Talk to Me Intelligibly: Investigating An Answer Space to Match the User's Language in Visual Analysis

Jan-Frederik Kassel
Volkswagen Group
Munich, Germany
jan-frederik.kassel@volkswagen.de

Michael Rohs
Leibniz University Hannover
Hannover, Germany
michael.rohs@hci.uni-hannover.de

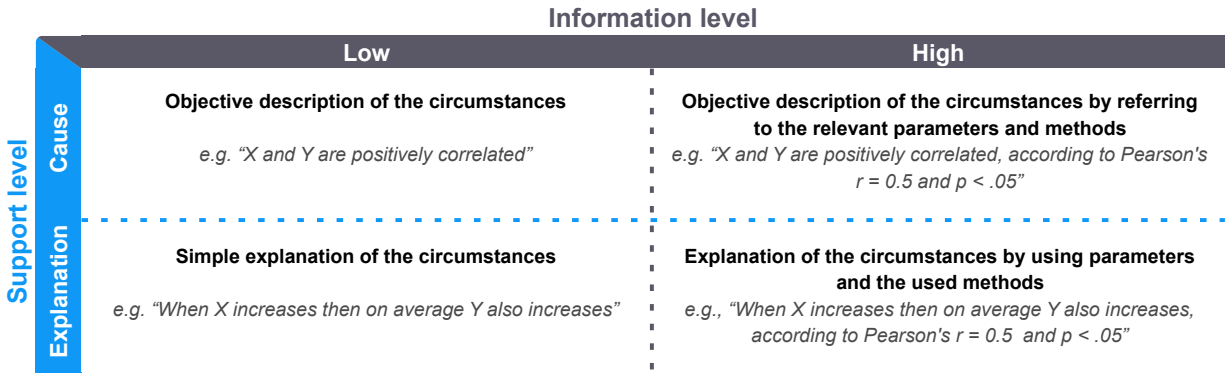

Figure 1: The two-dimensional answer space varies in the level of both information and support. The upper left cell comprises descriptions containing a statistical method's results in an objective manner. The lower right corner comprises explanations of a statistical method's results while additionally reporting the corresponding parameters. The example illustrates (grey) how the answer space implements the dependencies between two quantitative data attributes.

ABSTRACT

Conversational interfaces (CIs) have the potential to empower a broader spectrum of users to independently conduct visual analysis. Yet, recent approaches do not fully consider the user's characteristics. In particular, the objective of matching the user's language has been understudied in visual analysis. In order to close this gap, we introduce an *answer space* motivated by Grice's cooperative principle for framing personalized communication in complex data situations. We conducted both an online survey ($N = 76$) to analyze communication preferences and a qualitative experiment ($N = 10$) to investigate personalized conversations with an existing CI. In order to match the user's language properly, our results suggest to consider additional user characteristics along with their knowledge level. While mismatching communication preferences triggers negative reactions, a preference-aligned communication evokes positive reactions. As our analysis confirms the importance of matching the user's language in visual analysis, we provide design implications for future CIs.

DIS '19, June 23–28, 2019, San Diego, CA, USA

ISBN 978-1-4503-5850-7/19/06…$15.00

DOI: https://doi.org/10.1145/3322276.3322282

Author Keywords

Conversational interface; Visual analysis; Cooperative principle; Answer Space; Quantitative and Qualitative Analysis; Personalization; Conversational design

CCS Concepts

•**Human-centered computing** → **Natural language interfaces;** *User studies; Visualization systems and tools;*

INTRODUCTION

These days, the amount of data exponentially increases and so does the importance of data-related understanding. Visual analysis is one powerful approach to tackle this challenge. Yet, the focus is not any longer only in business [6], e.g., deriving valuable data insights for the company [25], but also in private life [8], e.g., challenging data-based conclusions presented in the media where the data is publicly available (see, e.g., [1]) [24]. Nevertheless, conducting visual analysis is a time consuming and complex task.

In order to increase the accessibility, and in conjunction with continuous improvements in natural language processing, conversational interfaces (CI) have been proposed for visual analysis [10, 13, 15, 21, 25, 26, 41, 44]. In this task-oriented scenario [42], a CI is essentially treated like a "virtual butler" [34]. Meaning that the user's expectation is not to have a human-like conversation, but more to have an easy-to-use system for fulfilling objectives [29]. Promising results suggest

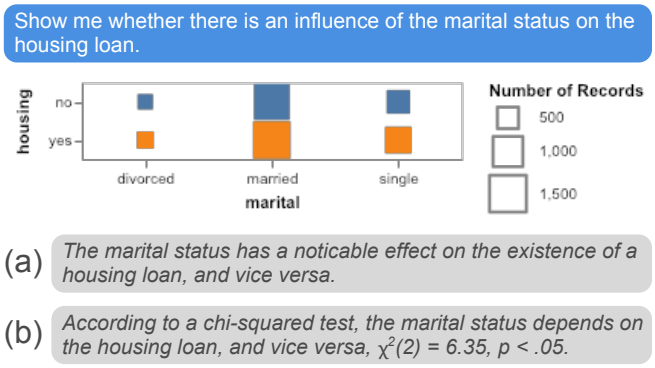

Figure 2: Assuming a user gives an utterance on the marital status and the existence of a housing loan (blue). Using the outcome of conducted χ^2-test in addition to the visualization helps to understand the data. However, how this information can be communicated both intelligibly and convincingly likely depends on the individual user.

that CIs do not only helo to accelerate visual analysis [13, 41], but also seem to lower the engagement boundaries for novice users [17]. Furthermore, the textual dialogue improves visual analysis by communicating supplementary information on the data [25] also referred to as "data facts" [43] to cover potential pitfalls [49].

Generally, such a dialogue consists of exchange sequences between parties [40] organized through "adjacency pairs" [40] of a user's utterance and the CI's response. The design of these sequences is important for a successful CI [20, 28]. This design directly affects a CI's usability and user experience [47]. Additionally, it influences the user's willingness to engage with the system [28]. In order to achieve a well-designed dialogue, Molich and Nielsen stressed the importance of matching the user's language [30]. Matching the user's language refers to using "words, phrases, and concepts familiar to the user" [30].

Human-human dialogues naturally follow this design recommendation, since a person unconsciously adjusts to their interlocutor's language [14, 18]. Consider, for instance, that the mathematical concept of derivation is to be explained to a student. A good teacher will surely adapt the explanation to the level of the student and provide different explanations depending on whether the student is a high school or a university student.

Although adjustment to the interlocutor is a necessary precondition for mutual understanding, this aspect has not been studied in detail in visual analysis. Little knowledge exists on how a CI should communicate facts about complex data situations in a user-specific way. The primary challenge for the CI is to put users in a position to draw valid conclusions from data sets of their interest. To achieve this the CI should dynamically adapt to the user's background and converse with the user both convincingly and intelligibly (see Figure 2).

