Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations

Arjun Srinivasan*
Tableau Research
Seattle, WA, USA
arjunsrinivasan@tableau.com

Nikhila Nyapathy
Georgia Institute of Technology
Atlanta, GA, USA
nikhila.nyapathy@gmail.com

Bongshin Lee
Microsoft Research
Redmond, WA, USA
bongshin@microsoft.com

Steven M. Drucker
Microsoft Research
Redmond, WA, USA
sdrucker@microsoft.com

John Stasko
Georgia Institute of Technology
Atlanta, GA, USA
john.stasko@cc.gatech.edu

ABSTRACT

Natural language interfaces (NLIs) for data visualization are becoming increasingly popular both in academic research and in commercial software. Yet, there is a lack of empirical understanding of how people specify visualizations through natural language. We conducted an online study (N = 102), showing participants a series of visualizations and asking them to provide utterances they would pose to generate the displayed charts. From the responses, we curated a dataset of 893 utterances and characterized the utterances according to (1) their phrasing (e.g., commands, queries, questions) and (2) the information they contained (e.g., chart types, data aggregations). To help guide future research and development, we contribute this utterance dataset and discuss its applications toward the creation and benchmarking of NLIs for visualization.

CCS CONCEPTS

• **Human-centered computing** → **Visualization**; *Empirical studies in visualization*; *Natural language interfaces.*

KEYWORDS

data visualization, natural language interfaces, natural language processing, natural language corpus, visualization specification

---

*This work was conducted while the author was affiliated with the Georgia Institute of Technology.

---

1 INTRODUCTION

Natural language (NL) is gaining traction as an input modality for data visualization tools. Both mainstream visualization systems (e.g., Microsoft Power BI [16], Tableau [27]) and research prototypes (e.g., [7, 10, 21, 24, 26, 30]) have demonstrated the potential natural language interfaces (NLIs) hold for supporting a rich visual analytic flow and to cater to broader audiences (e.g., visualization novices, people with disabilities). *Visualization specification* plays a central role in these tools, as people create charts to visually explore and analyze their data (e.g., creating histograms to check data distributions, creating scatterplots to observe correlations), as well as to communicate their findings with other stakeholders (e.g., creating grouped bar charts for quarterly reports).

However, beyond the space of utterances supported in current NLIs for visualization, there is limited empirical understanding about the nature of utterances people use to specify data visualizations through NL. For instance, how do people structure or phrase their utterances? Do they use systemic commands (e.g., *"Plot scatterplot of sales by profit."*) or high-level questions (e.g., *"Have cars gotten lighter over time?"*)? What type of information do people include in their utterances? Do they explicitly list chart types and aggregations (e.g., *"Bar chart showing average profit by state"*) or expect systems to infer such information for them (e.g., *"Visualize profit across states."*)? Furthermore, in cases where people provide explicit references, are these references direct (e.g., exact or partial matches to attribute names), or implied or semantic (e.g., using synonyms for attribute names, using phrases like *"How many"* instead of specifying a COUNT aggregation)? Answering these questions can help assess how well-aligned current tools are with people's expectations from NLIs for visualization. This, in turn, can help developers of NLIs for visualization validate and boost their systems' performance and improve system usability.

We conducted an online study with 102 participants, curating a dataset of 893 visualization specification-oriented NL utterances. Specifically, we showed participants a series of ten canonical visualizations (e.g., bar charts, line charts, scatterplots) and asked them to enter utterances they would use to create the shown charts. We then characterized the resulting utterances along two dimensions: 1) their phrasings, i.e., how the utterances were structured (e.g., questions, commands), and 2) the information contained in them (e.g., some utterances explicitly requested a visualization type). For

the remainder of this paper, we use the term **utterance** to refer to any NL command, statement, query, question, or instruction that one may issue to an NLI.

Leveraging this characterization, we provide a high-level summary of the types of utterances that people use to specify visualizations, discussing their coverage within current systems and revealing underexplored classes of utterances. We also distill our experiences and highlight critical considerations for designing and conducting online studies to collect data for NLIs for visualization. Finally, we discuss how the curated dataset of utterances can be used toward the creation and benchmarking of NLIs for visualization, and for collecting larger datasets.

The key contributions of this work are twofold: (1) a publicly available dataset of 893 visualization specification-oriented utterances and a discussion of its application toward creating and benchmarking NLIs, and (2) a characterization of NL utterances people use to specify data visualizations along with its implications for designing future NLIs for visualization specification.

2 RELATED WORK

Advances in NL recognition and understanding have led to a surge of interest in NLIs for data visualization. In recent years, a wide range of systems have been developed to support NL interaction during visual data exploration (e.g., [10, 12, 21, 23, 24, 26]), for question answering with charts [13], and for facilitating data-driven communication [7, 14]. A majority of these systems focus on specific visualizations and usage scenarios (e.g., network data exploration [25], dataflow diagram editing [30]) or technical aspects of system development (e.g., detecting ambiguities [10, 21], inferring underspecified queries [22]).

On the other hand, some projects have more closely investigated the structure of NL utterances and thus are highly relevant to our work. Metoyer et al. [15] conducted a study where 20 participants worked in pairs with one participant verbally describing a given chart to another participant who attempted to draw the chart based on that description. Through the collected data, the authors discuss implications for future NL systems including the need to support ambiguous instructions (e.g., *"a narrow bar"*) and the use of relative terms to arrange spatial layouts (e.g., *"the yellow bar is stacked on top of the blue bar"*), among others. Setlur et al. [21] conducted a study to collect utterances people may pose when interacting with a given chart. They identified different task categories people try to perform (e.g., search, filter, change chart type) and used the corresponding utterances to implement the Eviza system [21]. Tory and Setlur [28] conducted a wizard-of-oz study to understand people's expectations of the role of system intelligence in NLIs during visual analysis. Their findings highlight that NLIs can enable a richer visual analytic flow by supporting both explicit and implicit user intent, by prioritizing explicit intent over consistency in use of graphical encodings during chart transitions, and by performing proactive actions, such as visually encoding filters and implicitly applying data transformations, among others. Our work shares a similar goal with these studies in that we also seek to understand the nature of NL utterances within the context of data visualization. However, we focus on collecting and understanding utterances people may pose to a system for visualization specification (as opposed

to interacting with a given chart). In doing so, we provide additional evidence for findings from prior studies while also offering novel insights into classes of utterances not supported in current tools and discussing practical implications for system design.

