Words of Estimative Correlation: Studying Verbalizations of Scatterplots

Rafael Henkin and Cagatay Turkay Member, IEEE

Abstract—Multimodal approaches where interactive visualization and natural language are used in tandem are emerging as promising techniques for data analysis. A significant barrier in effectively designing such multimodal techniques is the lack of a systematic understanding of how people verbalize visual representations of data. Motivated by these gaps, this paper devises and applies a transferable, semi-automated methodology to systematically study the relation between visualization and natural language through two crowd-sourced experiments and natural language analysis. We describe these experiments, analyze the resulting corpus of utterances with natural language processing techniques and derive an empirically supported semantic lexicon for aligning visualizations and verbalizations of data. Our results indicate a wide range of vocabulary used to describe visualizations and led to a number of high level concepts to categorize the space of words and related utterances. We discuss how our findings can be used for natural language generation, also reflecting on the design of the experiments and the semi-automated methodology used in the analysis. We discuss further research directions and argue for a role for such multimodal experiments in advancing our understanding of how people work with visualizations and also at large.

Index Terms—Information visualization, natural language generation, natural language processing, human-computer interaction

1 INTRODUCTION

In his seminal, originally classified report, Central Intelligence Agency (CIA) analyst Sherman Kent introduces “Words of Estimative Probability” — a group of terms that are “advisable” to use when intelligence analysts talk about the likelihood of future forecasts with some known/estimated probabilities [1]. In order to establish this set of words, Kent, together with his team, investigated how terms such as “almost certainly”, “probably”, “unlikely” are perceived and understood by gathering empirical data through controlled studies (see p.79 on “Handbook for Decision Analysis” by Barclay et al. [2]) on how the perception of these terms match numeric probabilities. The insight this research provided on the links between the intended and perceived semantics of probability has been fundamental in how intelligence analysts report critical pieces of information. Inspired by these studies, this paper, through a novel data-intensive methodology, aims to understand how visual depictions of correlation are perceived, understood and verbalized by their observers. By doing so, we also aim to establish a more general role for natural language-based experiments both in understanding the inner workings of visual representations of data and also in providing empirical support to inform the design of solutions that combine visual and textual representations.

Such multimodal interactive systems where visualization is combined with natural language are gaining increased interest both in research [3], [4], [5], [6], [7], [8] and in industry as evidenced by implementations in modern business analytic systems [9], [10]. Natural language has been shown to allow more expressiveness in interactions [6] and as a means to complement or reinforce visualizations in data exploration and analysis [4], [11].

In order to design such systems effectively, the process of understanding and/or generating language should be informed by user requirements and align with the domain of the underlying analytical tasks [12]. This very need is also highlighted for developing natural language generation (NLG) systems in the Natural Language Processing (NLP) literature [13]. Reiter et al. refer to the knowledge acquisition stage where researchers try to acquire “relevant knowledge about the domain, the users, the language used in the texts” as “one of the biggest problems in applied NLG” [12]. Methods such as structured interviews with experts and learning from collections of accurate text corpora are often the proposed approaches to gathering required linguistic elements for building a natural language generation system [12]. In visualization research, this need has often been addressed in a case-by-case basis [3], [6], not always resorting to empirically supported knowledge on how people perceive and understand the natural language information related to visualizations.

This paper is motivated by this need for a methodical understanding of how natural language is used to verbalize visualizations, and takes the first steps towards building such an understanding by developing and conducting a novel methodology to study the interaction between visualizations and verbalizations of systematically varied data relationships. To inspire and facilitate further studies, we devise a semi-automated methodology that involves user experiments along with data and natural language analysis methods that will serve as an extendable approach for studying verbalizations of data visualizations and for generating semantic lexicons and utterances to serve as the basis for the design of multimodal solutions.
We introduce and apply our methodology by investigating how the visualization of correlation relationships between two variables is verbalized by people. We are interested in correlation analysis as a fundamental task in data science and as a representative of cases where people reason with visual representations, namely scatterplots in this instance. We describe two crowd-sourced experiments aimed at gaining insight on the use of natural language with visualizations, drawing inspiration from studies that investigate the perception of systematically varied correlation in scatterplots [14], [15]. The analysis of results is done using a number of natural language analysis and machine learning techniques to semi-automatically analyze the resulting data.

The first experiment is a visualization to natural language study, in which participants are asked to provide verbalizations of a data relationship when viewing scatterplots for the varied levels of correlation. We then analyze the results to construct a semantic lexicon that maps higher level concepts, related to data visualizations, to the vocabulary used by participants. The analysis is described as steps that can be later used to expand the constructed lexicon.

We then describe a second experiment, on natural language to visualization, where we prepare utterances based on the semantic lexicon and ask participants to select scatterplots that match the utterances. We analyze how the utterances resonate with participants and compare the results of the two experiments, demonstrating how the data collected across the experiments can be used as empirically evidenced templates for natural language generation.

The results indicate a wide range of vocabulary used by participants to refer to and describe relations, which we then organize under five concepts that indicate “what” participants refer to, and five traits that outline “how” they refer to them. We observe that responses often involved multiple concepts and traits with prominent particular combinations, and distinctive use of them over the whole correlation range. The comparative analysis between the first and the second experiments’ results reveals word combinations that are represented consistently and would serve as reliable vocabulary to construct utterances for constructing “empirically evidenced” descriptions of visual representations. The analyses in Sections 4 and 5 and the following discussions in Section 6 elaborate these points further.

The contributions of the paper are thus summarized as:

- Empirical data, analysis and observations to advance towards an understanding of how people process and verbalize correlation in scatterplots;
- A transferable methodology for systematic studies of verbalizations of visual representations of data through crowd-based studies and natural language processing;
- A semantic lexicon and empirically evidenced utterances that can help structuring templates for natural language generation in multimodal interactive data analysis systems
- A research agenda for further studies that systematically investigate the verbalizations of visual representations of data.

2 RELATED WORK

2.1 Visualization and natural language

The use of natural language generation (NLG) to complement visual information has been investigated in several application areas, as well as in domain-agnostic settings. As authors often describe the development of multimodal systems as a whole, the language realization step is rarely discussed, with issues about the use of certain words often appearing during evaluation. Srinivasan et al. [16] combined the generation of data-driven statements with the exploration of alternative visualizations to facilitate visual analysis; the authors report on study participants emphasizing the need to understand what words such as moderate and strong meant for the system. A slightly different approach, less focused on natural language, was used by Demiralp et al. [17] when providing an insight-based visualization exploration system. Murtaz et al. [18], who introduced a multimodal tool to help with code quality analysis, acknowledged this by including tooltips to clarify the boundaries when values change from low to medium. Jain and Keller [19] also reported changing their textual health alerts based on feedback on the use of words such as significant. We also identified instances of authors using synonym lists to keep statements interesting [20], but in none of these cases is there a thorough consideration of the alignment of data, visualizations and natural language from a language perspective.