In this paper we propose a novel framework (see Figure 1) – called *answer space* – based substantially on Grice's cooperative principle [18] to user-specifically communicate complex data facts as dialogue acts. This *answer space* conceptually supports users with diverse preferences and different knowledge levels in visual analysis. Through an extensive evaluation consisting of both an initial online survey ($N = 76$) and a controlled experiment ($N = 10$) with an existing CI enhanced by the *answer space*, we show the need for a personalized conversation in visual analysis as well as the usefulness of the *answer space*. Our statistical analysis reveals differences between users in how data facts should be communicated. We further show that adjusting to these preferences avoids negative impressions and is especially beneficial if a task exceeds the user's knowledge. Since dialogue acts are structuring the entire dialogue, our results provide novel insights on a user-specific conversation in visual analysis. Furthermore, we provide design implications for future approaches.

BACKGROUND

Over the last decades, research focused on investigating the conversation with intelligent systems and how to use CIs for visual analysis.

User-Specific Interactions with Conversational Interfaces

Shechtman and Horowitz [42] identified three main conversation styles between a user and a CI: task-oriented, communication-oriented, or relationship-oriented interaction. Aligned with these findings, Luger and Sellen [28] showed that regular CI users focus rather on simple tasks than complex tasks. For complex or sensitive tasks, users are lacking of trust in the CI. Additionally, Cowan et al. [9] confirmed the lack of trust in CIs for infrequent users as well. In terms of potential failures, Chen and Wang [7] showed that technical users think more about the reasons why a system failed. Nevertheless, Branigan et al. [4] generally highlighted that users tend to adapt their conversation style more when they believe to communicate with an intelligent system rather than with a human.

The user's background appears to have a considerable impact on the interaction with CIs in general. In contrast to previous work, however, we are investigating how the user's background particularly influences the communication preferences in visual analysis.

Conversational Interfaces for Visual Analysis

Intelligent agents help to reduce both work and information overload [29]. Additionally, using language as an interaction modality for visual analysis allows also novice users to analyze data [17]. As a result, conversational interfaces for visual analysis have recently been investigated more and more. These CIs hide the complexity of visual analysis from the user to lower the engagement boundaries. On the one hand, CIs like *DataTone* [15], *Valletto* [25], *Eviza* [41], *Aritculate2* [26], and *Evizeon* [21] are agnostic of the domain. The user can use these CIs for generating or adjusting visualizations for an arbitrary (structured) data set. As ambiguity in language is a challenge for CIs, reducing the ambiguity of user's utterances is proposed through either visual widgets [15, 21, 41] or dialogue sequences [10]. On the other hand, approaches like *Orko* [44] focus on visually exploring network visualizations through multitouch gestures combined with language.

Only a few approaches are considering the dialogue with user as an additional bidirectional communication channel. *Evizeon* [21] first uses pragmatics for allowing follow-up requests by the user in order to improve the dialogue flow in visual analysis. Additionally, a dialogue is used when the system's functionality is exceeded. Cox et al. [10] use the dialogue for clarification requests in case of ambiguity. *Orko* [44] verbally communicates its executed actions, e.g., "highlighted node X." Furthermore, the approaches of *Ava* [22] and *Iris* [13] compare themselves with Jupyter notebook [23]. They aim for reducing the complexity of data analysis by offering speech-executable functionality of scikit-learn [35]. Both systems implement state tracking for establishing a dialogue that helps creating the desired model step by step.

In contrast to our approach, all previously discussed approaches follow a generalized communication strategy. They do not dynamically adjust the conversation to an individual user, although matching the user's language is crucial for a successful CI [30].

Natural Language Generation in Visual Analysis

Natural language generation (NLG) is an important part of CIs. In visual analysis, however, only a few approaches exist that use NLG for textually describing the visualized data. Moreover, none of these approaches dynamically adapt their language to the user's background. Recently, *Voder* [43] investigated interactive "data facts" (textual descriptions of, e.g., correlation and distribution characteristics). They use an interactive widget for these data facts placed next to the visualization. *Voder* at its core is not a CI, however, its functionality can be seen as a response procedure of a CI for visual analysis. Furthermore, the commercial tool Tableau has an extension based on Quill [32]. This extension generates a summary for the shown dashboard. The idea of summarizing a complex situation is also applied in further domains, e.g., whether forecasts [37], or business process models [27].

As these automatic summaries are very convenient, they focus on key results (maximum, minimum, differences, etc.), but do not include advanced statistics. Although *Voder* [43] indeed reports e.g., whether two attributes are correlated, it does not show the corresponding ρ or the p-value. Since trust is an important factor in CIs [9, 28] communicating these parameters to those users who understand them might help establishing a trustworthy conversation [43]. On the other hand, bothering a statistically inexperienced user with those parameters will probably trigger confusion. Nevertheless, inexperienced users should be also made aware of the data relationships. Hence, CIs need a flexible framework in which they can communicate data facts in a way that is both intelligible and convincing.

To the best of our knowledge we are the first to investigate how to dynamically match the conversation needs of an individual user in CIs for visual analysis in order to properly support users with different backgrounds.

Conversation Theory

As in human-human dialogues [16], a CI should strive to make its dialogue optimally relevant for the addressee [30]. Generally, a fundamental theory for conversation analysis is the cooperative principle by Grice [18]. Grice devised how two persons should ideally communicate while cooperating. Grice precisely stated:

> "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." [18, p. 45]

He further derived normative guidelines consisting of four general maxims which have to hold for a cooperative conversation [18, p. 45 ff.]: (1) *Maxim of quantity*: make your contribution as informative as required, but not more than this; (2) *maxim of quality*: try to make your contribution one that is true; (3) *maxim of relation*: be relevant; and (4) *maxim of manner*: avoid ambiguity and be orderly. Violations of these maxims are called *implicatures* [18] and should be avoided in CIs for task-oriented scenarios. Otherwise a user might not be able to make sense of the dialogue act. However, previous approaches [10, 13, 15, 21, 25, 26, 41, 43, 44] do not consider the dialogue with the user under Grice's cooperative principle.

CONSTRUCTION OF THE ANSWER SPACE

CIs for visual analysis are typically answering to a user's utterance by generating visualizations [10, 15, 21, 25, 41, 44] and delivering additional information on the selected data through a dialogue [25, 44, 43]. These dialogue acts on data facts [43] help users to focus on the promising parts in the confirmatory analysis later on. However, differences in CIs exist between expert and novice [7, 9, 28]. Therefore, our primary objective is to provide a substantial construct which allows a CI to flexibly adjust dialogue acts in a user-specific way with an aim to avoid undesirable implicatures. In order to achieve this objective, we show how this construct is extracted from considering both Grice's cooperative principle and explainable user interfaces.

From psychology research we know that people perceive and communicate uncertainty differently [2, 5, 11, 38, 48]. Depending on user preferences, uncertainty should be communicated either numerically or verbally. However, these preferences additionally depend on whether a person trusts the person who is providing this information. In visual analysis through CIs, we hypothesize a similar effect.