Recent work has also begun to curate datasets of utterances to enable the use of machine learning techniques for NLI development. For instance, during an online study, Kim et al. [13] showed people a series of bar charts and line charts, and collected questions people pose to these charts along with the participants' answers (and explanations) to those questions. This data was then used to update Sempre [5], a model conventionally used for question answering within relational data tables, to answer questions in the context of data visualizations. Fu et al. [9] collected utterances mapping to Amar et al.'s 10 low-level analytic tasks (e.g., filter, sort, find extremum) [2]. Using the collected data to train a text classification model, the authors showed how one could detect analytic tasks from given NL utterances. Similar to these recent studies, we also aim to foster future research and development by curating a freely available dataset of utterances. To this end, we complement the existing datasets that focus on question answering and task detection, and contribute the first dataset that contains visualizations and utterances that were used to specify those visualizations. Besides discussing the applications of the presented dataset, we also offer insights into unique challenges we faced during data collection and curation as these can help extend the described study and generate larger datasets.

3 ONLINE STUDY

3.1 Visualizations and Datasets

Our study covered a total of 30 visualizations (10 visualizations x three datasets): Figure 1 shows the 10 visualizations that were presented during each session and Table 1 summarizes the three datasets used in the study. These visualizations covered three popular chart types: *bar charts*, *line charts*, and *scatterplots*, along with their variants (histograms, stacked & grouped bar charts, multiseries line charts, and colored & faceted scatterplots). Our choice of visualizations was motivated by the popularity of these charts across visualization tools in general [4] as well as within existing NLIs for visualization (e.g., [10, 21, 23, 26]). Furthermore, these visualizations also allowed us to cover three basic data types: *categorical* (both nominal and ordinal values), *quantitative*, and *temporal*, as well as their combinations involving one to three data attributes at a time. Collectively, this set of visualizations and attribute type combinations ensured that the utterances collected via the study would be broadly applicable and could be used in the context of general visualization tools.

Dataset (Cardinality)	Categorical Attributes	Quantitative Attributes	Temporal Attributes
Cars (303)	2	5	1
Movies (709)	3	5	1
Superstore (5899)	11	5	1

Table 1: Summary of three datasets used for the study (complete datasets are provided as supplementary material).
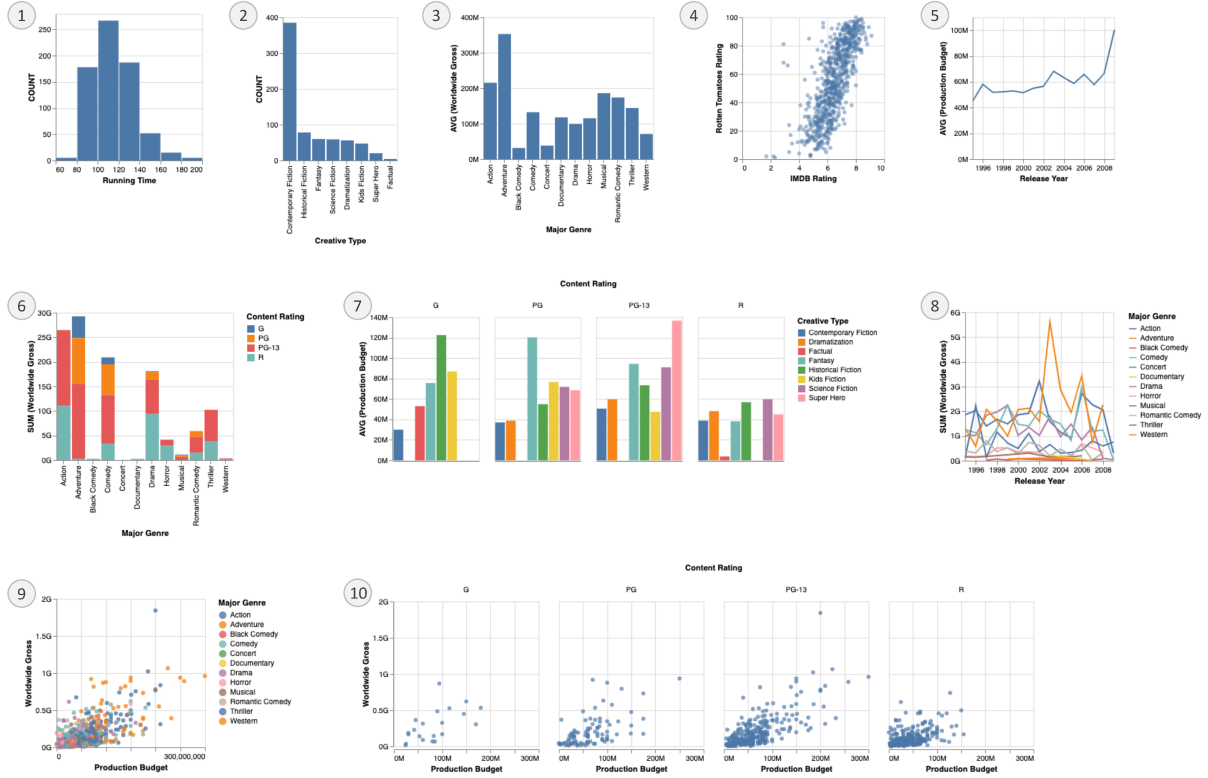
Figure 1: Ten visualizations used for the study (all charts were created using Vega-Lite [20]). Charts shown in the figure represent the movies dataset. Charts (1, 2) visualize one quantitative or categorical attribute using the x- and y-encoding channels. Charts (3-5) also only encode data via the x- and y-channels but visualize two attributes at a time. Finally, charts (6-10) use `color` and/or `column` (for faceting) as an additional encoding to visualize three attributes at a time. The corresponding ten charts for the cars and superstore data are included in the supplemental materials.