In addition to the importance of acquiring knowledge for this alignment, as emphasized in the literature (e.g., Reiter [12], [13]), we also motivate our work from the perspective of pre-design empiricism in visualization [21], which advocates gathering empirical evidence and documenting it before proceeding with designing multimodal tools. With this in mind, our contributions in this paper complement the described works through an open and reproducible methodology, that can be part of the pre-design process to improve NLG support.

Natural language is also used as an input in multimodal systems, as a means to provide alternative methods for hands-free interaction [3], [4] or as a complementary medium to traditional interaction modes [5], [6], [7], [8], allowing users to communicate with their own words rather than learning a potentially complex set of interactions. The methodology introduced in this paper also has a role in these cases, though in a more limited scope, to reveal the different manners through which users communicate about visualizations, which can then be adapted accordingly.

2.2 Perceptual studies

Our work is also related to the investigation of perception of visualization. Most prominently, researchers have long sought to investigate how humans perceive correlation in scatterplots, with various studies [22], [23] on exploring the ability to estimate varying levels of correlation. More recently, the information visualization community has approached this problem from a design perspective, focusing on the perception of correlation as a way to model the effectiveness of visualization design. Results from these studies showed how varying correlation for normally distributed variables can be modeled with a few parameters [14], and
also systematically demonstrated how different types of visualization are more or less suitable for these tasks [15], [24]. Beecham et al. [25] extended such studies into studying visual inferences from systematically varying levels of structure in visual representations of geospatial data. Sher et al. [26] reported results that brought new questions to the field, with results that demonstrated a difficulty in accurate estimations when the underlying data deviated from normal bivariate distributions. Correll and Heer [27] explored similar conditions when investigating how reliable are people’s estimates of trends and missing values. The idea in these studies, of systematically presenting varied levels of correlation, served as an inspiration for our experiments. In contrast to these studies, however, our experiments focus on the highly subjective application of natural language rather than the accuracy of estimations.

Other tasks were also tested regarding the depiction of correlation with scatterplots. Pandey et al. [28] investigated the perception of scatterplots based on judgment of similarity. Rather than modeling perception based on correlation, their aim was to understand which perceptual features of scatterplots were used when participants grouped them. The output of this study was a list of concepts that describe scatterplots, extracted from judgments of participants. Although their concepts partly overlap with our findings, our semantic lexicon includes a wider range of characteristics.

3 METHODOLOGY

As discussed in the introduction, the overarching goal of our study is to build a deeper and systematic understanding of how people verbalize correlation relations as seen through scatterplots. Through this investigation, we also aim to generate an empirically supported semantic lexicon [29], [30] to use as a basis for future multimodal interactive systems. To inspire and facilitate further research in this space, we devise a transferable, semi-automated methodology for conducting systematic studies of verbalizations of data visualizations. The methodology, illustrated in figure 1, is guided by two general objectives that relate verbalizations with visualizations and vice-versa:

**O1:** Determine the fundamental linguistic aspects of the association of verbalizations with visualizations;

**O2:** Identify the alignment between the association of verbalizations with visualizations and of visualizations with verbalizations.

In this paper, we use these objectives in the context of visualizing correlation relationships in scatterplots, however, it is important to note that the methods and analyses presented are neither limited to scatterplots nor correlation. Rather than pinning down preferences or accurate estimations through hypothesis-driven experiments, we design and conduct exploratory experiments aimed at collecting a wide range of verbalizations that will enable us to achieve the objectives. The methodology thus comprises two main experiments: **E1:** *Visualization to Natural Language*, in which we gather verbalizations of scatterplots of systematically varied levels of correlation between two variables, and **E2:** *Natural Language to Visualization*, in which we gather data on how participants associate scatterplots of varying correlations to natural language descriptions.

To achieve **O1**, E1 consists of three steps: (1) a user study to acquire utterances; (2) a semi-automated analysis pipeline, in which natural language processing (NLP) methods are alternated with manual coding steps, and (3)
Objective O2 is achieved through E2, in two steps: (4) a second user study, which asks participants to choose scatterplots that match statements describing a correlation relationship, and (5) the analysis and comparison between distributions of the selected scatterplots and the distribution of utterances matching concepts and traits from E1, based on the distance between histograms of these distributions. This second step also helps with the validation and refinement of the semantic lexicon. In addition to the analysis steps, Sections 4 and 5 also include descriptions of material preparation, participants and procedure for the two studies.

4 EXPERIMENT 1: VISUALIZATION TO NATURAL LANGUAGE

The first experiment involves asking participants to provide textual descriptions of scatterplots. The results are collection of utterances for the varying levels of correlation, from which a semantic lexicon is built. It includes three major steps of the methodology, which are the study itself, the analysis of utterances and the construction of the semantic lexicon by extracting the main concepts and traits used in the utterances. This enables completing the first objective (O1). We describe material preparation and task procedure, discuss details about the analysis methods used and present the semantic lexicon.

4.1 User study #1

Materials and participants: We generated 9 scatterplots (fig. 3) inspired by the scatterplots used in previous research [14], [15], [24]. They were generated by drawing samples from a bivariate normal distribution, with correlations ranging from -0.8 to -0.8, in increments of 0.2. Scatterplot figures are 288 pixels in both dimensions, with axes named for abstract variables A and B and ticks drawn for even numbers from 0 to 10. We recruited 160 participants through the Prolific online platform; our only pre-screening filter was self-declaration of native language as English. We used abstract variable names instead of real concepts (such as birth rate, GDP, etc.) so that participants would not relate more/less to the data, and any potential bias due to familiarity with the data is avoided. Moreover, this avoids additional layers of interpretation whilst the scatterplots are verbalized and the resulting data is analyzed.

Procedure: Each participant was asked to write an answer to the instruction “Please describe the relationship between the variables A and B in this chart” for each scatterplot, in a randomized order (see Supplement1 for the online surveys submitted as additional material). Before proceeding with the task, participants were given a description on the use of scatterplots to analyze the data – we note that this description, while it included the word relationship, did not include the term correlation. We carefully avoided biasing participants in this training by not providing any interpretation or description of the relation between the variables, and instructed participants to give complete answers and avoid answering with same as before or same.

4.2 Analysis

We collected 1428 answers out of 1440 potential answers; 11 participants from an initial test run saw 8 instead of 9 scatterplots and we removed one answer which was “n/a”. As part of the data cleaning process, we corrected spelling errors when sensible, such as “realtionship”, but did not change words such as “correlationship”. We carefully avoided biasing participants in this training by not providing any interpretation or description of the relation between the variables, and instructed participants to give complete answers and avoid answering with same as before or same.
on the other hand, our aim was to map concepts to answers rather than classify answers by concepts, thus we decided not to use classification or clustering techniques such as topic modeling. In the following, we go over the NLP analysis, followed by the derivation of conceptual categories before we move on to the semantic lexicon construction. All the steps described in the paper are accessible in the Python notebooks included in the supplementary material (also in the online repository at https://gitlab.city.ac.uk/nlvis/wec_supplement).

