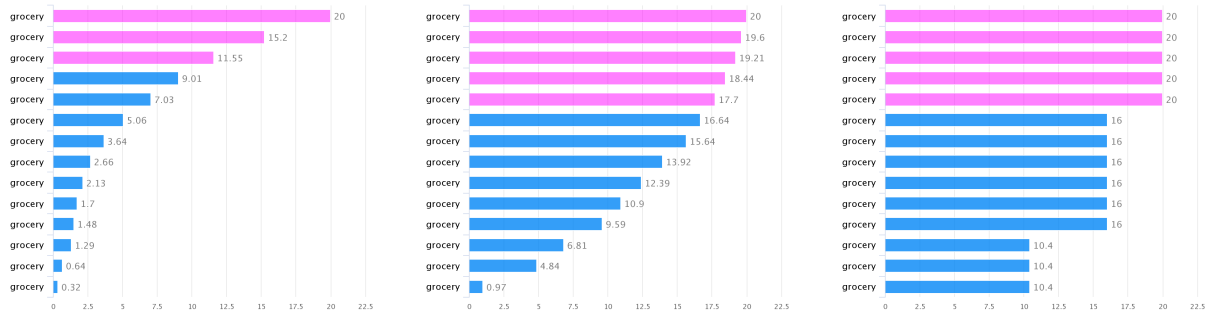
Toward Interface Defaults for Vague Modifiers in Natural Language Interfaces for Visual Analysis

Marti Hearst*
UC Berkeley

Melanie Tory†
Tableau Software, Inc.

Vidya Setlur‡
Tableau Research.



Shape 1: Exponential dropoff | Shape 2: Inverse exponential | Shape 3: Plateaus

Figure 1: The stimuli shapes (distributions) used in the experiment, showing the Highlight Many view. The other possible views were Highlight One (same view, but with only the top item highlighted), Filter Many (only highlighted bars shown), and Filter One (only the topmost bar shown). Participants choose their preferred view in response to seeing one (shape, question) pair. Questions started with: "Which of my grocery expenses..." and completed with one of: (Single Superlative) "is the highest this month?" (Plural Superlative) "are highest this month?" or (Ambiguous Modifer)"are high this month?"

Abstract

Natural language interfaces for data visualizations tools are growing in importance, but little research has been done on how a system should respond to questions that contain vague modifiers like "high" and "expensive." This paper makes a first step toward design guidelines for this problem, based on existing research from cognitive linguistics and the results of a new empirical study with 274 crowdsourcing participants. A comparison of four bar chart-based views finds that highlighting the top items according to distribution-sensitive values is preferred in most cases and is a good starting point as a design guideline.

Index Terms: Human-centered computing—Visualization;

1 Introduction

There has been a recent surge in research in natural language interfaces for visual analytics tools. Gao *et al.* note [3] that natural language questions offer a compelling alternative since the user does not have to learn a sophisticated query language to build charts and graphs. Recently released commercial products such as ThoughtSpot and Tableau's Ask Data allow users to type questions about data and see results expressed as visualizations, with inferencing to handle underspecification [20]. However, little research has been done on how users might express uncertainty or vague intent via language to an information visualization system, and how a system should best respond.

Vagueness appears throughout language. Although it can be convenient to think in terms of extremes, many concepts are expressed on a spectrum, from unsafe to safe, from expensive to cheap, from better to worse [8, 16, 17]. Research from linguistics puts forward evidence that people deliberately use imprecise language as a way to better communicate. They do this for many reasons, including [25]: (i) to avoid error, (ii) because of the absence of a mutually understood metric, (iii) to reduce cognitive effort, or (iv) because precision is not relevant.

One of the most prevalent kinds of vague language are adjectives and other modifiers. The use of vague modifiers in expression of questions may signal vagueness in underlying intent. Consider these variations of a question:

(a) What are the *available* apartments?