For the data underlying these charts, we decided to use datasets that: (1) contain a mix of categorical, quantitative, and temporal values, and (2) represent a generally understood domain to engage a broad participant pool. With these criteria in mind, we chose three tabular datasets covering different domains: *cars*, *movies*, and *superstore sales* (Table 1).

3.2 Participants

To collect utterances that were representative of what end-users of a visualization system might provide, we sought to recruit participants who are actual users of visualization tools or are interested in visual data analysis. With this target population in mind, we shared the call for participation for the study via mailing lists at universities and software companies. We specifically used mailing lists that included data systems-related users (e.g., database and visualization courses, dashboard tool user groups, visualization tool developers and product managers). We also posted the call for participation on public social platforms such as LinkedIn and Reddit. On these social platforms, we again posted only in targeted

interest groups pertaining to data visualization and visual analytics (e.g., 'r/visualization' and 'r/PowerBI' communities in Reddit, Tableau user groups on LinkedIn, the Data Visualization Society Slack channel). Participation was voluntary and no compensation was provided (neither financially nor via other reward mechanisms such as course credit). Participants could choose to exit the study at any point. Participants were also informed beforehand that the utterances collected through the study will be made publicly available and that no personal information (contact details or demographic information) will be captured during the study.

Over a span of ∼60 days, a total of 202 participants visited the study URL among which 102 participants took part in the study (i.e., completed at least one trial). Seventy-six participants completed the entire session, entering utterances for all ten visualizations while 26 participants exited the study amidst a session.

3.3 Procedure and Task

Each session included four major phases: (1) *introduction & consent*, (2) *dataset & task description*, (3) *task trials*, and (4) *providing prior experience with visualization tools*, and consisted of 10 trials

corresponding to the 10 visualizations displayed in Figure 1. Each participant saw all 10 visualizations from only one of the three datasets (cars, movies, superstore). Both the dataset selected and the order of visualizations were randomized across participants.

Participants were first informed about the study goal, compensation, public sharing of collected utterances, etc. and were asked to provide their consent. Next, participants were given a brief introduction to the dataset (e.g., number of rows and columns, what the rows and columns represent) and the task, which required participants to enter (type in) one or more utterances they would pose to a system like Tableau or Microsoft Power BI to specify a displayed visualization (Figure 2). Note that to ensure participants were not biased to phrase utterances in a specific way, we did not provide any example utterances as part of the task description. Participants were then presented with the 10 trials in random order. Lastly, participants who completed the trials were given the option to provide their prior experience level using visualization tools on a five-point scale ranging from 'Never' to 'Very Frequently'. All utterances entered during the study were logged for further analysis.

3.4 Study Design Considerations

We iteratively developed the final study interface and procedure described above through 41 pilot sessions, including two in-person group sessions[1] (with 11 and 14 participants, respectively) and 16 distributed online sessions. To the best of our knowledge, this is the first online study designed specifically for collecting NL utterances for visualization specification. Correspondingly, below we briefly reflect on the pilot feedback and the evolution of our study design to highlight key considerations for future online studies investigating NL input for data visualization.

3.4.1 No training trials or sample inputs. As we required participants to enter utterances that naturally occurred to them, we did not present a "training phase" or sample inputs at the start of the study to avoid biasing the utterances participants entered. This absence of training trials is one crucial consideration compared to more traditional perception- and interaction-focused online visualization experiments that use training trials to acclimatize participants to the study tasks and interface.

3.4.2 Phrasing task prompts. One important factor for future studies to consider is the phrasing of the task prompt (Figure 2C). To cite an anecdote, during the pilots, we initially used the prompt "*Enter one or more natural language commands or utterances that you would pose to a system to create the visualization shown on the right.*" With this prompt, a majority of the input utterances were phrased along the lines of *"Create a [chart type] of [attributes]"* where the start *"Create a"* was consistent across almost all participants. While these were natural utterances for some participants, other participants said that the prompt nudged them towards phrasing their utterances in a particular way. Correspondingly, to repeatedly remind participants that they were free to phrase utterances per their individual preferences, we updated the task

---

[1]During the group sessions, participants completed the study individually on their laptops and then collaboratively provided feedback on issues they faced (e.g., incorrect chart sizes) or instructions that were confusing to them (e.g., task phrasing).

prompt to (*"Type one or more natural language..."* as shown in Figure 2C. Specifically, we included a variety of terms to refer to participant input (e.g., questions/statements/queries/...) and randomized the ordering of these terms (e.g., utterances/queries/statements/..., queries/questions/commands/...), and also varied the verbs used in the prompt (e.g., create, generate, specify).

3.4.3 Choosing the number of required utterances. Setting the minimum number of utterances that participants need to enter is also an important study design factor to consider. We initially asked the pilot participants to enter *two or more* utterances they would use to specify the given chart. However, participants commented that it was challenging to provide more than one "natural" utterance, leading them to create a contrived second utterance by paraphrasing their first utterance. Furthermore, instead of reading "*two or more*" as two mutually exclusive utterances to specify the shown chart, some participants misinterpreted the instructions as a mandate for entering at least two utterances that collectively specified the shown chart (e.g., *"Show me a bar chart"* > *"Add states and average profit"*). We therefore asked participants to enter *one or more* utterances in the final study (Figure 2C).