4.2.1 NLP pipeline

NLP processing: the first step taken was to replace occurrences of references to the variables A and B to identifiable proper nouns VarA and VarB. Following this, we proceeded with the tokenization (a process to break down text into individual words), part-of-speech (POS) tagging (marking up a word in a text as corresponding to a part of speech such as nouns, adverbs or verbs), and lemmatization (the process to group together the different inflected forms of a word into a single word, e.g., “am”, “is”, “are” being mapped to “be”) of each answer [31]. After a few tries with different tokenization methods, we decided to use regular expressions so that hyphenated words would be preserved. Tagging was done using the Stanford CoreNLP library [32] and lemmatization was done using the WordNet lemmatizer [33] applied only to the supported POS tags (nouns, adjectives and verbs). In the processed results, the average number of tokens per answer was 12.01; across correlations, the average number of tokens did not vary much, falling between 11.14 and 12.64 tokens.

Filtering and summarization: after the initial processing, the analysis proceeded with an exploratory stage to further understand the results, supported by other NLP-related methods. First, we excluded stop words and summarized the remaining words to find out which words and parts of speech were common and relevant in our data, to decide how to proceed with the exploration. The summary, shown in figure 4a, revealed that a large number of answers made direct references to the variable names, followed by variable, relationship (a word that was part of the instruction), correlation, increase, high and low. In the figure, words that might be different parts of speech are aggregated – increase includes both nouns (singular and plural) and verbs (third person singular or plural). Nonetheless, we identified a high number of nouns, which is in line with the focus of the rest of analysis.

Collocation analysis: as the next step, we analyzed collocations – word combinations that co-occur frequently [31], initially only with adjectives and nouns and focusing on bigrams and trigrams. Here, it is important to note that our analysis had a different final objective compared to traditional NLP scenarios where collocations are used. In these scenarios, distinguishing collocations are important to differentiate documents (or utterances). However, in our case, the objective was not to characterize individual answers, but to understand how language was used in these answers across dataset. From the resulting list of bigrams and trigrams, we extracted a list of the most common adjectives and compared their distribution across correlations to investigate how these words were used.

The analysis of the types of adjectives and how often they appeared provided an overview of how the varying levels of correlation were verbalized. Although not many adjectives seemed unexpected, such as positive and negative, our task procedure did not really suggest that people should focus on one or the other characteristic of the relationship, and the variety of adjectives reflects this. Another observation in our analysis was the low occurrence of some
potentially common adjectives, which suggested that collocational analysis was not enough as a method to use for categorization and counting of certain words.

### 4.2.2 Derivation of categories

The previous two steps indicated the types of words that participants used to describe the relationship between variables, in a general manner and also across correlations. This preliminary knowledge allowed us to proceed to capture higher level ideas that would be used in the lexicon. For this stage, rather than using automated approaches such as topic modeling or analyzing whole sentences and manually tagging them, we decided to create categories with the collocation analysis as a starting point. For that, adjectives and nouns were separated into two lists of groups: in one list, each adjective was paired with the groups of nouns that it appeared next to in a collocation; in the other list, the inverse was done: each noun was paired with the groups of adjectives that appeared next to it. The same procedure was done with adverbs and verbs.

Based on their functions, we organized nouns and verbs under concepts and the adjectives and adverbs that modify these concepts under the traits, which became the two axes in the resulting semantic lexicon [30]. Here, rather than starting with a complete bottom-up approach, we focused on names for concepts and traits from the continuum between data visualization and statistical relationships. This continuum includes words that describe elements of charts and visual channels – shapes, position, transparency, etc. – and words related to data/statistics – relationships, arrangements, characteristics of data. Our objective was to define categories that would be sufficiently distinct, but that could also potentially encompass sub-categories and be open for future changes. Several rounds of manual coding, refinement and discussion between authors were done until a final set of core concepts and traits was defined. Most of the resulting words were categorized in this manner – words that did not fit into our refined continuum were left without a category and deemed unhelpful for the lexicon.

The following concepts, summarized with examples in table 2, emerged from our analysis:

- **Relation**: words in this category describe the relationship between variables, including primarily relation and correlation. References to the variables A and B and the word variable are also included in this category, as they were used often in answers that define a correlation relationship (e.g. “as A increases, so does B”). This concept also includes link, association, connection, degree, level, proportion;
- **Behaviour**: words that describe arrangements or motion of data points, including trend, pattern, variation, rate, etc. This category also includes words that rather explicitly suggest direction, such as increase – the key difference between behaviour and direction (which is one of the traits we discuss next) is the part of speech, or how the word is used in a sentence: nouns and verbs are categorized as concepts, whereas adjectives and adverbs are categorized as traits.
- **Graphic**: words that refer to visual features, including dots, line, points, chart, graph, plot, space, curve;
- **Space**: words that refer to spatial distribution, grouping of items or measurement scale, including values, group, location, anomaly, cluster, outlier, distribution, concentration;
- **Inferences**: words that are used when the participant described additional information or gave meaning to the visual features, including confidence, data, results, information, likelihood, majority, meaning, significance.

These concepts appeared along the following traits, summarized with examples in table 1:

- **Magnitude**: mapped adjectives and adverbs in this category refer to strength or amount, including words such as strong, many, moderate, slight, weak, minimal;
- **Direction**: this category includes references to signals such as positive and negative, as well as upwards, downwards, increasing, rising, declining;
- **Position**: this category includes spatial positions, such as high, low or bottom, as well as descriptions of spread of distribution, such as random, widely or outside, among others.
- **Discernibility**: this category refers to how easy it is to identify a concept, including words such as clear, identifiable, discernible, noticeable, diffuse, distinct, vague;
- **Regularity**: this category describes the stability or the form of a concept, including words such as tight, linear, steady, direct, inconsistent, loose, uneven.

We additionally defined a special category Negate to categorize answers between affirmative and negative statements, as a quick glance over the collected utterances revealed the need to account for the negation of some of the concepts or traits, in sentences such as “no clear relationship”.

### 4.3 Lexicon construction

This process resulted in four lists of pairs of word and category, for adjectives, nouns, verbs and adverbs, which were used to map the concepts and traits to the answers. Each answer was tagged only once per occurrence of a concept or trait. This led to a further refinement of the categories: once a first round of analysis was done, the remaining uncategorized answers were analyzed and new words were added to the lists and categorized, until all answers had at least one category, either concept or trait – a final set of 8 utterances were not tagged as their words did not map to a concept or trait.

The iterative process also included prepositions, such as above or below, that were not captured by the collocation analysis. The complete set of tagged utterances range from one concept to long, multi-sentence answers that include up to 9 concepts or traits. The distribution of the number of answers per number of concepts and/or traits tagged, for the complete set of answers, is shown in Fig. 5. The distribution shows that, while the majority of answers referred to 2-4 concepts or traits, a high number of participants provided very long answers even if they were not asked to write a specific number of words or cover all aspects of the chart.

