website*") and softens the response by suggesting that anyone could have made this mistake, i.e., even the tourist assistant had to "*double-check*". Preliminary research has shown evidence that making interactions more personal may express politeness [23]. However, identifying FTAs is challenging and further research is needed to address FTAs identification and acts. In Table 7, guideline G4 describes how to convey politeness, which can be achieved by using first-person pronoun, active verbs and emphatics.

6 LIMITATIONS

A potential threat to the validity is the interlocutors' representativeness. All the human interlocutors were female from a specific place. For the chatbots, although they fit to the selection criteria, they target varying audience. These aspects may introduce bias based on differences in characteristics of language use. However, previous research suggests that register is the strongest predictor of language variation [5]; thus, the influence of these factors would be mainly stylistic, and the register analysis can still provide powerful conclusions. We acknowledge, however, that further investigation with larger and more diverse pool of subjects would be useful.

The permutation model assumes independence among observations. Texts either within or across human interlocutors may include repetitive questions and similar responses. However, the effect of such violations should be minimal, since every human's utterances were produced on-the-fly, and hence, they are unique. In any case, the method applied here is considered best practice and is commonly used in register studies [3] in sociolinguistics.

7 CONCLUSIONS

Large effort has been invested in improving the ability of chatbots to interact in natural language with human users. As the popularity of chatbots increases, however, end-users will expect chatbot conversations that also use a register that is appropriate to the conversational context. To date, there has been research showing that register has impact on credibility, trust, and persuasiveness, but chatbot designers have had little or no concrete guidance on how to think rationally about the issue.

In this paper, we explored how register analysis, a concept from sociolinguistics, might be used to address this shortcoming, by providing a framework for characterizing linguistic register, as well as providing a basis for identifying its impact on user perceptions. The results show how the register analysis can characterize the patterns of language for particular contexts, and reveal distinct linguistic characteristics within conversations in the tourism domain. The analysis of these characteristics revealed not only variations and similarities between interlocutors, but also within conversations. We noted the importance of sub-registers as a mechanism used by humans to accomplish dynamic adaptation, and the inability of chatbots to follow suit. Clearly, the ability to track conversational purpose and adapt register is an important feature of natural human conversation, at least in the studied domain. Our future research will make use of our results to drive development of a practical framework to analyze and classify target situations, and then develop chatbot conversational engines both tailored to the target context and able to adapt register to the specific communicative purpose. Additionally, we plan on performing a side-by-side evaluation on the collected utterances when the tourist assistants answered the exact same question. In the meantime, we have provided a set of preliminary guidelines that designers might consider to improving the naturalness of language in the current chatbots.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No.: 1815503.