3.4.4 Providing a dataset preview. One subtle but important design consideration is the inclusion of a dataset preview as part of each trial (Figure 2A). In the initial pilots, we included the dataset preview only during the introductory phase but not for the actual trials. However, we observed that including a dataset preview alongside the chart gave participants a better understanding of the dataset, helping them formulate more naturalistic utterances involving semantic and value-based references to data attributes (e.g., using phrases like *"over time"* or *"movie length"* instead of the attribute names *Year* or *Running Time*, or values like *"furniture, office supplies, and technology"* to refer to *Category*).

4 ANALYSIS AND RESULTS

4.1 Data Cleaning

A total of 803 logs were generated from the 102 sessions (76 complete + 26 partial sessions, one log per trial). We manually inspected these logs to curate a set of logs containing only valid utterances that could reliably be used for future research and development. We discarded 159 trials, leaving a total of 644 trials: we discarded trials if they contained utterances that fell into the following categories:

- **Arbitrary text** (45): Trials that contained utterances composed of an arbitrary collection of characters. Examples include *"njknjk,"* *"xyz,"* *"fdsfs,"* and *"na."*
- **Code or SQL** (43): Trials that contained SQL commands or code in programming languages, such as R and Python. Examples include *"SELECT Year, Origin, AVG(Horsepower) FROM dataset WHERE Origin IN ('Europe', 'Japan', 'USA') GROUP BY Origin, Year"* and *"ggplot(dataset) + geom_point(mapping= aes(x= Horsepower, y= Acceleration)"*.
- **Incomplete utterances** (37): Trials with utterances that were incomplete in the context of recreating the displayed visualizations. For instance, for the trial shown in Figure 2, we excluded the utterance *"show me displacement vs miles per gallon"* because it does not specify, in any way, that *Origin* should be visualized as a third data attribute and used for coloring the
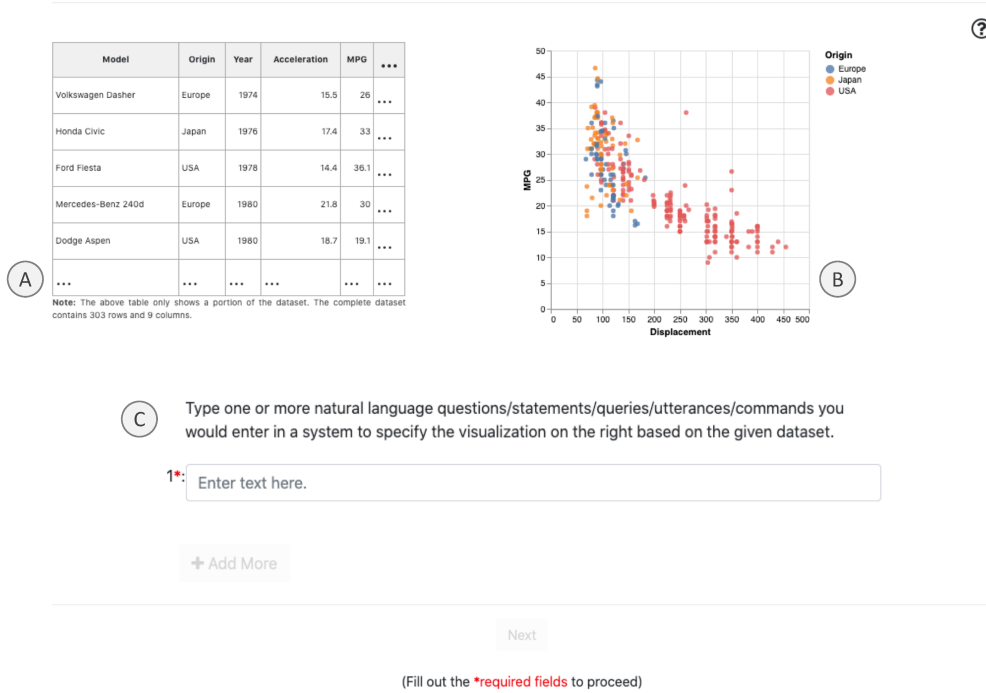
Figure 2: Screenshot from the study illustrating a single task trial. The example shown is from a session visualizing the cars dataset. (A) Dataset preview, (B) chart to specify, and (C) task prompt and input. Besides entering utterances, participants could also click the ⊙ icon on the top-right corner to revisit the dataset and task overview.

points. While incomplete utterances may be considered valid for the given datasets during general data exploration, they lacked sufficient information for a visualization system to adequately infer the shown chart from the utterance alone. Other discarded examples under this category include *"Bar graph"* (for a bar chart that needs two attributes similar to Figure 1-3), *"average of horsepower"* (for a multi-series line chart similar to Figure 1-8), and *"Show me bar graphs of each ship mode"* (for a stacked bar chart with three attributes similar to Figure 1-6).

- **Miscellaneous utterances** (34): Trials containing utterances where it looked like participants did not follow the instructions and entered utterances that were not relevant to the task of specifying a displayed chart. For example, for the chart in Figure 2, we discarded a trial composed of only the following two utterances: *"Overlay Legend top right"* and *"Add Exponential Trend."* Other discarded examples include *"add data labels, inside end,"* *"Change y-axis major units to 2,"* and *"Title start with A."*

4.2 Results

Among the filtered group of 644 trials, 252 corresponded to the cars dataset, 205 to the movies dataset, and 187 to the superstore dataset. The number of trials for each [dataset x chart type] combination ranged 15–28 ($M = 21$). The 76 participants who completed all 10 trials also provided their prior working experience with visualization tools (Never: 3, Somewhat infrequently: 9, Occasionally: 15,

Somewhat frequently: 23, Very frequently: 26). In terms of utterances, from the 644 trials, we extracted a total of 893 utterances with individual trials containing one to seven utterances.

Characterizing Utterances. We qualitatively analyzed the 893 utterances through an open coding process. Specifically, two of the authors collaboratively inspected the data to identify high-level themes and generated a set of codes to characterize and group utterances. We then individually tagged random subsets of utterances using the initial coding scheme and compared the coding results for agreement. The resulting codes were then collectively discussed and refined for consistency until an 85% agreement was reached using the Jaccard Index. Subsequently, all 893 utterances were tagged using the mutually agreed-upon codes.