---

**Table 2**: Emerge from our analysis:
- Table 2, emerged from our analysis:
- Table 2, emerged from our analysis:
- Table 2, emerged from our analysis:
4.4 Analysis of concepts and traits

After arriving at a list of concepts and traits, we proceeded to analyze the categorized utterances according to the varying levels of correlation. As an initial overview, we looked at the combinations of concepts, traits and negative statements for the whole data, revealing 419 distinct combinations of concepts, traits or their negations, with almost half of them (210) being unique to a single answer. Fig. 6 displays the ten most common combinations of categories for the whole data; describing the absence of relationship with a negation was the most common combination (e.g. “no clear relationship”), followed by description of behaviour and inferences (e.g. “as A increases, B decreases”). The low number of answers for the most common combination (52), relative to the total number of answers (1428), demonstrates the diversity of combinations for describing the data relationships.

Figure 7 shows the co-occurrence of concepts and traits for the varying levels of correlations; each stacked bar corresponds to affirmative (in blue) and negative statements (in gold). As the categorization is a many-to-many process, each chart in the figure is potentially depicting utterances that appear in other plots as well; as such, the aggregated distributions of concepts and traits, above and to the left, are not the marginal distributions for each row or column.

As seen in the figure, certain concepts, traits or their combinations have representations concentrated on particular ranges of correlation. Two pronounced examples are direction and behaviour that are over-represented for more extreme values of correlation, i.e., bi-modal distribution with peaks around higher and lower correlation values. This indicates that participants identify strong correlations with their inclinations, e.g. positive, negative, increasing and how the points are arranged, e.g. trend, pattern, rate. Some others, such as magnitude or inferences, are used uniformly regardless of the correlation value. Such “uniform” categories indicate language that is regularly used by participants to describe the scatterplots. For instance, the magnitude of the correlations is often important to mention and participants

| Trait | Definition | Main words | Examples of Answers |
|-------|------------|------------|---------------------|
| Magnitude | Adjectives and adverbs in this category refer to strength or amount | Strong, many, moderate, slight, weak, minimal, ... | • There is a slight negative correlation • Linear, slopes downward, looks pretty strong |
| Direction | References to actual direction and also signal | Positive, negative, upward, increasing, rising, ... | • Weak positive correlation • Downward sloping correlation between 10A and 10B • Linear, sloping upward, not very strong |
| Position | Descriptions of spatial location or spread | High, low, bottom, random, widely, outside, ... | • The variables are reaching towards a higher point • Both start low and rise at about the same rate • Fairly random but correlated positively overall |
| Discriminability | Adjectives that refer to how easy it is to identify a concept in the scatterplot. | Clear, identifiable, discernible, noticeable, diffuse, distinct, vague, ... | • No clear relationship between the two variables • Relationship is random with no obvious pattern • Vague trend line • Both variables increase at a noticeable and consistent rate |
| Regularity | Stability or the form of a concept. | Tight, linear, steady, direct, inconsistent, loose, uneven, ... | • Both variables increase at a noticeable and consistent rate • An equal pattern but higher B reducing over A • More of a linear relationship where the greater A is the less A is |

| Concept | Definition | Main words | Examples of Answers |
|---------|------------|------------|---------------------|
| Relation | Words that describe the relationship between variables, including primarily relationship and correlation. | Link, association, connection, degree, level, proportion, ... | • There is positive correlation between A and B • Clear and definitive correlation in a diagonal line • There is a slight negative correlation • Appears to be a weak positive correlation • No clear relationship |
| Behaviour | Nouns and verbs that describe arrangements or motion of data points. | Trend, pattern, variation, rate, increase (verb form), decrease (verb form), ... | • There is a suggestion of exponential growth as the variables increase but there are several anomalies • There is a downward trend • There is no regular trend that I see |
| Graphic | Words that refer to visual features. | Dots, line, points, chart, graph, plot, space, curve, ... | • Both A and B have dots scattered mostly in the middle of the chart with a few random dots in the lower numbers • Looks like a random plot |
| Space | A broad category including words that refer to the spatial distribution of points, clustering or grouping or references to the axes/measurement scales. | Values, group, location, anomaly, cluster, outlier, distribution, concentration, ... | • The results appear to show a less concentrated spread of results with high points. It does not appear to have consistent results. • Seem to have a higher concentration of results near B5 than A5 • Higher values of A seem to result in a lower incidence of B. • Evenly distributed between A and B |
| Inferences | Words that are used when the participant described additional information or gave meaning to the visual features, as well as references to the variables A and B. | Confidence, data, results, information, likelihood, majority, meaning, significance, ... | • Low confidence in a trend of decreasing A with increasing B • There seems to be a predictable relationship between A and B except in certain points of the chart. |

TABLE 1
Definition, words and examples of answers for the traits identified through the analysis of results.

TABLE 2
Definition, words and examples of answers for the concepts identified through the analysis of results.
indicate inferences and try to interpret the “meaning” of the scatterplots irrespective of the strength of relations. An increasing use of negative statements for position, combined with inferences and relations, for the lower levels of correlation, is also noticeable. An example of utterance with this combination is “the relationship between A and B is random and there is no correlation”. Discernibility also contrasts with the other traits for its common combination with negative statements rather than affirmative (e.g. “no clear relationship”).

In addition to these prominent observations, we also identified what we consider a few peculiarities on how participants used some of the concepts. An interesting use of the concept of space was in describing a coordinate system, with the variables A and B as references. Examples of such utterances are “loose cluster of data points distributed around A6, B6 with a few wild outliers” and “scattered points concentrated most densely between 6A and 4-8B”. What stands out regarding this is that the scatterplots do not contain gridlines, and ticks are included only for even numbers (which at least explains why most of the grid references are indeed for the even positions).

5 EXPERIMENT 2: NATURAL LANGUAGE TO VISUALIZATION

This section describes the second experiment, in which natural language statements are presented to participants, who then select the scatterplots to which the statement applies. The results are the frequency of choices for varying levels of correlation, providing empirical evidence for natural language generation. As part of the overall methodology, the statements are prepared based on the combinations of concepts and traits synthesized in E1. As such, the experiment serves as a tool to validate and refine the semantic lexicon, as the results can be compared with the distribution of concepts and traits across correlations from E1. However, the statements and the charts themselves can also come from other sources, so the experiment can also be conducted independently. We describe the steps of material preparation, task procedure and analysis.

5.1 User study #2

5.1.1 Material preparation

For this experiment, we prepared 15 statements based on seven combinations of concepts and traits. Four of these groupings include the relation concept, which includes terms such as correlation and relationship, combined with four traits: magnitude, direction, regularity and discernibility – this was additionally combined with a negation. Another pair of statements was the combination of behaviour and direction, whilst the last two pairs of statements combine inferences with space and position. This resulted in the following statements:

(A) Relation and Magnitude:
- There is strong correlation between the variables A and B
- There is moderate correlation between the variables A and B
- There is weak correlation between the variables A and B

(B) Relation and Direction:
- There is a positive correlation between the variables A and B
- There is a negative correlation between the variables A and B

(C) Relation and Discernibility with Negation:
- There is a clear relationship between A and B
- There is no obvious relationship between A and B

(D) Relation and Regularity:
- There is a tight relationship between A and B
- There is a loose relationship between A and B

(E) Inferences and Position
- There is an upward trend
- There is a downward trend

(G) Inferences, Space and Position:
- When the values of A are high, so are the values of B
- When the values of A are high, the values of B are low

We would like to note that here we identified a set of representative categories and associated utterances to demonstrate how we conduct E2 and compare results from E1 and E2. Even though we aimed to select the most common category combinations and utterance structures, we used our judgment to identify utterances that we thought would resonate with participants, while covering a wide range of concepts and traits. An alternative and a much more comprehensive study could have selected utterances from all the most common combinations, as seen in figure 7 – in this case this would mean 25 distinct pairs of categories.