According to Grice's *maxim of quantity*, a dialogue act should always contain the right amount of information. However, the question remains, how much is the right amount of information in the domain of visual analysis. For instance, that two attributes X and Y are positively correlated can be stated in multiple ways. One may, for example, say: "X and Y are positively correlated" (casually) or "X and Y are positively correlated according to Pearson's $\rho = 0.5$ and $p < .05$" (APA style). The difference is not only in numerical values, but also in the choice of statistical method and metric (e.g., Pearson's ρ) and supplementary parameters (e.g., p-value). Nevertheless, both statements are true, but how convincing and understandable they are exclusively depends on the user. In order to enable a user-specific communication and to avoid implicatures (violations of Grics's maxims) the model has to vary in information level.

Moreover, a CI may not only report a certain computed result, but may also actively help the user in interpreting it. In data analysis the terms *cause* and *causality* are often used in this context. *Cause* describes the circumstances, whereas *causality* is the conclusion based on the circumstances. However, *causality* usually requires domain knowledge to be sure that a conclusion is generally true in the domain. Since a CI should always communicate points for which evidence exists (see *maxim of quality*), we believe heading for statistical causality could be rather harmful for a user's conclusion process. Instead, we focus on the (causal) explanation [19]. Remembering the above example of correlation, the explanation would be "When X increases then on average Y also increases." Hence, explaining the consequence of correlated data supports users to draw valid conclusions, although they might not know the concept of a correlation coefficient. Consequently, our model should vary in support level as well.

Incorporating both dimensions into one framework leads to our two-dimensional answer space (see Figure 1). It provides a methodology to design dialogue acts on data facts in a way either descriptive with lower information (*LC*), descriptive with higher information (*HC*), explanatory with lower information (*LE*), or explanatory with higher information (*HE*).

Our initial objective is to identify coarse differences between users as well as to show how these differences should impact the conversation. Therefore, we initially model both dimensions as binary. Nevertheless, the information level might not be black or white, but rather a continuum. In order to analyze this assumption, we later discuss whether a binary categorization is already sufficient.

SITUATIONS IN DATA ANALYSIS

In order to effectively evaluate the answer space through actual user studies, we initially conducted unstructured interviews with three experienced data scientists (more than five years of professional work experience). The objective was to uncover a broad picture of a data scientist's typical work flow, rather than a complete analysis of all conceivable analysis situations. For framing the interviews, we used the following questions: *"How do you conduct an initial data analysis?"*, *"What are you focusing on?"*, and *"Which methods do you use?"*.

Dependencies

The first category is about dependencies between data attributes. The participants said that they are typically interested in finding first preliminary relationships in the data. Regarding pairwise relationships, the chosen measure depends on the characteristics of the data.

For two quantitative attributes, they compute a correlation coefficient. The Spearman's rank correlation coefficient is an appropriate measure, since it has fewer assumptions on the underlying data distribution, in contrast to Pearson's correlation coefficient. For describing the relationship between two categorical attributes, the χ^2-test of independence is selected. However, we add the normalized mutual information (NMI) [45] to the measures mentioned by the experts. NMI is based on entropy and can be applied to any discrete data without any assumptions on the level of measurement or the distribution. Therefore, it can be used to summarize relationships between two attributes X and Y with different levels of measurements: $NMI(X;Y) = \frac{H(X) - H(X|Y)}{\sqrt{H(X)H(Y)}}$ with $H(X) = -\sum_{\forall x \in X} p(x) \log p(x)$. As a result, we receive at least one measure that quantifies the relationship between two arbitrary data attributes.

Warnings

After finding relationships the next step of the participants' analyses is to identify differences within certain attributes. In this phase errors can arise to due method selection such as applying an unsuitable statistical test. Furthermore, there are also process errors such as the the multiple comparisons problem. Zgraggen et al. [49] show that this problem does exist in visual analysis. Potential false conclusions in this respect may cause critical consequences. This situation is hard to solve solely from the visualization. However, by computing a Kruskal-Wallis H test on all groups, followed by a Bonferroni correction on the pairs would reveal evidence.

Recommendations

In the search for meaningful attribute combinations an analyst often has to use a trial and error approach, as they mentioned. Therefore, we propose a recommendation of meaningful attribute combinations by the system. A useful measure to compute these potential recommendations is the NMI. Based on the visualized attribute the system computes all pairwise combinations with the remaining attributes of the data set [24]. By maximizing the NMI the top-k recommended attributes are those with the strongest relationship to the currently visualized one.

Interaction

In visual analysis through intelligent agents interaction with the data is an important aspect. Highlighting of certain areas in the visualization or applying filters on the underlying data via speech or gestures is supported by many systems [15, 25, 26, 41, 43, 44]. These interactions influence the user's decision process [49]. Highlighting and visual cues are shown in GUI widgets and are easily missed.

RESEARCH QUESTIONS

After deducing the answer space from conversation theory and showing relevant situations in the data analysis process, we investigate the utility of the answer space by a sequence of research questions:

Q1: What are the influencing factors for matching the user's language through the answer space?
Q2: Can the users' preferred communication style be accurately predicted by a probabilistic model?
Q3: How does the answer space work in practice? What are the reactions given an implicature vs. a user-specific answer?
Q4: Is the granularity of the answer space adequate?

ID	Situation	Method	Condition
In_1	Interaction	Filtering	one filter
In_2			two filters
Dq_1	Dependency	Spearman	positive
Dq_2			negative
Dc_1		χ^2-test	independent
Dc_2			dependent
Da_1		*NMI*	low
Da_2			high
Wa_1	Warning	H test & Bonf.	significant
Wa_2			$\neg$ significant
Re_1	Recommendation	*NMI*	categorical
Re_2			quantitative

Table 1: Concrete data analysis situations that occurred in the online survey.

USER STUDY 1: ONLINE SURVEY

The objective of the online survey is to identify preferences in the answer space depending on the participants' characteristics and a given data analysis situation (cf. **Q1**).

User Study Procedure

The online survey started with a randomly assigned sequence of 12 different concrete data analysis situations (see Table 1). Each situation – a concrete situation with a corresponding statistical method – was represented by an effective visualization (bar, line, or scatter plot) and two options from the answer space. A participant had to decide on the preferred option w.r.t. the visualization. After they made a decision, they received two new options from the answer space, however, the visualization remained the same. The participant's first choice – *preferences* in the answer space – determined the answer space options for the second decision. This means: if explanatory answers are shown, but one with low (*LE*) and the other with high information (*HE*), and the participant selects *LE*, then the explanatory answer with low information stays, but the other answer will be exchanged with a descriptive answer with low information (*LC*). The initial row and column of the answer space are randomly chosen. This procedure represents a greedy approach, which allows us to identify the participants' preferences within the dimensions (e.g., rating low over high information) and not only in the overall answer space. Therefore, we can derive further insights on the participants' preferences. Prior to the study, the questions and answers in the survey were reviewed for correctness by an experienced data scientist. Furthermore, this study design reduced the risk to overwhelm a participant, which could have been the case, had four options of the answer space been shown simultaneously.