From the 893 utterances, we identified a total of 814 ***utterance sets***: we define an utterance set as a set of one or more utterances that collectively map to a specific visualization. Utterance sets can either be *singleton* (755) or *sequential* (59). A singleton utterance set contains only one utterance intended to specify one chart. Examples of singleton utterance sets for the chart shown in Figure 2 include (1) *"Scatterplot mpg vs displacement color by origin"* and (2) *"What is the correlation between displacement and MPG of cars with different origins?"* On the other hand, a sequential utterance set is composed of multiple utterances that collectively create one chart. Going back to the chart in Figure 2, an example of a sequential utterance set (with two utterances) is *"plot displacement by mpg"* > *"color by origin"*

(the symbol '>' denotes a follow-up utterance). The 59 sequential utterances were composed of 2–5 utterances ($M = 2$). The lengths of utterance sets ranged 2–61 words ($M = 10$), with individual utterances containing up to 38 words ($M = 9$), see Figure 3.

To better understand the nature of utterances that people use to specify visualizations, we characterized the 814 utterance sets along two key dimensions: **How** the utterances are phrased and **What** dataset or visualization-specific information they contain. Figure 4 summarizes the results of this characterization.

In terms of phrasing variations, we tagged each of the 814 utterance sets as one of: *commands* (368), *queries* (260), *questions* (114), or *others* (72). Figure 4-top presents examples of utterances under each of these phrasing categories. Specifically, we defined 'commands' as utterances that were phrased similar to instructions or requests from one person to another. We tagged an utterance set as a 'query' when the underlying utterances were phrased as keywords or terse web search-like queries. 'Questions' were data-driven questions that participants expected to see the displayed visualization as a response to. If an utterance set did not fall under one of the three aforementioned phrasing categories, we tagged it as 'others' because such phrasings were relatively infrequent. Broadly speaking, the types of phrasings in the 'others' category included caption-like statements describing a chart, commands that were primarily in NL but also used special characters like '=' and '( )' that have a programming connotation, or detailed chart rendering instructions.

With respect to the information contained within utterances, we identified five types of references that would be relevant to an NLI when interpreting utterances to create visualizations: *attribute*, *chart type*, *encoding*, *aggregation*, and *design* references. Figure 4-top highlights these references in the context of sample utterances illustrating the different phrasing variations. Attribute references are essentially words in a query that map to a data attribute. Attribute references can be further divided into four sub-categories: (i) explicit references, where words in a query directly match a portion of a data attribute (e.g., *"mpg"* → MPG, *"genre"* → Major Genre), (ii) semantic references, where words in a query are synonyms or semantically similar to the dataset attributes (e.g., *"heavy"* → Weight, *"fuel economy"* → MPG, *"over time"* → Year), (iii) value-based references, where words in the query refer to cell values instead of column names (e.g., *"1995 to 2010"* → Release Year, *"furniture, office supplies, and technology"* → Category), and (iv) implicit references, where attributes are requested indirectly through a visualization type (e.g., if there is only one temporal attribute, requesting a line chart implies a reference to that attribute). Chart type references include explicit requests for specific visualization types (e.g., *"scatterplot"*, *"bar chart"*, *"histogram"*). A subset of utterances also included explicit, unambiguous references to graphical encoding channels that should be bound to different attributes (e.g., *"color by category,"* *"facet by origin"*). Utterances with aggregation references included one or more words that mapped to the type of mathematical transform that needed to be applied to create a chart (e.g., sum, average, count). Aggregation references were further sub-divided as explicit (e.g., *"total gross"*->SUM(Worldwide Gross), *"number of orders"* → COUNT) or implicit via a requested chart type or phrasing (e.g., *"histogram"* or *"How many"* → COUNT, *"stacked bar chart"* → SUM or COUNT). Finally, utterances with explicit design references specified additional details such as mark color, chart
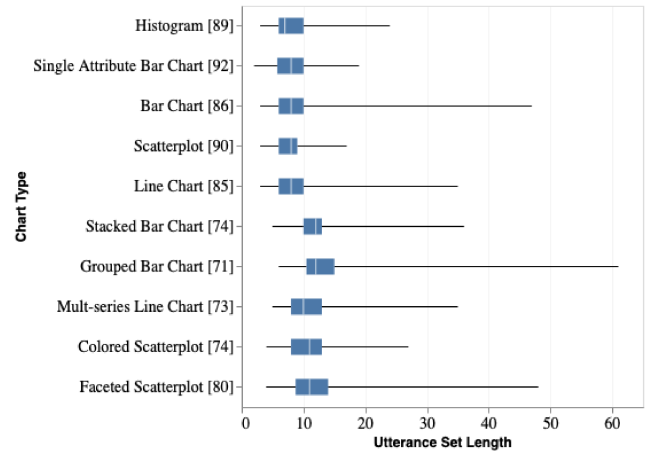

Figure 3: Summary of the lengths of the 614 utterance sets across the 10 chart types (Figure 1) and three datasets (Table 1) used in the study. The [count] alongside the chart types indicates the number of utterance sets.

orientation, sorting order, or axis tick windows that could be leveraged by visualization NLIs when rendering charts (e.g., *"set origin colors to: europe, blue, japan, orange, usa, red,"* *"Histogram of running time, in 20 minute bands"*).

Note that the aforementioned phrasing and information types are by no means an exhaustive list, nor are they mutually exclusive. For instance, while we tagged them as commands, one could cross-list the utterances *"Faceting on Origin, plot Weight by Acceleration"* and *"Can you create a graph showing sales and profit by region?"* as command+query and command+question, respectively. Similarly, many utterances included keywords like *'correlation,' 'distribution,' 'compare'* that could be used to identify intended analytic tasks [2], which is another potentially relevant factor for a system when determining which visualizations to render. Instead, the categorization described in this section is primarily to provide a subjective overview of the curated dataset and to serve as a reference point for more detailed discussions regarding the nature of utterances within visualization NLIs going forward.