5.1.2 Participants and procedure

We recruited 120 participants through the Prolific online platform; our pre-screening filter was self-reported English as a native language and no participation in the first experiment. For the survey itself, we again used the Qualtrics system. The scatterplots from experiment 1 were positioned on a 3x3 grid and labeled sequentially Chart 1, . . . , Chart 9. A randomizer was set up so that, on average, each participant
would see 4-5 statements and thus each sentence would have been seen by a similar number of participants.

Each participant was asked to follow the instruction: "Please select all the charts that apply to the following statement", followed by the utterance. Before proceeding with the task, participants were shown a description about the use of scatterplots to analyze data, similarly to the first experiment.

There are three points in our procedure here that need highlighting as potential issues: participants’ lack of enough knowledge about statistics, lack of attention to the task at hand or lack of engagement with the experiment. We included one knowledge check question about correlation, which would not interrupt the task. After comparing the answers to the question with the overall answers, we decided not to filter any participants. This was also due to the fact that we are not judging accurate estimations and that some of the statements involved highly subjective interpretations. The answers to this knowledge check question are included in the complete data in the supplementary materials.

5.2 Analysis

We collected 550 answers from 120 participants. We first present and discuss the results of this experiment, and then compare them with the results of the first experiment and discuss the implications. In the following, we organize the discussions according to the categories of utterances introduced in Section 5.1.1.

**Relation and Magnitude:** this category included three utterances, with magnitude varying between strong, moderate and weak. For strong, the distribution for the second experiment is bimodal and symmetric, with peaks in the extreme values (right side in fig. 8), indicating that participants have an almost unambiguous association of strong with minimum and maximum values of correlation. For moderate, although the distribution is also bimodal and symmetric, it is not skewed and the peaks are located towards the mid-values of correlation. The shape of the distribution suggests that, when all scatterplots are seen together, moderate correlation is more ambiguous. Lastly, the distribution for weak is relatively uniform, except for the low occurrence in extreme values, indicating that even the absence of correlation is also associated with the word.

**Relation and Discernibility with Negation:** for this combination, utterances included one affirmative and one negative utterance. Overall, the results (top side in fig. 9) suggest that participants associate less ambiguously the existence of a clear relationship with the scatterplots that depict higher levels of correlation (0.6 and 0.8, positive and negative) and, conversely, no obvious relationship from -0.4 to 0.4.

**Relation and Regularity:** the utterances for this combination referred to tight and loose relationships. The distribution of results (bottom right side in fig. 9) show that participants strongly associate tight relationship with the extreme values of correlation with little ambiguity. In contrast, the distribution for the choices associated with loose is almost uniform, with very small peaks around ±0.6, suggesting that it is an ambiguous term. Although we did not ask participants regarding their views on the meaning of words, we also speculate that tight and loose are not interpreted as having exact opposite meanings in relation to a data relationship such as correlation.

The results for the three combinations described so far characterize the empirical support for terms for which the meanings are not directly linked to the underlying mathematical definitions or to the visual forms of the charts. The next four combinations include traits that have a closer association with the mathematical definitions of correlation, such as direction with positive and negative. This means that for such statements, it is possible to assign a correct answer. However, as our experiment was not designed to obtain accurate estimations, our analysis focus on the interpretation of the statements in relation to the scatterplots.

| Combination                        | N   | Examples                                                                 |
|------------------------------------|-----|--------------------------------------------------------------------------|
| NEGATE, RELATION                   | 60  | • "No correlation"                                                      |
|                                    |     | • "There is no correlation between the variables. This could be two variables that have nothing to do with each other" |
| BEHAVIOUR, INFERENCE               | 45  | • "B increases when A decreases"                                       |
|                                    |     | • "A and B appear to increase relative to one another"                 |
| DIRECTION, MAGNITUDE, RELATION     | 41  | • "Weak negative correlation"                                           |
|                                    |     | • "There is slight suggestion of a negative correlation"               |
| DIRECTION, RELATION                | 35  | • "Positive correlation"                                               |
|                                    |     | • "There is a negative correlation"                                    |
| INFERENCE, NEGATE, RELATION        | 32  | • "No correlation between A and B"                                      |
|                                    |     | • "There seems to be no predictable relationship between the variables A and B" |
| INFERENCE, POSITION                | 30  | • "When A is low, B is high"                                           |
|                                    |     | • "Concentrated in the centre of A above 4 on B"                       |
| BEHAVIOUR, DIRECTION, INFERENCE, MAGNITUDE, RELATION | 21 | • "There is a fairly strong positive correlation between variables. As A increases so does B" |
|                                    |     | • "Slightly positive relationship between variables A and B, fairly steep incline" |
| DISCERNIBILITY, RELATION           | 21  | • "There is no clear correlation between the two variables"            |
|                                    |     | • "Correlation is visible but not so definitive"                        |
| POSITION                           | 21  | • "Random"                                                              |
|                                    |     | • "Widespread"                                                          |
| SPACE                              | 19  | • "Scattered"                                                           |
|                                    |     | • "All over the place"                                                 |

Fig. 6. Most common combinations of concepts, traits and negative statements with examples of utterances. With exception of one combination, most other common combinations are for 2 or 3 concepts or traits. The lower number of answers, for the bottom four combinations in the table, reinforces the notion that participants used a wide variety of unique combinations of concepts and traits.
Fig. 7. Counts of co-occurrences of concepts and traits per level of correlation, with color separating affirmative statements (blue) from negative statements (in gold). The upper row and the leftmost column are the individual distributions of occurrences for concepts and traits, across correlations. In the central region, the charts show the co-occurrence of pairs of concepts and traits; the counted utterances may also include other concepts and traits. The figure shows the dominance of the combination of magnitude with inferences and also relation. As inferences include correlations. In the central region, the charts show the co-occurrence of pairs of concepts and traits; the counted utterances may also include statements (in gold). The upper row and the leftmost column are the individual distributions of occurrences for concepts and traits, across correlations. In the central region, the charts show the co-occurrence of pairs of concepts and traits; the counted utterances may also include other concepts and traits. The figure shows the dominance of the combination of magnitude with inferences and also relation. As inferences include references to the variable names, and relation includes references to “correlation” or “relationship”, the distribution matches well with the most common words identified (as seen in fig. 4a).
Relation + Magnitude

![Graphs showing iterative steps of semantic grouping for Relation and Magnitude.](image)

**Fig. 8.** Iterative steps of semantic grouping for Relation and Magnitude. The first groups are defined based on the intensity-driven word vector distances, with acknowledgment that it is not the optimal configuration. After step one, words are moved between groups or dropped until the distances between the two experiments, for each group, are subjectively acceptable. In the example, the best configuration of answers from E1 which are mapped to *Strong* is when the word *strong* itself is the only word left in the group. That is, it is the only time when the Hellinger distance between the distributions of E1 and E2, for *Strong*, is the shortest, based on the words elicited in E1.