(b) What is the *cheapest* apartment? (singular superlative)

(c) What are the *cheapest* apartments? (plural superlative)

(d) What are the *cheap* apartments? (plural graded adjective)

A system that interprets questions of this sort must determine the underlying intent. Does the questioner of (a) really want to see all available apartments if there are hundreds? Does the questioner of (b) really only want to see the single apartment that costs the least (as implied by the use of the superlative modifier), or do they also want to see some other apartments for context? Question (c) is only slightly different from (b), but its use of the plural verb seems to imply a desire to see more than one apartment – but how many? A reasonable assumption is fewer than (a), but how far down the list of apartments is appropriate? And finally, does (d)'s use of a graded adjective with a plural verb differ from (c) in terms of which results to show, and how to show them? Another consideration is that the questioner may not *literally* mean the lowest price, but may also mean within the bounds of decency for apartment rental, which can be an implicit requirement when shopping.

This paper consists of:

- An introduction to the problem of vague modifiers in natural language interfaces for visualization systems.
- A brief discussion of relevant empirical work from cognitive linguistics.
- An analysis of a previously conducted query elicitation study

*e-mail: hearst@berkeley.edu; This work was conducted while Hearst was a visiting researcher at Tableau Research.

†e-mail: mtory@tableau.com

‡e-mail: vsetlur@tableau.com

showing the prevalence of such modifiers.

- Main contribution: The first empirical study comparing visualizations in response to questions containing vague modifiers.

2 Related Work

2.1 Natural Language Interfaces to Visual Analytics

This paper is the first to consider the questions of imprecise modifiers for natural language interfaces to data retrieval interfaces that show visualizations in response. Handling vague modifiers, which involves determining the degree of match, is a different problem than disambiguating different competing meanings of interpretation. The research systems Eviza [19], Evizeon [5], and Orko [24] provide natural language interfaces and mixed initiative interaction to help users refine an existing visualization. The DataTone system [3] combines a sophisticated semantic parser with user interface widgets to help disambiguate complex natural language queries. However, none of these systems consider the issues surrounding vague or uncertain language.

Although there has been extensive work in natural language interfaces to database systems, those systems show results as tables of information, not visualizations, and the query languages do not allow imprecise modifiers that do not match the query schema [1, 14, 15]. There has also been work on disambiguating ambiguous queries in that literature, and this was also a focus of the DataTone work [3].

2.2 Cognitive and Computational Linguistics

In this work we focus on adjectives such as "expensive" and "tall" and other modifiers that indicate superlatives (on either end of the scale) such as "most," "least," and "best". Researchers from cognitive linguistics, and computational linguistics have contributed insights into the behavior of such adjectives. They exhibit several properties that pose challenges to natural language interfaces:

Gradedness: Many adjectives are considered to be graded, that is, they can be interpreted on a scale (e.g., from more expensive to less expensive). The interpretation is both context-specific and sensitive to the distribution of the values [8], and both relative and absolute as in this example from Kennedy [7]:

> Kyle's car is an expensive BMW, though it's not expensive for a BMW. In fact, it's the least expensive model they make.

Experimental evidence has shown that people are sensitive to the relative distribution of items being compared by a modifier [16, 17, 23]; this is discussed in more detail below.

Antonyms vs. "Not"-Adjective Antonyms behave (slightly) differently than not-adjective. For instance, the set of items labeled "not expensive" may be different than those labeled "cheap" [22].

Components: Certain adjectives have multiple attributes associated with them (e.g., cheap can be defined by both price and quality while tall is solely height) [8, 12, 27].

Subjectivity: Some adjectives are more subjective than others, and this can be determined empirically (at least out of context). Scontras *et al.* [18] compared two different human-rating tests, and found a strong correlation between the two ($r^2 = .91$).

Computational linguists have shown how to automatically infer adjective grade ordering from very large corpora [2] and how to automatically detect subjective adjectives [26]. These results can be used in applications, such as for identifying which concepts within medical records are modified by gradable adjectives and mapping their meaning into more actionable interpretations [21]. Computational linguists have also produced results on automatically analyzing sentences with vague modifiers [9, 10, 13].