5 APPLICATIONS

The central contribution of this work is the curated dataset of 893 utterances (814 utterance sets) that are publicly available at: https://nlvcorpus.github.io. To lay the groundwork and help foster ideas for future research and development, below we highlight some exemplary applications of this dataset.

5.1 Benchmarking Existing NL-based Visualization Tools

The curated utterances can serve as a benchmarking dataset to evaluate the performance of existing systems. As an example, we used the dataset to evaluate the NL4DV toolkit [18] which returns a list of Vega-Lite specifications in response to NL utterances. To benchmark NL4DV's performance, we configured NL4DV with the three datasets used for the study and executed the 755 singleton

Commands
- draw a line chart of daily sales forecasts (s) (e)
- Please show me a histogram of weights with 500 intervals. (i) (e)
- Show me a bar graph of the profit for each region > (e) (e) Make the bars stacked with the ship status (e)
- show me the average weight in a line chart (e) (e) (i)

Queries
- Cylinders average mpg (e) (e) (e)
- Scatter Production Budget vs Worldwide Gross > (e) (e) Slice by Content Rating (e)
- mpg vs displacement > as scatter chart (e) (e)
- Count by origin (e) (e)
- Gross versus genre (e) (e)

Questions
- What is our profit based on shipping mode by (e) (e) customer segment? (e)
- How much do various cars weigh? (s)
- How does displacement relate to fuel economy for cars (e) (s) from Europe v. USA? (v) (v)
- What major genre had the highest average worldwide (e) (e) gross? (e)

Others
- scatter(x=production budget, y=worldwide gross) for (e) (e) content rating (e)
- Trend for average horsepower over time across (e) (e) (s) different origin (e)
- 3 scatter plots, one each for every Europe, Japan, and (v) (v) USA, to show weight on X Axis and Acceleration on Y (v) (e) (e) Axis

		Attributes				Chart Types	Encodings	Design	Aggregations	
		Explicit (e)	Semantic (s)	Value-based (v)	Implicit (i)				Explicit (e)	Implicit (i)
Commands (368)	Singleton (336)	322	46	9	7	152	57	28	173	28
	Sequential (32)	32				22	17	3	15	2
Queries (260)	Singleton (249)	246	15	3		108	33	19	133	18
	Sequential (11)	11		1		8	2	1	6	
Questions (114)	Singleton (113)	105	32	3	1	1	2	1	28	28
	Sequential (1)	1	1						1	
Others (72)	Singleton (57)	56	9	1		36	23	1	24	6
	Sequential (15)	15		1		12	3	6	10	1
	Total (814)	788	103	18	8	339	137	59	390	83

(out of 570)

Figure 4: An overview of the curated dataset of 814 utterance sets composed of 893 utterances from the 644 trials. (Top) Utterance phrasing types we observed along with examples of each type. (Bottom) A distribution of information references within each utterance set. The color codes of the columns headers are also used to highlight the corresponding references in examples shown at the top. Note that specifically for attribute references, a single utterance set can have multiple types of references and hence the sub-category totals add up to more than 814 (e.g., "How does displacement relate to fuel economy for cars from Europe v. USA" contains an explicit reference to Displacement, a semantic reference to MPG, and a value-based reference to Origin). Additionally, data aggregations were only applicable to 570/814 utterance sets that were issued in the context of visualizations other than scatterplots (since scatterplots did not show aggregated values) and 97 of these 570 utterance sets did not include explicit or implicit aggregation references (e.g., "bar chart for mpg v/s cylinders").

Utterances Used	Accuracy (Exact)	Accuracy (Partial)	Execution Time (Avg.)
All (755)	485 (64%)	569 (75%)	1.04
Cars (290)	203 (70%)	225 (78%)	0.21
Movies (257)	168 (65%)	204 (79%)	0.15
Superstore (208)	114 (55%)	140 (67%)	3.30

Table 2: NL4DV's accuracy and execution time (in seconds) for 755 singleton utterance sets from the curated dataset. Note: results are based on a beta version of NL4DV 0.0.1.

Technique	Accuracy
Logistic Regression	88%
Random Forest	88%
Linear Support Vector Classification	87%
Multinomial Naive Bayes	85%

Table 3: Performance of different classification techniques using the 814 curated utterance sets. All numbers are reported based on a 10-fold cross validation.

utterance sets curated from the study. We only used singletons since NL4DV currently does not support sequential utterances.

Table 2 summarizes the toolkit's performance on the curated utterance dataset. Execution times for individual utterances ranged between 0.01s–14.11s (mean: 1.04s)[2]. To compute the accuracy, we compared the list of Vega-Lite specifications returned by NL4DV to the Vega-Lite chart shown during the study. In our first iteration, we checked for exact matches between the mark type (`point`, `bar`, `line`), encodings (`x`, `y`, `color`, `column`), data bindings (i.e., mappings between attributes and encodings), and data aggregations (`count`, `mean`, `sum`) with one exception: we marked two scatterplots as "equal" even if their `x`, `y` data bindings were interchanged. Inspecting the failed cases, we noticed that some mismatches were not interpretation or visualization recommendation errors but rather resulted from default data binding and aggregation choices made within NL4DV's recommendation engine (e.g., reversed `x` and `color` bindings for a stacked bar chart, setting a `mean` aggregation instead of a `sum` when no aggregation is specified in an utterance). Thus, to account for these default choices, we also computed a second "partial match" accuracy score ignoring such mismatches.

Looking through the remaining failed utterances, we noticed that errors occurred due to NL4DV's keyword-based task detection logic (e.g., the word 'relationship' would map to a *correlation*, resulting in a `point` mark while the utterance was in the context of a bar chart), and undetected semantic attribute references (e.g., the phrase '*fuel economy*' did not get mapped to the *MPG* attribute), among others. While not a critical review of NL4DV's performance, this example shows how the presented utterance dataset can help benchmark and identify areas for improvement within current tools.