**Relation and Direction:** this combination had two utterances, for *positive correlation* and *negative correlation*. The distribution of answers (bottom right fig. 10) for negative is right-skewed, decreasing across correlations. For positive, the distribution is left-skewed, being uniformly low from -0.8 until 0.4 and then sharply increasing at 0.6. We speculate that the differences between the results are likely the result of an anchoring or baseline effect, as participants saw all stimuli at the same time. However, further experiments would be needed to confirm such effect or clear up the differences between positive and negative.

**Behaviour and Direction:** for this combination, the statements mentioned either an *upward trend* or a *downward trend*. Unlike the previous combination, the results for both (top right side in fig. 10) seem to be more closely aligned, being right-skewed for downward and left-skewed for upward in a similar manner. We speculate that the difference between the categories is due to the fact that *upward* and *downward* are associated with the visual form rather than the mathematical definition of correlation.

**Inferences and Position and Space and Position:** in these two pairs of utterances, the difference is the use of the words *variable* and *values*. The results for both combinations are very similar (see fig. 11), with a high number of selections in the range of ±0.6-0.8. The two categories are more focused on the visual descriptions rather than an interpretation associated with degrees of correlation. For both categories, all scatterplots were chosen at least once; this stands out considering that some of the scatterplots clearly do not match the descriptions. Although we did not ask for justifications or timed the answers, we speculate that participants might have disengaged with the task given the longer form of these utterances compared to the other simpler statements.

### 5.3 Comparison with E1

The connection with the previous experiment through the statements and scatterplots enables different analyses based on the types of traits:

(i) for traits that have a more subjective interpretation, the comparison enables going back to the answers collected
in E1 and iteratively grouping statements and words based on their similarity – this is the case for the magnitude trait, in which words such as moderate do not have an anticipated meaning for correlation;

(ii) the comparison also serves to further validate the definition of certain traits, in particular for their use in NLG, complementing the overall methodology. This is the case for discernibility and regularity, which display differences between experiments.

The following approaches were used here to compare the results for these different cases; we also discuss the value and limitations of the experiment further in this section.

**Semantic grouping:** the semantic grouping involves matching words, identified from the utterances in E1, with the smaller set of words used in the statements in E2. This is required for traits for which there is no clear quantifiable distinction between adjectives. In the case of magnitude, for example, there might be clearer distinctions between strong, moderate and weak. However, deciding if an utterance that describes slight correlation corresponds to moderate or weak is more challenging. Such mapping is important in any case of using natural language in data analysis: if users describe slight correlation, the system must have ways to deal with it; if an output describes slight correlation, it is expected that a majority of users should interpret it appropriately.

There are several ways to do this. In this paper, we illustrate one data-driven method and use an existing vector space of words [34] where the distance between adjectives take the semantic intensity of the adjectives into account. This means that, in this space, vectors representing words such as weak and strong have a longer distance between them than in unweighted vector spaces. A 2D projection of this space, using the t-SNE algorithm, is used to determine initial groupings of words (included in the supplementary materials); iterative refinements of the groups, based on a quantified similarity (see below), are done until an adequate set of interchangeable words is found.

**Affirmative-negative matching:** for binary traits such as discernibility (clear or vague), utterances that include them in combination with a negative statement configure a pair of similar utterances. An example is vague relationship be-
ing similar to no noticeable relationship. For the comparison, distributions of both kinds of answers should be combined to better reflect the meaning of the utterances. Further down this section, this is what is done to compare discernibility and regularity traits.

Quantifying similarities: To compare the distributions of the answers quantitatively, we use normalized distributions of charts chosen by participants, for each of the nine levels of correlation across the statements. As the resulting distributions are discrete, we use Hellinger distance [35] to quantify the similarity of the distributions from the two experiments, where applicable. The distance is a metric that produces results from 0.0 to 1.0, where values closer to 0.0 mean that the distributions are more similar and values closer to 1.0 mean that the distributions are more different. The metric is also advantageous as it is nonparametric and distribution-free, therefore it does not impose any assumptions about the shape of the distributions that are being compared.

Relation and Magnitude: as described above, the semantic grouping step is required to be able to compare results of the magnitude trait between experiments. Figure 8 illustrates four iterations of magnitude groupings: the first step consists of swapping the initial assignment of few and slight, from moderate to weak and vice-versa. This results in a stronger group for weak, with the Hellinger distance between the experiments, which is already quite low at 0.1481, going down to 0.0968. However, the distance for moderate increases significantly to 0.2929. The next step to improve this is moving three words, broad, condensed and concentrated from strong to moderate, which has the effect of improving the distances for both groups. Lastly, we remove the word large, leaving strong itself as the only word for its category, with a final distance of 0.1166. Here it is important to emphasize again that the comparison aims at finding groups that are good enough, a highly subjective decision for which data-driven methods can help with.

The comparison of results indicate that strong is a suitable category, aligned between experiments, but it is also best represented by the word strong alone, without other potential synonyms found in E1. Moderate also aligns well between experiments and include some interchangeable words. Weak too aligns well, but it is quite ambiguous for mid-range and lower levels of correlation.

Relation and Discernibility and Relation and Regularity: the comparison of these two pairs aims at validating the definition of the traits and giving insight into how the traits vary between the two scenarios – visualization to verbalization and verbalization to visualization. For discernibility, from the answers of E1, statements were matched to clear or vague, for vague, negated statements for clear were also grouped together. The Hellinger distances between the distributions are 0.2917 for clear relationship and 0.1260 for no obvious relationship, validating the choice for the traits. The slightly higher distance for clear relationship suggests that the absence of a relationship is a stronger concept than the existence of it, in a visualization to verbalization scenario.

For regularity, the distributions of answers from the first experiments were combined in the following manner: answers mapped to “loose” were combined with the negated utterances for tight, whilst answers mapped to tight were combined with negated utterances for loose. The Hellinger distances of 0.2554 and 0.2514 for loose and tight, respectively, validate the choice for the trait. However, as discussed above, a tight relationship is a much less ambiguous concept than a loose relationship in the verbalization to visualization scenario. As the second experiment only covered the word “loose”, further experiments focused on linguistic aspects would help to clarify the differences.

Relation and Direction and Behaviour and Direction: For these two categories, the comparison is again aimed at both validating the traits and understanding the differences in the scenarios. For the first combination, the distances between the distributions are 0.2509 for negative correlation and 0.2444 for positive correlation. The main difference between results is the shape of the distribution for positive correlation, which in the second experiment is concentrated in 0.6 and 0.8, whereas in the first experiment, participants identified the scatterplots from 0.2.

For behaviour and direction, the distances are slightly lower, with 0.1555 for downward trend and 0.1691 for upward trend. As discussed in the previous section, it is possible that the small difference between the two categories relates to the fact that upward and downward are linked to the visual form, but the data collected do not support further conclusions. From the results, both combinations seem appropriate for describing correlation in scatterplots.