2.3 Judgments Vary Based on Data Distribution

Studies from cognitive linguistics show that human judgments of gradedness of adjectives vary depending on the distribution, or shape, of a set of presented data items.

most	117	biggest	15	largest	5	expensive	3
highest	32	best	10	new	5	high	3
more	28	lowest	8	worst	4	bigger	3
top	21	greater	7	long	4	well	3
last	16	least	6	far	3	higher	3

Table 1: Most frequent vague modifiers from a query elicitation task.

Schmidt *et al.* [17] experimented with the word "tall", assessing a wide range of distributions, sampling both randomly and at regular intervals from one or multiple distributions of different types, including Gaussian, uniform, and exponential. Solt & Gotzner [23] studied four adjectives and four manually determined distributions. For both studies, crowd worker participants were shown bars of differing heights, organized into a matrix, and presented in random order. They were asked to mark which items were considered "tall" and leave the rest blank. Qinq & Franke [16] replicated most of Solt & Gotzner's study.

The authors of all three studies used cognitive modeling techniques to model the results, but no one method fits all distributions. Viewers were quite sensitive to the distribution of the relative sizes of the items, how many appear in plateaus adjacent to one another, and if the relative values formed a convex or a concave shape when placed in sorted order. Interestingly, the responses for different adjectives in the latter two studies were quite similar to one another, suggesting that for this task at least, only the relative sizes of the visual items mattered.

These studies were deliberately designed to remove external contextual clues; rather than compare, say, basketball players' heights, participants judged the height of fictitious objects. Furthermore, the bars were not assigned numerical scores nor were they shown against an axis. Finally, the items were shown in random order. Therefore, these results might not generalize to real-world usage in a visualization interface.

The cognitive linguistics results suggest that a rigid design guideline, such as "show the top item" or "show the top 5 items" is likely to result in a user interface that violates users' expectations. We hypothesize that a more flexible approach, that is responsive to the shape of the data, and that allows room for error and disagreement around the boundaries of the proper cutoff, is more likely to succeed.

3 Vague Modifiers from a Query Elicitation Task

As part of a study to assess natural language input to visualization systems [20], participants were asked to look at the metadata for datasets and write natural language queries that to ask of the underlying datasets. 75 participants wrote 578 natural language queries from 5 different metadata profiles (bird strikes, world indicators, superstore, mutual funds, and Olympic medals). The words in the queries were lowercased and tokenized and compared to the adjectives in WordNet [11]. Words that are not superlatives or vague modifiers, in the usage described above, were manually removed.

Table 1 shows the most frequent of these and their frequency within the collection as a whole. Example queries include "Is there a destination that has the most shipping problems?", "Which country has more female medalists?", and "Which aircraft model has the highest repair cost?" This data shows that when participants were not restricted in the format of expression, they often chose to use vague modifiers to state their underlying intent. Presumably, they write the query in a vague way in the hope that the system will handle the meaning of "more" and "highest."

4 Experiment

No guidelines exist for the appropriate way to show information visualization results in response to questions that contain vague modifiers like "most" or "cheap." The default for superlative questions for a commercial system (Ask Data) is to show only the top item.

	Shape 1	Shape 2	Shape 3
Singular Superlatives:	H1	none	none
Plural Superlatives:	HM	none	HM
Ambiguous Modifiers	HM	none	HM

Table 2: Hypothesized outcomes. None = no strong agreement.

However, it is unclear if users wish to see only a subset of results to which the modifiers apply, or perhaps the full results with a subset highlighted, or some other option.

Our goal is to probe some of the boundaries of what an appropriate interface is for responding to queries that contain imprecise modifiers, with the ultimate goal of formulating sensible defaults for natural language interfaces to visual analytics tools. As a step in that direction, we performed a crowdsourcing study that presented participants differing views for (question type, distribution type) pairs and asked their preferences.

Objective measures in terms of speed or accuracy are not applicable for this study, since our goals are to determine what people think is the most *appropriate* response under these circumstances. A good outcome, therefore, is if there are conditions under which there is strong agreement among participants, since this can lead to reliable design guidelines.