5.2 Developing New Models for NL-driven Data Visualization

The curated utterance dataset can also be used to develop new techniques to generate visualizations from NL. As a preliminary application, we explored how the curated utterance dataset can be used to develop a classification model that predicts chart types (with labels being the ten chart types used in the study) based on NL utterances. To this aim, we experimented with different classification techniques including logistic regression, random forests, support vector machines, and Naive Bayes available within Python's scikit-learn package [19].

---
[2]Reported based on a MacBook Pro with a Dual-Core 2.6GHz processor and 8GB RAM running MacOS Catalina version 10.15.6

We summarize the classification accuracy of these models trained on TF-IDF vectorized representations of the 814 utterance sets (singletons + sequential) in Table 3. Although we need substantially larger datasets to make these models practically useful, this initial setup illustrates the potential of the presented dataset for supporting future development. Ultimately, such classifiers can be coupled with other NL understanding and visualization recommendation modules to generate visualizations in response to given utterances. For example, one could combine the output from a classifier (i.e., chart type), detect attributes in input utterances with toolkits like NL4DV [18], and then determine the most perceptually effective bindings for the classified chart type's encoding channels using tools like CompassQL [29] and Draco [17]. Similarly, the curated utterances can also be used as input to contemporary few-shot learning models such as GPT-3 [6]. For instance, developers have already begun to experiment with GPT-3 for generating visualizations from NL using only tens of hand-crafted example utterances [1]. To this end, the presented dataset can help fuel these recent efforts and significantly improve performance by providing a broader training set covering multiple chart types and dataset domains.

6 DISCUSSION

6.1 Implications for System Design

The spectrum of heavily specified utterances (e.g., *"Show me a bar chart of the total profitability of each region with a breakdown based on shipping time"*) to highly underspecified keyword-based utterances (e.g., *"Cylinders average mpg"*) highlights the varying level of expectations that people have from NLIs for data visualization. Closely examining the collected utterances can help current systems expand their underlying grammar configurations to support more natural utterances, or help develop new systems based on the themes observed from this data. As actionable takeaways, below we highlight some key design considerations for developing NLIs for visualization based on our observations from the curated set of utterances.

6.1.1 Accommodating natural phrasings as part of user input in visualization tools. While not an exhaustive or definitive list, the examples shown in Figure 4-top illustrate the diverse forms in which people naturally phrase NL utterances. However, given practical implementation challenges, current NLIs often suggest utterances as users provide their input (e.g., Figure 5), asking users to select from these utterances. While these suggestions can aid discoverability and help improve system accuracy, their phrasing may mislead people about the system's interpretation capabilities or bias users
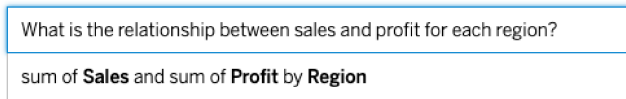
Figure 5: A phrasing suggestion in Tableau's Ask Data [27] while entering an utterance collected during the study.

towards phrasing their input in a specific format (e.g., query-like utterances as in the suggestion in Figure 5). Furthermore, mentally mapping different phrasings of input utterances and system suggestions can be tedious and potentially confusing (e.g., the input utterance in Figure 5 does not contain aggregations but the system suggestion does). Hence, going forward, to enable truly "natural" language-based interactions and support flexible input phrasings, one consideration for visualization systems is to explore ways to internally translate input utterances into system-friendly phrasings without exposing these phrasings to users by default.

6.1.2 Inferring different types of attribute references. Prior work by Grammel et al. [11] has highlighted how people, especially novices, heavily rely on attributes to specify visualizations (deferring underlying visualization design decisions onto the system). This observation has played a key role in motivating the design of NLIs for visualization as NL allows users (novices or experts) to freely specify attributes amidst naturally phrased questions or as part of targeted commands. The collected utterances not only reaffirm this fact but also highlight the flexibility that NL affords in specifying attributes (explicit, semantic, value-based, and implicit references in Figure 4-bottom). However, as also highlighted by prior work [10, 22], to be effective, NLIs must complement this input flexibility with robust interpretation techniques to infer these different references, particularly focusing on ambiguous (e.g., partial attribute matches) and semantic references (e.g., synonyms, domain-specific terminology such as *"fuel economy"* for MPG).

6.1.3 Balancing automated and manual view specification. Given the overarching task of visualization specification, an important aspect for supporting NL interaction is making automated visualization design choices when an utterance is underspecified (e.g., choosing chart types based on attribute types, selecting a default aggregation when there is none specified in the query) [22]. However, in this study, we observed that participants oftentimes provided explicit references to one or more encoding channels (e.g., *"color by Origin," "faceted by region"*) that may not match a system's inherent design choices (e.g., a visualization system might select a colored scatterplot over a faceted one by default). Yet, such custom specification is seldom supported in current systems that strive to provide agency through implicit decisions. Thus, in line with recommendations from Tory and Setlur on prioritizing explicit intent during analytic conversations [28], in the context of visualization specification, future systems must accommodate explicit encoding requests, allowing users to override implicit visualization design choices made by current systems.

6.2 Limitations and Future Work

Although our work presents implications for the general design of NLIs for visualization, the collected data and our findings should be interpreted with the study's constraints and assumptions in mind. Specifically, we only considered typed input and did not allow participants to use voice to enter utterances. Furthermore, we did not screen participants or collect any demographic information, and only distributed the study to participants outside the European region (due to privacy laws and IRB requirements). Hence, conducting follow-up studies to collect spoken utterances or targeting focused user groups (e.g., visualization novices) can help enrich the data and findings.