After completing the sequence of actual situations, the participants were asked to state their experience with data analysis, chatbots, the methods used in the system (e.g., χ^2-test), and their conversation preferences. The questions were extended by a "honeypot" question in order to find unreliable answers. [33].

Actual situation (ID)	$\chi^2(9)$	p	*Cramer's V*
In	26.64	$< .01$	.24
Dc	8.92	$= .44$	-
Dq	18.11	$< .05$	.20
Da	22.30	$< .01$	.22
Wa	17.41	$< .05$	.20
Re	21.57	$< .05$	.22

Table 2: Results of the χ^2-tests between the self-reported knowledge and the preferences in the answer space.

Participants

We recruited both mathematicians and computer scientists at a university, as well as data scientists in an industry company. Additionally, we used Amazon's Mechanical Turk (AMT) restricted to participants with a US bachelor's degree (ensured by AMT), but not necessarily in a mathematical subject. We received 76 complete and reliable answers (87 in total), whereof 23 were from industry.

According to the self-reported statistics knowledge level the participants fall into 38% novices, 15% advanced beginners, 25% competent users, and 22% experts. Overall, the majority of the participants at least heard of the mathematical methods which our answer space supports and were used in the survey. Only the Bonferroni correction was new to 51% of the participants. On the question whether they would like to be guided in data analysis, 77% of the participants agreed or strongly agree to the idea of using a digital assistant. Furthermore, 75% considered recommendations of further attribute combinations as helpful. While control over on intelligent system is an important factor, we received no clear preferences on this aspect from the participants. 73% agreed that at least one of the given options fitted their preferences, in each situation of the survey.

Results

The following results on the participants' preferences in the different tasks are based on their second (final) decision, except when noted otherwise. Additionally, the analysis focuses on the participants' preferences on method level, since a conducted McNemar's test showed no significant differences between the conditions of each method. Generally, the following statistical analyses were conducted against $\alpha = .05$.

Our analysis first addresses the self-reported knowledge level of the participants as an influencing factor on their preferences. For each actual situation a χ^2-test of independence was separately performed. As Table 2 shows significant differences exist in almost every actual situation.

In order to better understand these differences, we conducted post-hoc tests with Šidák correction ($\alpha^* = .012$) for each significant actual situation. Regarding the information level, participants with little knowledge significantly prefer answers with lower information (*LC* or *LE*), whereas participants with more knowledge prefer those with high information (*HC* or *HE*) in both situations *Da* ($\chi^2(3) = 18.76$, $p < \alpha^*$) and *In* ($\chi^2(3) = 15.78$, $p < \alpha^*$). The statistical effect is large in *Da* (*Cramer's V* $= .35$) and medium in *In* (*Cramer's V* $=$

Figure 3: For each data fact and self-reported knowledge level the percentage-wise distribution (color saturation from 0 to 1) of the participants' preferences in the answer space is shown. The participants' preferences in the different data facts (*In* = filtering, *Dc* = dependency between categorical attributes, *Dq* = dependency between quantitative attributes, *Da* = dependency between attributes of arbitrary levels of measurement, *Wa* = multiple comparison problem, *Re* = recommendation of attributes to combine) are (*significantly) different, depending on the self-reported statistical knowledge.

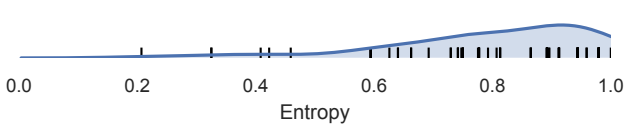

Figure 4: The participants' variation in their preferences depending on the individually computed entropy.

.32) respectively. Regarding the support level, participants with little knowledge chose with a chance of 3.25 (Fisher's exact test $p < \alpha^*$) explanatory over descriptive answers in *Wa* ($\chi^2(3) = 14.38$, $p < \alpha^*$, *Cramer's* $V = .31$)

Furthermore, the particular knowledge about the used mathematical method in the actual situations are influencing the participants' preferences. Participants with higher conceptual knowledge in Spearman's ρ significantly prefer answers with higher information (*HC* or *HE*), $\chi^2(3) = 13.62$, $p < .001$ with a large effect *Cramer's* $V = .31$.

In addition to the participants' knowledge, our analysis further shows significant differences in the participants' preferences based on the degree to which a participant wants to be guided in data analysis, $\chi^2(12) = 25.62$, $p < .05$, *Cramer's* $V = .10$. Furthermore, the participants' need for transparency on the system's computations significantly influences the preferences as well, $\chi^2(12) = 30.74$, $p < .01$, *Cramer's* $V = .11$.

By combining these individual participants' characteristics, further significant differences appear. On the one hand, answers with lower information (*LC* or *LE*) are significantly preferred by novice users ($\chi^2(4) = 13.50$, $p < .01$, *Cramer's* $V = .20$) and competent users ($\chi^2(4) = 14.52$, $p < .01$, *Cramer's* $V = .25$) when they care less about the used method (transparency). On the other hand, novice users ($\chi^2(4) = 24.52$, $p < .001$, *Cramer's* $V = .28$) as well as expert users ($\chi^2(4) = 15.47$, $p < .01$, *Cramer's* $V = .29$) significantly choose answers with higher information (*HC* or *HE*) when the would like to be in control of the entire analysis process.

Overall, the participants' preferences in the answer space are very different, according to the overall pairwise Hamming distance $\mu = .72$, $\sigma = .14$. Additionally, the individual preferences of the participants change from one to the other situation (see Figure 4).

Discussion

The online survey results show that participants' preferences in the answer space are only partially explained by the their knowledge. Although participants with little statistical knowledge tend to prefer explanatory answers with lower information, we do not see this general preference on other knowledge levels. Furthermore, the participants' preferences seem to depend on the given data situation too. Therefore, we argue that the preference in the answer space depends rather on the particular knowledge about the used statistical concept than how to communicate numbers.

Additionally, the self-reported level of needed transparency influences the preferences for dialogue acts with higher information. Since these dialogue acts comprise the used measure by the system as well as the corresponding parameters, it appears that transparency-oriented participants value these information. Finally, participants with a higher self-reported level of acceptance of a digital assistant are more likely to prefer an explanatory to a descriptive answer. A reason could be the general acceptance of intelligent systems and their use in the participants' daily life.