REFERENCES

- [1] Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat Detection in an Intelligent Assistant: Combining Task-oriented and Non-task-oriented Spoken Dialogue Systems. *CoRR* abs/1705.00746 (2017).
- [2] Marti J Anderson. 2014. Permutational multivariate analysis of variance (PERMANOVA). *Wiley statsref: statistics reference online* (2014), 1–15.
- [3] Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.
- [4] Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- [5] Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8, 1 (2012), 9–37.
- [6] Douglas Biber. 2017. MAT - Multidimensional Analysis Tagger. Available at: <https://goo.gl/u7h9gb>.
- [7] Douglas Biber and Jesse Egbert. 2016. Using Multi-Dimensional Analysis to Study Register Variation on the Searchable Web. *Corpus Linguistics Research* 2 (2016), 1–23.
- [8] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman grammar of spoken and written English*. Vol. 2. MIT Press Cambridge, MA.
- [9] Marion Boiteux. 2019. Messenger at F8 2018. Retrieved October 18, 2019 from <https://bit.ly/2zXVnPH>. Messenger Developer Blog.
- [10] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.
- [11] Ta-Chung Chi, Po-Chun Chen, Shang-Yu Su, and Yun-Nung Chen. 2017. Speaker Role Contextual Modeling for Language Understanding and Dialogue Policy Learning. *arXiv preprint arXiv:1710.00164* (2017).
- [12] David Coniam. 2008. Evaluating the language resources of chatbots for their potential in English as a second language learning. *ReCALL* 20, 1 (2008), 99–117.
- [13] Susan Conrad and Douglas Biber. 2009. *Register, genre, and style*. Cambridge University Press, New York, NY, USA.
- [14] Susan Conrad and Douglas Biber. 2014. *Multi-dimensional Studies of Register Variation in English*. Routledge.
- [15] Daniëlle Duijst. 2017. *Can we Improve the User Experience of Chatbots with Personalisation*. Master's thesis. University of Amsterdam.
- [16] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement. In *Conference on Human Information Interaction and Retrieval*. ACM, 301–305.
- [17] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.
- [18] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *International Conf. on Internet Science*. Springer, 194–208.
- [19] Piedad Garrido, Javier Barrachina, Francisco J Martinez, and Francisco J Seron. 2017. Smart tourist information points by combining agents, semantics and AI techniques. *Computer Science and Information Systems* 14, 1 (2017), 1–23.
- [20] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *International Conference on Information Systems*. 13.
- [21] Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science* 349, 6245 (2015), 261–266.
- [22] 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 *INVTUR Conference*.
- [23] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*. ACM, 895–906.
- [24] 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.
- [25] Marie-Claire Jenkins, Richard Churchill, Stephen Cox, and Dan Smith. 2007. Analysis of user interaction with service oriented chatbot systems. In *International Conference on Human-Computer Interaction*, Julie A. Jacko (Ed.). 76–83.
- [26] Ridong Jiang and Rafael E Banchs. 2017. Towards Improving the Performance of Chat Oriented Dialogue System. In *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 23–26.
- [27] Grace I Lin and Marilyn A Walker. 2017. Stylistic Variation in Television Dialogue for Natural Language Generation. *EMNLP 2017* (2017), 85.
- [28] François Mairesse and Marilyn A Walker. 2009. Can Conversational Agents Express Big Five Personality Traits through Language?: Evaluating a Psychologically-Informed Language Generator. Cambridge University Engineering Department & Department of Computer Science, University of Sheffield.
- [29] Thomas W. Morris. 2002. Conversational agents for game-like virtual environments. In *Artificial Intelligence and Interactive Entertainment*. AAAI, 82–86.
- [30] Kellie Morrissey and Jurek Kirakowski. 2013. 'Realness' in Chatbots: Establishing Quantifiable Criteria. In *International Conference on Human-Computer Interaction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 87–96.
- [31] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *J Lang Soc Psychol*. 21, 4 (2002), 337–360.
- [32] Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K Chandrasekaran. 2017. A Survey of Design Techniques for Conversational Agents. In *International Conference on Information, Communication and Computing Technology*. 336–350.
- [33] Georg Rehm, Marina Santini, Alexander Mehler, Pavel Braslavski, Rüdiger Gleim, Andrea Stubbe, Svetlana Symonenko, Mirko Tavasani, and Vedrana Vidulin. 2008. Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems.. In *LREC*.
- [34] Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press, New York, NY, USA.
- [35] Albert V.D. Sano, Tanto D. Imanuel, Mega I. Calista, Hendro Nindito, and Andreas R. Condrobimo. 2018. The Application of AGNES Algorithm to Optimize Knowledge Base for Tourism Chatbot. In *2018 International Conference on Information Management and Technology*. IEEE, 1–9.
- [36] Marina Santini. 2004. *State-of-the-art on automatic genre identification*. techreport ITRI. University of Brighton, UK.
- [37] 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, 42–51.
- [38] MS Walgama and B Hettige. 2017. Chatbots: The next generation in computer interfacing—A Review. *KDU International Research Conference* (2017).
- [39] 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).