6 DISCUSSION
6.1 Study design
We presented two possibilities of experimental setups for achieving our main goal, each addressing a different objective. For the first objective, determining the fundamental linguistic aspects of the association of verbalizations with visualizations, we devised an experiment aimed at collecting utterances without biasing participants in particular directions.
or language. This meant, for example, choosing a dataset with abstract attributes. While this fulfills our initial aim, it does not reflect real-world situations, which could include analysts dealing with either familiar or unfamiliar data. It is possible that with such data, participants would refer to the variables more often and would try to include additional interpretations of the scatterplots.

For the first experiment, an important aspect is participants’ knowledge about correlation, which affects the variety of utterances and the related concepts and traits. Being the first published lexicon of this kind, we argue that it is preferable to have a wide range of vocabulary representative of a diverse cohort. Therefore, we did not restrict participants based on any criteria besides self-reported English as a native language. We also argue that knowledge or attention checks during the experiment would likely influence participants’ answers. We recognize that some of the observations and language would be different for different cohorts, that might be more or less numerate. In future work that expands the lexicon for specific types of visualization or that targets a narrower set of participants, a more rigorous screening procedure would be welcome.

For the second experiment, the task procedure is also important. Our setup presents a small number of statements to each participant, each at a time, from the whole pool of statements. One of the aims in doing so was to collect results that would mirror the structure of results of E1, of answers per level of correlation, so that a comparison between experiments would be possible as part of the methodology. Variations of the presentation of the statements and scatterplots include, for example, showing one scatterplot at a time, to potentially minimize comparison between scatterplots.

In terms of overall design, an alternative study that would elicit a similar kind of information is matching statements to scatterplots via a proxy measurement, such as level of agreement. While this would work well for correlation estimation (such as those by Rensink [14]), for comparing results our preference was to ask for direct mappings between the statements and the scatterplots, aware of the fact that this hides potential uncertainties about the answers.

Another further question is whether a similar setup to the original “Words of Estimative Probability” study [1] could have been conducted here; that is, present participants with some words and ask them to enter a number for the correlation strength they think the word represents. Unlike the notion of probability, where there is a more widely understood link between concept and number, the concept of correlation is open to interpretation and a much less established understanding could be expected, and to further complicate matters, there is even inconsistency between different numerical metrics that measure correlation [36]. Motivated by these, we did not expect participants to represent their interpretations through a number or a level of confidence as we discussed above.

6.2 Context of the concepts and traits

The concepts and traits that we derived from our results are limited by our context of correlations in scatterplots. Although it is likely that, as higher level categories, they will generalize to other visualizations and relationships, it is uncertain how the same words would be used in different experimental setups. Nonetheless, as we discuss in sec. 7, our aim is to provide a baseline for future studies, which could extend or replace parts of our overall categorization and word mappings.

6.3 Employing a semi-automated approach

As outlined in the previous sections, the proposed methodology employs automated natural language processing and machine learning techniques to facilitate the analysis of the resulting data. Methods such as thematic analysis [37] could be considered more suitable for this setting, where the whole data set is coded and categorized by experts manually. With the crowd-based studies, however, conducting such thematic analysis on the whole set is much less feasible due to the large volume of responses and the automated steps reduce the volume of data to be manually analyzed to more manageable levels. And compared to the conventional settings where thematic analysis is applied, the data we operate on in our setting is much more structured and responses are much more targeted, making it suitable for automated analysis such as, part-of-speech tagging or n-gram extraction. The methodology and second experiment are also a way to provide a scalable alternative to identifying empirical support for words and concepts. A more conservative method would be to run E2 for as many words as needed; the aim of the comparison between experiments is to mitigate the need for that.

7 Research Roadmap

The research presented in this paper is a starting point towards the wider goal of understanding how natural language and visualization work together in data analysis contexts. In this section, we present a research agenda to facilitate the further exploration of this topic and we think that the experiment designs, computational analysis routines, and the resulting data, which are all available in an online repository (http://gitlab.city.ac.uk/nlvis/wec_supplement), will serve as stepping stones.

7.1 Verbalization as an evaluation methodology

We observe and demonstrate that our verbalization study into scatterplots provide unprecedented insights on how participants internalize the information conveyed by the scatterplots and how they then externalize this information through language. Systematically conducted verbalization studies have the potential to serve as a new lens to look into visualization designs directly, providing insights into how they work and how well they meet the intended design goals. Visualization researchers already work with natural language utterances of people using their designs collected through methods such as think-aloud protocols or interviews [38]. However, the scope of language used in data from such studies is broad and requires extensive analysis and interpretation from researchers. We argue that visual-verbalization studies will complement existing visualization evaluation methodologies, by providing researchers with systematically structured data on the language used by people reading and processing information communicated...
through visual representations of data. This should eventually point out to the limitations and strengths of the visualization designs. A future challenge to address here is to conduct such studies for more complex tasks while still maintaining the applicability of the natural language processing techniques as described in this paper.

Another potentially interesting avenue is to consider interaction, and how researchers can gather utterances from users that inform them on how well the interactive process is running in a visualization solution. One direction could be to investigate multimodal interaction interfaces, such as chatbots, as an experiment medium to gather data for the evaluation of interaction between systems and users.

### 7.2 Moving towards interactive systems

One motivation for this work is the increasing prominence of interactive, multimodal data analysis systems [3], [6]. In this paper we produce a semantic lexicon and empirically evidenced utterances to help structure such multimodal interfaces. Our approach here, however, is limited to one-off utterances, i.e. not part of a dialogue but rather as responses to a single stimulus. The structure of dialogues, how the interactive narrative flow could be designed is also a key consideration in the design of multimodal interfaces [39] and the semantic lexicon we produce could be considered a part of what Mitchell et al. refers to as “dialogue seed corpus” [40]. A future direction is to extend these visual verbalization studies to interactive settings where the sequential nature of interactive dialogues could also be captured.

### 7.3 Linguistic analysis and other languages

The scope of the analysis of utterances in E1 was targeted at understanding the connection between vocabulary and the varying levels of correlation, of which the implications for NLG-empowered visualization are more immediate. Although the methodology employs natural language processing methods, linguistic aspects related to sentence construction or the association with the varied backgrounds of participants were not examined. Investigating this perspective can be part of an investigation about the cognitive processes related to verbalization in data analysis contexts. Such an in-depth understanding, in particular if related to the theoretical frameworks emerging from cognitive science in similar vein to what Padilla et al. did within the context of decision-making [41], has the potential to inform designers on respecting and acting on cognitive limitations and strengths of viewers of visualizations.

Expanding the lexicon and the experiments to other languages beyond English is also an interesting path to take. Given the different ways of constructing words and sentences, such as declensions in German, it is likely that a wholly different analysis process would emerge, as well as other combinations of concepts and traits. As related cultural and educational aspects would also be part of this, it would be interesting to see how visual literacy and numeracy affect the language used.