These results suggest to avoid not only a purely binary distinction between experts and novices, but also a generalized dialogue design in visual analysis (cf. Figure 3). In fact, a combination of all influencing factors should be considered to achieve a dialogue that is both convincing and intelligible (cf. **Q1**). Additionally, as the preferences seem to vary along the

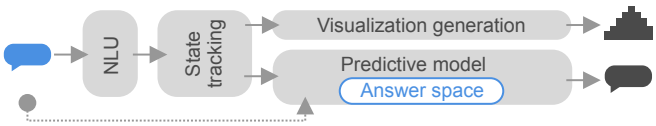

Figure 5: Conceptual embedding of the answer space into a conversational interface for visual analysis.

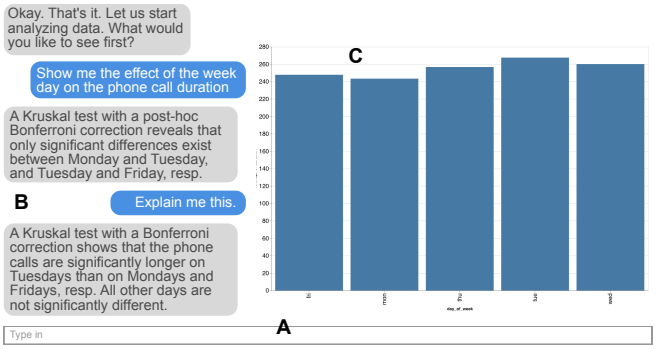

Figure 6: The GUI of the prototype for the controlled experiment optimized for a 13" notebook. Based on user's entered utterance (A), the CI provides both a presumably preferred answer (B) and a generated visualization (C).

analysis process (cf. 4), a CI should adjust to the user. This adjustment is potentially helpful in both avoiding implicatures and strengthening the user's communication with the CI.

USER STUDY 2: CONTROLLED EXPERIMENT

The user study's objective is to primarily investigate both **Q3** and **Q4**. Therefore, we implemented a prototypical web-based CI enhanced by a probabilistic model for predicting user's preferred answers from the *answer space*.

Prototype

The implemented prototype for the controlled experiment follows the design of *Valletto* [25] and incorporates the answer space (see Figure 5). For generating the visualizations, we use Vega-lite [39, 46]. While the frond end handles the user's interactions, all computational tasks are processed on a Flask server. Figure 6 shows how an interaction could look like.

Predicting User Preferences

The online survey results reveal that preferences in the answer space depend on a combination of user's multiple characteristics and the actual data situation. Therefore, we need a predictive model to retrieve a most suitable answer (cf. **Q2**) given those influencing factors. The primary objective is to avoid implicatures regarding the *maxim of quality* (predicting correctly the user's preferences) and the *maxim of relation* (predicting the right situation).

In either case, a user needs to actively provide information to the system in the first place. From a UX point of view, though, we decrease the effort on the user's side by primarily considering factors which have a significant influence on the

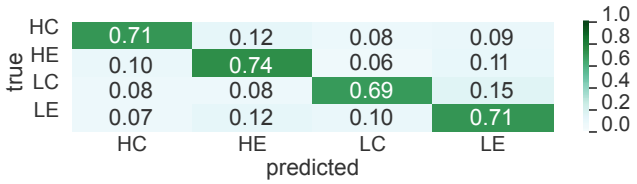

Figure 7: The confusion matrix of the classification results of the mLR + FS classifier highlights two things. First, the predictive performance is almost identical for each answer (diagonal = true positives). Second, the classifier can distinguish equally well between the answers.

user's preferences: self-reported knowledge, need for transparency as well as assistant, and the knowledge about the used measures.

Predicting a Preferred Answer

The problem of predicting an element from the answer space is a classification problem. However, the selection of an appropriate modelling technique has to be done under the constraint of the reduced feature space and the sample size. Our data set consists of 912 samples and 7 features. Therefore, we focus on classical approaches, since our problem does not fulfill the requirements of neural networks in terms of number of samples as well as number of features.

First, we investigated the performance of a multinomial Naïve Bayes (NB) as well as a multinomial logistic regression (mLR), both implemented in Python with scikit-learn [35]. Both models performed insufficiently for our objective (mLR: $\mu = .305$, $\sigma = .03$, NB: $\mu = .331$, $\sigma = .02$). We then extended our predictive models by a supervised feature selection (FS) via a random forest. This leads to a new space of binary features where high order feature interactions are included. Using these new features for the linear models, we receive a better performance with a highest accuracy by the mLR + FS classifier, $\mu = .697$, $\sigma = .02$ (vs. NB + FS: $\mu = .629$, $\sigma = .02$). Figure 7 illustrates the corresponding classification results.

Comparing the performance on the selected feature space to the potential entire feature space (all survey questions), we see that the performance slightly increases by 1% when using all features ($\mu = .706$, $\sigma = .01$). Hence, **Q2** is supported not only by the achieved results, but especially by the trade-off of requiring less information from the user and simultaneously achieving almost identical performance. This reduced need for asking the user potentially improves the user experience.

Prediction in an Actual Setting

In order to integrate the predictive model into a tool, two aspects have to be considered: The extraction needed information from the user as well as the prediction of the right situation during the analysis. From a linguistic perspective, both aspects are referring to the communication accommodation theory (CAT) [3]. CAT describes generally how people adapt their communication behavior to match the skills of their interlocutor. Two propositions of CAT are of particular interest: *accommodative orientation*, and *immediate situation* [14].

In *accommodative orientation*, the "initial orientation" [14] describes a prior adjustment of the conversation style to an interlocutor. In our case, we establish an initial dialogue between the CI and the user in order to gather the user's previous experience. The questions asked by the CI are according to the predictive model's feature space and are identically to the questions as in the online survey. This routine can be considered as users familiarize themselves with the system, as it would also happen in a conversation with a previously unknown person. Based on the given answers, the CI automatically adapts to the user. However, we presume the user's knowledge and preferences to be static for the duration of a session.

In *immediate situation*, the "goals and addressee focus" [14] aligns with Grice's *maxim of relation*. In our cases, it conceptually refers to how to predict the right data analysis situation based on the user's both recent utterance and analysis flow. Furthermore, a user might ask for a different answer, e.g., for additional parameters or an explanation. One the one hand, these circumstances might occur since probabilistic models are generally not flawless. On the other hand, the triggered reformulation requests contain additional information about the user. Therefore, they should be treated differently.

User Study Procedure

We conducted a qualitative user study, which lasted approximately 50 min per participant. The study took place as a face-to-face meeting between an investigator and a participant in a quiet room. The prototype was running on a 13" notebook.

The study began with a standardized introduction (10 min) to the study's design, the prototype, and the to be explored data set. The data set summarized a bank's marketing campaign and included numerical as well as categorical attributes [12, 31].