As with most online studies, cleaning and filtering the data collected from the trials was an integral aspect of our work. For instance, even after incorporating the feedback from the pilot sessions, we still encountered a breadth of trials that had to be discarded (Section 4.1). Furthermore, for individual trials, participants often entered two or more singleton utterance sets, one or more sequential utterance sets, or a combination of singleton and sequential utterance sets. While this highlights the input flexibility that the study interface provided, it also required us to manually inspect utterances and detect utterance sets. Given the relatively small scale of the data, performing these cleaning steps manually was feasible, albeit tedious. However, generalizing the data cleaning process for larger utterance datasets and investigating interactive data cleaning systems that function based on preconfigured rules or a set of examples are important areas for future work to explore.

Besides the immediate applications of the current dataset (Section 5), our study can also be extended in different ways. First, a short-term extension could be to consider additional visualization types (e.g., parallel coordinates, heatmaps) or focusing on non-tabular data (e.g., network visualizations, maps). Second, following the methodology used to construct popular NL-to-SQL datasets such as WikiSQL [33], a crowdsourced task could be setup to paraphrase the current set of utterances and then to validate the paraphrased utterances. This paraphrasing can help generate a larger dataset which could subsequently enable more quantitative analyses of utterances and the development of NL-to-Visualization models, complementing research trends in the computer vision (e.g., [3, 8]) and database communities (e.g., [31, 32]). Lastly, we asked participants to specify only one chart at a time. However, NLIs have the potential to offer a more fluid analytic experience which goes beyond specifying a single chart. To this end, future studies could combine ideas from the current study and prior research on conversational visual analysis [12, 28] to collect data on how people specify dashboards or sequences of charts (as opposed to specifying each chart from scratch).

7 CONCLUSION

We conducted an online study with 102 participants to collect NL utterances people use to specify data visualizations. To guide future research and development, we make the curated dataset of 893 utterances publicly available at https://nlvcorpus.github.io and illustrate its application toward the creation and benchmarking of NLIs for data visualization. In this paper, providing insight into the

nature of these utterances, we characterize them based on their phrasing type (e.g., commands, queries, questions) as well as the visualization specification-relevant information they provide (e.g., chart types, encodings, aggregation functions). Additionally, we present the system design implications of the observed utterance patterns and briefly reflect on design considerations for conducting subsequent online studies in the space. Finally, we discuss the constraints of our study along with future research opportunities that can complement our study, enriching the data and findings.

ACKNOWLEDGMENTS

REFERENCES

[1] Ken Acquah. 2020. Examples using GPT-3 to create data visualizations. https://twitter.com/aquariusacquah/status/1285415144017797126.
[2] Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*. IEEE, Piscataway, NJ, USA, 111–117.
[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Piscataway, NJ, USA, 2425–2433.
[4] Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. 2018. Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 594:1–594:8.
[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Westfield, NJ, USA, 1533–1544.
[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv:2005.14165 [cs.CL]
[7] Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang. 2019. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 906–916.
[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ, USA, 326–335.
[9] Siwei Fu, Kai Xiong, Xiaodong Ge, Siliang Tang, Wei Chen, and Yingcai Wu. 2020. Quda: Natural Language Queries for Visual Data Analytics. arXiv:2005.03257v2 [cs.CL]
[10] 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 ACM Symposium on User Interface Software & Technology (UIST)*. ACM, New York, NY, USA, 489–500.
[11] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 943–952.
[12] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2017. Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 309–318.
[13] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering Questions about Charts and Generating Visual Explanations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 340:1–340:13.
[14] Chufan Lai, Zhixian Lin, Ruike Jiang, Yun Han, Can Liu, and Xiaoru Yuan. 2020. Automatic Annotation Synchronizing with Textual Description for Visualization. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 316:1–316:13.
[15] Ronald Metoyer, Bongshin Lee, Nathalie Henry Riche, and Mary Czerwinski. 2012. Understanding the verbal language and structure of end-user descriptions of data visualizations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 1659–1662.
[16] Microsoft. 2020. Microsoft Power BI Q&A. https://docs.microsoft.com/en-us/power-bi/consumer/end-user-q-and-a.
[17] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 438–448.
[18] Arpit Narechania, Arjun Srinivasan, and John Stasko. 2021. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. *IEEE Transactions on Visualization and Computer Graphics* 27, 1 (2021), 10 pages.
[19] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12 (2011), 2825–2830.
[20] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 341–350.
[21] 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 ACM Symposium on User Interface Software & Technology (UIST)*. ACM, New York, NY, USA, 365–377.
[22] Vidya Setlur, Melanie Tory, and Alex Djalali. 2019. Inferencing underspecified natural language utterances in visual analysis. In *Proceedings of ACM Conference on Intelligent User Interfaces (IUI)*. ACM, New York, NY, USA, 40–51.
[23] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M Drucker, and Ken Hinckley. 2020. InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 653:1–653:13.
[24] Arjun Srinivasan, Bongshin Lee, and John T. Stasko. 2020. Interweaving Multimodal Interaction with Flexible Unit Visualizations for Data Exploration. *IEEE Transactions on Visualization and Computer Graphics* 27 (2020), 15 pages. https://doi.org/10.1109/TVCG.2020.2978050
[25] Arjun Srinivasan and John Stasko. 2017. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 511–521.
[26] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*. Springer, Berlin, Heidelberg, 184–195.
[27] Tableau. 2020. Tableau Ask Data. https://www.tableau.com/products/new-features/ask-data.
[28] Melanie Tory and Vidya Setlur. 2019. Do what I mean, not what I say! Design considerations for supporting intent and context in analytical conversation. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Piscataway, NJ, USA, 93–103.
[29] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Towards a general-purpose query language for visualization recommendation. In *Proceedings of the HILDA Workshop*. ACM, New York, NY, USA, 1–6.
[30] Bowen Yu and Cláudio T Silva. 2019. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1–11.
[31] Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. 2019. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. arXiv:1909.05378 [cs.CL]
[32] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2019. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. arXiv:1809.08887 [cs.CL]
[33] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. arXiv:1709.00103 [cs.CL]