### 7.4 Visualization and data relationship path

Immediate follow-up experiments can be based on varying the visualizations or the underlying data relationship. The semantic lexicon produced in this paper can be enriched by new experiments, or used as a baseline for comparing how people respond to different stimuli.

**Types of visualization:** systematic variations of visualizations based on visual variables can be used to compare results with the baseline correlation in scatterplots lexicon. Experiments can be set up to investigate changes across point-based visualizations, such as dot plots, or with more different encodings, such as bar or line charts. Using one lexicon as a baseline will enable quantifying the differences in results and understanding how the language used by participants vary based on the encodings. For NLG-assisted visual analysis, acquiring knowledge about how concepts and traits overlap or differ across lexicons will help to generate statements that can be generalized to different types of visualizations.

**Types of statistical relationships and characteristics:** follow-up experiments can also vary the data relationship under focus or characteristics of the data, such as those characteristics governing cluster separation [42]. Proxy measures for scagnostics [43] can be used to simulate data for further studies, also enabling both comparison and expansion of the lexicon. As correlation studies are generally focused on bivariate datasets, investigating multivariate or univariate data, as well as temporal or spatial attributes, would also greatly expand the scope of the lexicon and the potential for applying the knowledge into a wider range of NLG-supported tools.

### 8 Conclusion

We described two experiments aimed at advancing the understanding of how people relate verbalizations to visualizations of correlation in scatterplots. We contribute with data, analysis and observations towards that goal, in the form of a collection of utterances and a semantic lexicon from a visualization to verbalization experiment. Our findings present a diversity of vocabulary for the different levels of correlation, with some concepts more prevalent than others in different ranges. We also show that that our organization of concepts into utterances also resonates with users in a verbalization to visualization experiment.

We conducted our analysis through a transferable methodology to support a research roadmap towards a further understanding on the use of natural language with visualization for other types of visualization, statistical relationships and languages, as well as arguing for a role for systematic natural language experiments as a means of evaluating visualizations and advancing multimodal interactive systems. The potential success of future work relies heavily on a multi-disciplinary thinking with concerted effort from researchers in visualization and linguistics, as well as those in cognitive sciences to relate to theories of cognition. With the eventual aim to build a comprehensive and multifaceted understanding of how visual and textual representations work in human reasoning and decision making, this strand of research is likely to open up new directions and further the impact of human-data interaction systems in our increasingly data-rich society.
ACKNOWLEDGMENTS

We thank Johannes Lien for the help with setting up the online studies and discussions. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) with grant number EP/P025501/1.

REFERENCES

[1] K. Sherman, “Words of estimative probability,” Central Intelligence Agency, Tech. Rep., 1964.
[2] S. Barclay, R. V. Brown, C. W. Kelly III, C. R. Peterson, and L. D. Phillips, “Handbook for decision analysis,” DECISIONS AND DESIGNS INC MCLEAN VA, Tech. Rep., 1977.
[3] A. Srivisan and J. Stasko, “Orko : Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks,” InfoVis TVC, vol. 24, no. SI, p. 11, 2017.
[4] A. Srivisan and J. Stasko, “Natural Language Interfaces for Data Analysis with Visualization: Considering What Has and Could Be Asked,” in EuroVis 2017 - Short Papers, B. Kozlikova, T. Schreck, and T. Wischgoll, Eds. The Eurographics Association, 2017.
[5] E. Hoque, V. Seltur, M. Tory, and I. Dykeman, “Applying Explanations and Principles for Interaction with Visual Analytics,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 309–318, Jan 2018.
[6] V. Seltur, S. E. Battersby, M. Tory, R. Goosweiler, and A. X. Chang, “Eviza: A Natural Language Interface for Visual Analysis,” Proceedings of the 29th Annual Symposium on User Interface Software and Technology, vol. 10, no. 16, pp. 365–377, 2016.
[7] Y. Sun, J. Leigh, A. Johnson, and S. Lee, “Articulate: A semi-automated model for translating natural language queries into meaningful visualizations,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6133 LNCs, pp. 184–195, 2010.
[8] T. Gao, M. Doncicheva, E. Adar, Z. Liu, K. Karaballois, and A. Azbor, “DataTone : Managing Ambiguity in Natural Language Interface for Data Visualization,” in Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, Charlotte, NC, USA: ACM, 2015, pp. 489–500.
[9] Tableau Software, “Tableau.” [Online]. Available: http://www.tableau.com
[10] Qlik Software, “Qlik Insight Bot.” [Online]. Available: https://wwwqlik.com/us/products/qlik-insight-bot
[11] R. Sevastjanova, A. Srinivasan, and J. Stasko, “Visualizations of Correlation Using Weber’s Law,” IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 3, pp. 1203–1210, Aug 2010.
[12] W. Livnat and G. Balbridge, “The Perception of Correlation in Scatterplots,” Computer Graphics Forum, vol. 29, no. 3, pp. 1203–1210, Aug 2010.
[13] L. Harrison, F. Yang, S. Franconeri, and R. Chang, “Ranking Visualizations of Correlation Using Weber's Law,” IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1937–1940, Dec 2014.
[14] A. Srivisan, S. M. Drucker, A. Endert, and J. Stasko, “Augmenting visualizations with interactive data facts to facilitate interpretation and communication,” IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 672–681, 2019.
[15] c. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati, “Foresight: Recommending visual insights,” Proc. VLD Endow, vol. 33, no. 2, pp. 1937–1940, Aug 2017.
[16] H. Mumtaz, S. Latif, F. Beck, and D. Weiskopf, “Exploratorian Code Quality Documents,” IEEE Transactions on Visualization and Computer Graphics, pp. 1–1, 2019.
[17] A. Jain and J. M. Keller, “Textual summarization of events leading to health alerts,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2015-November, pp. 7634–7637, 2015.
[18] S. Latif and F. Beck, “Interactive map reports summarizing bivariate geographic data,” Visual Informatics, vol. 3, no. 1, pp. 27–37, 2019.
Rafael Henkin received a PhD in Computer Science from City, University of London. He is a Research Associate in Visual Data Science at the giCentre at the Computer Science Department of City, University of London. He is interested in understanding how information visualization, machine learning and human-computer interaction can be combined in the data analysis process, and which methods facilitate the integration from a design perspective and enable scientists to apply them in practice. He has been a reviewer for journals such as *IEEE Transactions in Visualization and Computer Graphics* and conferences such as *INTERACT*.

Cagatay Turkay is an Associate Professor at the Centre for Interdisciplinary Methodologies at University of Warwick. Before that he was a Senior Lecturer in Applied Data Science at the giCentre at the Department of Computer Science at City, University of London. His research focuses on designing visualisations, interactions and computational methods to enable an effective combination of human and machine capabilities to facilitate data-intensive problem solving. He serves as a committee member for several conferences including VIS and EuroVis, and part of the organising committee for IEEE VIS on 2017, 2018 and 2019. He served as a guest editor for *ACM Transactions on Interactive and Intelligent Systems* and *IEEE Computer Graphics and Applications*, and an editorial board member for *Computers and Graphics* and *Machine Learning and Knowledge Extraction* journals.