The interactive phase (30 min) started with a dialogue between the CI and the user in order to collect the user's previous experience w.r.t. the information required by the predictive model. Afterwards, the participants executed a randomly assigned sequence of six tasks (one for each supported situation). In order to better understand the participant's thoughts while conversing with the prototype, we incipiently encouraged each participant to think-aloud during the tasks. In each task, the objective was to make sense of the data (open exploration) by considering both visualization and answer(s). After each task, participants rated their satisfaction on 5-point Likert scale on both the visualization and the best-fitting answer. Furthermore, each participant had the chance to state how a preferred answer should look like in the specific task.

The study closed via a questionnaire (10 min) regarding the participant's feedback on the overall conversation. The questionnaire also asked for demographic information.

Participants

We recruited 10 participants (3 novices, 3 advanced beginners, 3 competent users, and 1 experts) aged between 25 and 40 within an industry company with an average professional work experience of 4 years. They had a background either in natural science, computer science, or economics. Furthermore, all have used some CIs before.

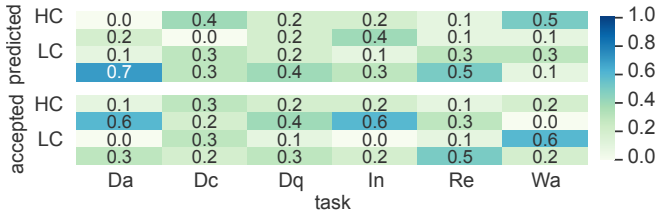

Figure 8: The heatmaps show the percentage-wise distributions in the answer space for each task of the predicted answers (top) and the accepted answer (bottom).

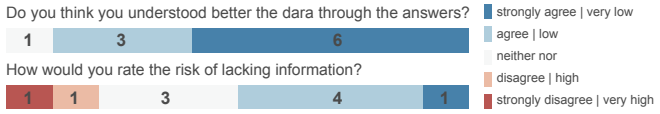

Figure 9: Count of participants' ratings on a 5-point scale on their data comprehension and risk estimation of hidden information by the agent.

Results

The participants rated overall the agent's answer as helpful or very helpful. In 59% of all cases, the first given answers were already considered to be of acceptable quality (30% after one reformulation, and 11% two reformulations). In case of a reformulation, the system in the end gave an explanatory answer with higher information in 44% of the cases. Thereby, the participants considered the agent's reformulation as positive by, e.g.,"*I liked that the tool changed the formulation when I requested it.*". By far the most reformulations were requested in *Dq*. However, the ratings of the usefulness of the last given answers were nearly equivalently high (4.3 on average) in all tasks. The difference in the first predicted answer and the finally accepted answer is shown in Figure 8.

When the agent gave an implicature, we tracked negative comments by the participants. One participant who did not know Spearman's ρ said "*I do not understand what the bot is saying*". The bot descriptively reported Spearman's ρ with corresponding $p - value$, which is arguably an implicature. This changed immediately to a positive reaction when the bot gave an explanatory answer with lower information "*...now I understand*". Furthermore, a participant commented, "*the first answer was too sketchy, but the second was much better*". We further observed that participants started thinking for a while when a potential implicature occurred: "*Well, I have to decode this (descriptive answer with lower information) first*" (novice user). However, when the explanatory answer was given instead the participant was much more satisfied.

If a participant considered an answer as helpful, the comments were rather positive: "*I really like the answers*". A participant who received mainly answers with low information said "*I liked the simple and short answers.*" Moreover, the explanation of certain data situations was welcomed *"the bot tried to explain things, that was good"*. Furthermore, the partici-

pants also trusted the agent in the conversation. The reasons, however, were different: "*I assume the system used the right tests*", or "*I trust the answer, because I had the same notion*". The number of comments depends on the task. The dialog act given in the *Wa* task received the most enthusiastic comments.

Regarding the system's initial questions, the participants generally commented positively, but not without restrictions. When the agent was asking if it is okay to ask seven questions, one participant promptly said: "*oh my good, seven questions*". However, the same participant as well as others told us afterwards that the questions were "*...not too much, but only okay for the first use*", but also "*not more questions*". Furthermore, some participants put the questions in the right context. One participant said: "I like that it tried to evaluate me in the beginning".

The relatively experienced data analysis participants tested the agent's functional boundaries. Finally, Figure 9 shows that the participants subjectively benefited from the given answers as well as assumed to receive all relevant information.

DISCUSSION

As the online survey initially revealed different preferences w.r.t. the answer space, the qualitative user study results empirically confirm the importance of actually matching the user's language in visual analysis. While dialogue acts given in accordance with the user's preferences evoke positive reactions, negative reactions are triggered when the preferences are mismatched. Additionally, these results confirm the sufficient flexibility of the answer space to cover the diverse preferences of the users. Hence, the answer space appears to be a lightweight and effective framework to personalize dialogue acts in visual analysis.

Experience with the Answer Space

Under **Q3**, we investigate the interaction with a user-specific conversation. Considering both the online survey results and participants' reactions, it makes a substantial difference how the data situations are verbally communicated.

Figures 3 and 8 show the different preferences in the answer space. Hence, violating these preferences in the conversation trigger probably inadvertent users' reactions. The participants' comments in the user study support further this assumption. In situations of potential implicatures, the reactions / comments were rather negative. Especially when a requested reformulation better fitted the preferences, the differences in the reactions can be clearly seen. As implicatures triggered negative reactions, the reactions were rather positive when a given answer matched the preferences.

Generally, the agent's answers were predominately seen as very positive according to the participants' comments which further supports the performance of the predictive model. Interestingly, the answer space seems to become even more relevant when a task gets more complicated. Considering the reactions in the tasks *Dq* and *Wa*: The participants' statements were mainly emotionally neutral in task *Dq*, while they were positively surprised in task *Wa*. These observations arguably corroborate the dependency to a task's complexity, since *Dq* is known by many users, but *Wa* is probably not. Hence, a possible system's generalized communication strategy would have failed in these situations.

Both circumstances imply that the UX (e.g. frustration / satisfaction) of CIs for visual analysis is certainly affected by the way the system communicates. However, not only the UX is affected, but also the usability. Since some participants needed more time in order to understand an agent's potentially too complicated answer, using the answer space is likely to speed up the analysis process. Communicating in accordance to the user's preferences makes a reformulation obsolete as well as reduces the time to make sense of the agent's dialogue act.

Acceptance of an Initial Orientation

In order to achieve the best possible prediction of preferred answers, the initial orientation of the system is needed. The participants' comments point out that answering a few questions in the beginning of the first use is acceptable when the user assumes to get a benefit in return, although it turned out that the seven questions were at the limit of getting a burden for the user. On the other hand, users are often not fully aware of a CI's functionality. Hence, implementing an initial mutual introduction between the user and the system potentially improves the conversation strategy on both sides.

Granularity of the Answer Space

Under **Q4**, we investigate the granularity of the answer space. The participants suggested only a few alternative formulations for the responses of CIs. Additionally, these formulations focused essentially on the wording. The comments did not indicate a need for changing the answer space's granularity. Therefore, we infer that the 2×2 answer space is sufficient, at least for the analyzed situations.

LIMITATIONS

Our objective with introducing the answer space was to provide a framework for communicating data relationships in a user-specific way. We focused on identifying differences in the users' preferences as well as make the answer space applicable in practice. We did not focus on investigating a certain wording.

Effectiveness of the Answer Space

The question still remains whether the users' preferences in the answer space depend on the corresponding visualization. In our case we only used scatter plots, line, and bar charts. However, there are more visualizations available. Thus, it is an open question whether the preferences would be different for other visualizations. Finally, our aim is to improve visual analysis through a user-specific conversation. As the qualitative results suggest that the answer space can improve the user's data comprehension, we lack though of quantitative evidence.

Limited Number of Used Methods

One particular limitation is the number of evaluated actual situations. Although we focused on the main occurring challenges in data analysis, we did not cover all relevant data analysis methods. For instance, there are at least two additional methods for estimating the relationship between quantitative

variables: Pearson's ρ and Kendall's τ. Since Pearson is more commonly known than Spearman, this knowledge difference is likely to affect the users' preference in the answer space.

Performance of the Predictive Model

Our probabilistic model can predict a user's preferences in the answer space at an accuracy of approximately 70%. However, in the user study we achieved only an initial accuracy of 59%. One reason may be that certain participants' backgrounds are underrepresented in our online survey, which may have an impact on the training data for the model. For those participants, the uncertainty for a predicted answer is likely higher.

DESIGN IMPLICATIONS

Based on our findings, we derive the following three design implications for CIs in visual analysis:

DI1: *Design multiple dialogue acts for communicating data facts:* A CI should avoid a generalized conversation in visual analysis, since users have diverse preferences regarding the communication of data facts. For each supported data fact, hence, a set of different dialogue acts should exist in order to give the CI the opportunity to communicate in a personalized way. Otherwise undesirable implicatures are triggered, which will negatively affect the user experience.

DI2: *Do not only rely on the user's knowledge as a basis for the conversation design:* As the user's preferences are only partially explained by their statistical knowledge level, additional characteristics (e.g., need for transparency) have to be taken into account for the conversation design as well as for user modelling in an intelligent system.

DI3: *Realize a mutual introduction:* In human-human dialogues, people mutually introduce themselves when they meet the first time. As our findings show that users are actually willing to provide initial information about their background under the assumption of a subsequent benefit, a tool should conceptually realize an initial mutual introduction. This would not only help the user to understand the functionality of the tool, but also give the tool the opportunity to adjust its functionality to the user.

Although our findings primarily address the dialogue acts of CIs, these derived design implications might additionally exceed the boundaries of the subject of CIs. Other concepts combine visualizations with textual data descriptions as well, e.g., interactive data facts [43], annotations, or storytelling. These concepts can apply DI1 and DI2 for their design, since a dialogue act that is not embedded in an entire dialogue is conceptually similar. Finally, the design implication DI3 essentially focuses on CIs but makes no additional assumptions on the particular domain. Hence, a CI that acts in a different domain than visual analysis, but has in common to serve users on different domain knowledge levels, might also consider DI3 for design.

FUTURE CHALLENGES

In addition to these design implication, our framework opens up many research avenues for further elaborating the potentials of CIs in individually supporting users. As we initially focused on personalized dialogue acts, future work could explore the design of entire personalized dialogue sequences to eventually transform the interaction with CIs in visual analysis from a task-oriented "virtual butler" [34] to an actual conversation.

A first challenge is trustworthiness. In order to be entirely useful, a CI does not only have to be sure whether a data fact is true (e.g., based on p-value), but also has to strictly avoid potential implicatures. As an expert is able to meticulously challenge a CI's response, an inexperienced user might not, which leads to wrong conclusions in the end.

A second challenge is how to directly educate the user in the analysis process, e.g., by interactively explaining a concept like ordinary least squares regression [36]. The explanation would of course be adapted to the data at hand, which potentially would increase the user's motivation for dealing with the new concept or method. Methodologies are needed to identify opportune moments in a user's analysis for effectively placing these brief educational interludes. However, an educational session should only be activated when a user really needs it. Hence, a personalized methodology should sense a user's individual need for education. For an intelligent sensing, DI2 appears to be directly relevant. For the conversation design itself, DI1 should be taken into account. Additionally, combining our answer space with the pragmatics-based approach by Hoque et al. [21] could further cover challenges along an interaction sequence in visual analysis [49].

As the language is a CI's primary modality, CIs will eventually become able to effectively support users with very different backgrounds. They may become even more effective than classical interaction concepts for visual analysis, but only when matching the user's conversational needs.

CONCLUSION

Motivated by the idea of improving CIs for visual analysis for a broader spectrum of users, we focused on matching the user's language in a personalized way by constructing a linguistically motivated answer space. Through an online survey ($N = 76$) as well as a qualitative user study ($N = 10$) we empirically showed that it makes a considerable difference how a CI communicates in visual analysis. First, the diverse user's preferences in communicating data facts significantly depend on both the user's characteristics and the analysis situation. Second, neglecting the user's characteristics in the conversation design in visual analysis triggers negative reactions while user-specific answers evoke positive reactions. Hence, adjusting to those preferences improves both the user experience and the usability of a CI for visual analysis. By providing design implications grounded in our empirical findings, our contribution helps to better understand matching the user's language in visual analysis.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions expressed in this paper do not necessarily reflect the views of the Volkswagen Group.

REFERENCES

[1] U.S. General Services Administration. 2018. The home of the U.S. Government's open data. (2018). https://www.data.gov

[2] Alan Barnes. 2016. Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intelligence and National Security* 31, 3 (2016), 327–344. DOI: http://dx.doi.org/10.1080/02684527.2014.994955

[3] Zhang Yan Bing and Giles Howard. 2017. *Communication Accommodation Theory*. American Cancer Society, 1–14. DOI: http://dx.doi.org/10.1002/9781118783665.ieicc0156

[4] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (2011), 41–57. DOI: http://dx.doi.org/10.1016/j.cognition.2011.05.011

[5] David V. Budescu, Shalva Weinberg, and Thomas S. Wallsten. 1988. Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance* 14, 2 (1988), 281–294.

[6] Hsinchun Chen, Roger H L Chiang, and Veda C. Storey. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36, 4 (2012), 1165–1188.

[7] Mei-Ling Chen and Hao-Chuan Wang. 2018. How Personal Experience and Technical Knowledge Affect Using Conversational Agents. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion)*. ACM, New York, NY, USA, Article 53, 2 pages. DOI: http://dx.doi.org/10.1145/3180308.3180362

[8] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding Quantified-selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1143–1152. DOI:http://dx.doi.org/10.1145/2556288.2557372

[9] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. DOI: http://dx.doi.org/10.1145/3098279.3098539

[10] Kenneth Cox, Rebecca E. Grinter, Stacie L. Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A Multi-Modal Natural Language Interface to an Information Visualization Environment. *International Journal of Speech Technology* 4, 3 (July 2001), 297–314. DOI:http://dx.doi.org/10.1023/A:1011368926479

[11] Mandeep K. Dhami, David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. Improving Intelligence Analysis With Decision Science. *Perspectives on Psychological Science* 10, 6 (2015), 753–757. DOI: http://dx.doi.org/10.1177/1745691615598511

[12] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. (2017). http://archive.ics.uci.edu/ml

[13] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. 2018. Iris: A Conversational Agent for Complex Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 473, 12 pages. DOI: http://dx.doi.org/10.1145/3173574.3174047

[14] Cindy Gallois, Tania Ogay, and Howard Giles. 2005. Communication accommodation theory: A look back and a look ahead. *Theorizing About Intercultural Communication* (Jan 2005).

[15] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology (UIST '15)*. ACM, Charlotte, NC, USA, 489–500. DOI: http://dx.doi.org/10.1145/2807442.2807478

[16] Raymond W. Gibbs and Gregory A. Bryant. 2008. Striving for optimal relevance when answering questions. *Cognition* 106, 1 (2008), 345 – 369. DOI: http://dx.doi.org/10.1016/j.cognition.2007.02.008

[17] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov 2010), 943–952. DOI: http://dx.doi.org/10.1109/TVCG.2010.164

[18] H. Paul Grice. 1975. Logic and Conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, Peter Cole and Jerry L. Morgan (Eds.). Academic Press, New York, 41–58.

[19] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science* 56, 4 (2005), 889–911.

[20] Matthew Henderson. 2015. Machine Learning for Dialog State Tracking: A Review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

[21] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 309–318. DOI: http://dx.doi.org/10.1109/TVCG.2017.2744684

[22] Rogers Jeffrey Leo John, Navneet Potti, and Jignesh M Patel. 2017. Ava: From Data to Insights Through Conversation. In *Biennial Conference on Innovative Data Systems Research (CIDR 2017)*.

[23] Project Jupyter. 2018. The Jupyter Notebook. (2018). https://jupyter.org

[24] Jan-Frederik Kassel and Michael Rohs. 2017. Immersive Navigation in Visualization Spaces through Swipe Gestures and Optimal Attribute Selection. In *Workshop on Immersive Analytics: Exploring Future Interaction and Visualization Technologies for Data Analytics (IEEE VIS '17)*. Phoenix, AZ, USA.

[25] Jan-Frederik Kassel and Michael Rohs. 2018. Valletto: A Multimodal Interface for Ubiquitous Visual Analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW005, 6 pages. DOI:http://dx.doi.org/10.1145/3170427.3188445

[26] Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano, Andrew Johnson, Abeer Alsaiari, Nigel Flowers, Alberto Gonzalez, and Jason Leigh. 2017. Towards Multimodal Coreference Resolution for Exploratory Data Visualization Dialogue: Context-Based Annotation and Gesture Identification. *The 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017 – SaarDial)* (August 2017), 48.

[27] Henrik Leopold, Jan Mendling, and Artem Polyvyanyy. 2014. Supporting Process Model Validation through Natural Language Generation. *IEEE Transactions on Software Engineering* 40, 8 (Aug 2014), 818–840. DOI: http://dx.doi.org/10.1109/TSE.2014.2327044

[28] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. DOI: http://dx.doi.org/10.1145/2858036.2858288

[29] Pattie Maes. 1994. Agents That Reduce Work and Information Overload. *Commun. ACM* 37, 7 (July 1994), 30–40. DOI:http://dx.doi.org/10.1145/176789.176792

[30] Rolf Molich and Jakob Nielsen. 1990. Improving a Human-computer Dialogue. *Commun. ACM* 33, 3 (March 1990), 338–348. DOI: http://dx.doi.org/10.1145/77481.77486

[31] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22 – 31. DOI:http://dx.doi.org/10.1016/j.dss.2014.03.001

[32] narrative science. 2018. Quill. (2018). https://narrativescience.com/products/quill

[33] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (June 2010), 411–419.

[34] Sabine Payr. 2013. Your Virtual Butler. Springer-Verlag, Berlin, Heidelberg, Chapter Virtual Butlers and Real People: Styles and Practices in Long-term Use of a Companion, 134–178.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Nov. 2011), 2825–2830.

[36] Victor Powell and Lewis Lehe. 2018. Ordinary Least Squares Regression. (2018). http://setosa.io/ev/ordinary-least-squares-regression

[37] Alejandro Ramos-Soto, Alberto Jose Bugarín, Senén Barro, and Juan Taboada. 2015. Linguistic Descriptions for Automatic Generation of Textual Short-Term Weather Forecasts on Real Prediction Data. *IEEE Transactions on Fuzzy Systems* 23, 1 (Feb 2015), 44–57. DOI:http://dx.doi.org/10.1109/TFUZZ.2014.2328011

[38] Silja Renooij and Cilia Witteman. 1999. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning* 22, 3 (1999), 169 – 194. DOI: http://dx.doi.org/10.1016/S0888-613X(99)00027-4

[39] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 341–350. DOI: http://dx.doi.org/10.1109/TVCG.2016.2599030

[40] Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Vol. 1. Cambridge University Press.

[41] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, Tokyo, Japan, 365–377. DOI: http://dx.doi.org/10.1145/2984511.2984588

[42] Nicole Shechtman and Leonard M. Horowitz. 2003. Media Inequality in Conversation: How People Behave Differently when Interacting with Computers and People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 281–288. DOI: http://dx.doi.org/10.1145/642611.642661

[43] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 672–681. DOI:http://dx.doi.org/10.1109/TVCG.2018.2865145

[44] Arjun Srinivasan and John Stasko. 2018. Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 511–521. DOI: http://dx.doi.org/10.1109/TVCG.2017.2745219

[45] Alexander Strehl and Joydeep Ghosh. 2003. Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. *The Journal of Machine Learning Research* 3 (March 2003), 583–617. DOI: http://dx.doi.org/10.1162/153244303321897735

[46] Jacob VanderPlas, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. Altair: Interactive Statistical Visualizations for Python. *Journal of Open Source Software* (dec 2018). DOI:http://dx.doi.org/10.21105/joss.01057

[47] Giorgio Vassallo, Giovanni Pilato, Agnese Augello, and Salvatore Gaglio. 2010. *Phase Coherence in Conceptual Spaces for Conversational Agents*. John Wiley & Sons, Ltd, Chapter 18, 357–371. DOI: http://dx.doi.org/10.1002/9780470588222.ch18

[48] Thomas S. Wallsten, David V. Budescu, Rami Zwick, and Steven M. Kemp. 1993. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society* 31, 2 (01 Feb 1993), 135–138. DOI: http://dx.doi.org/10.3758/BF03334162

[49] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 479, 12 pages. DOI: http://dx.doi.org/10.1145/3173574.3174053