4.1 Question Types

In the experiment, we compare question types (b), (c), and (d) of Section 1. We use question type (a) to orient the participants, as a baseline for what the results look like without a vague modifier. Our hypotheses are that:

- A question of type (b) will result in a preference for a different visualization than for (c) and (d); that is a question for "What is the highest one?" will in general have a different preference than "What are the highest ones?".
- Questions of type (c) and (d) will result in similar preferences; that is, a question for "What are the highest ones?" and "What are the high ones?" will result in a preference for the same type of view.

These are further refined by shape and view type below.

4.2 Data Shapes

As discussed in Section 2, prior work has shown that people's judgments of which items correspond to which adjectives (i.e., which out of a set of bars are "tall" and which are not) are highly influenced by the other bars shown when asked to judge in a visual comparison context. Since visualizations using bar charts are similar to the contexts of this prior work, we expect similar sensitivities to arise.

We adapt distributions from Schmidt *et al.* [17][1] (see Figure 1):

Shape 1: A roughly exponential dropoff. The first three items were marked as tall by nearly all participants, with the rest marked not-tall. *Hypothesis:* because there is a distinguishing largest item, participants will want different views for singular superlative vs. plural/graded adjective. Because there was strong agreement on tall vs not-tall, there will also be strong agreement for the best views for (c) and (d). Note however that Schmidt *et al.* [17] did not ask about "tallest," so this hypothesis is an extrapolation from their data.

Shape 2: A roughly inverse exponential curve, with no clear visual markers as to where to distinguish between tall and not-tall. *Hypothesis:* participants will be uncertain as to which design is best and so will show disagreement for all question types.

[1]Distributions adapted from Figures 3c right, 3c left, and 4a right. We adjust these slightly, doubling the length of Shape 1 and 2, and shortening 3 so they all have the same number of items. All stimuli begin with the value 20. To keep values within meaningful ranges, we reduce the dropoff of the last 6 values of Shape 1, but retain its essential shape.

Shape 3: A first plateau of items followed by an 80% drop to a second plateau of items, followed by a 65% drop. About 60% of participants agreed that the first drop was the end of tall, but some noise occurred about how far along the first plateau to mark as tall. *Hypothesis:* disagreement about designs for the singular superlative but agreement for plural and graded adjective.

4.3 Visualization Alternatives

We compared four different visualization views: two filtering options and two highlighting options. Filtering refers to reducing the set of bars shown, while highlighting refers to visually annotating the bars to draw attention to those that merit special consideration. Four views were designed for each data shape as follows:

F1 (Filter One) Filtered to top item (most or least)

H1 (Highlight One) All items, with the top one highlighted

HM (Highlight Many) All items, with the top several highlighted

FM (Filter Many) Filtered to top several items

The views were shown all on one page, but in randomized order. Participants were asked to select the one they would most prefer to see in response to the question. Our hypotheses for which views will be preferred by question type and shape are shown in Table 2. The number filtered or highlighted in the HM and FM views were based on Schmidt *et al.* [17]. In the filtered view, a bar chart is shown with only the selected bars, in the blue color of Figure 1, with no highlighting.

4.4 Experiment Design

Participants were recruited from the Amazon Mechanical Turk crowd sourcing platform, as done in much contemporary research in information visualization (e.g., [4, 6]). Participation was to English speakers in the U.S. with at least a 95% acceptance rate and 500 approved tasks. We payed a rate equivalent to $1.50 for 10 minutes of effort. The stimuli did not require excluding participants for color deficiencies.

The experiment was a 3 (data shapes) x 3 (question types) design, with each trial having 4 possible views. Because we are seeking subjective responses, participants can complete only one trial, to avoid biases that might arise from repeated exposure to the task. We used financial applications (credit card bill balance, someone monitoring their grocery store budget) since this is likely to be understandable to a broad population. The experimental procedure was:

- Training task: Credit card bill example, with practice questions to ensure understanding of the meaning of bar charts.
- A new page, the task description and question (a): "Show me how much I spent on groceries this month." In all cases, the participant is shown a view of the full data shape on this page.
- A new page, and a second question, one of type (b-d). Question format is "Which of my grocery expenses" and then one of (b) "is the highest this month?" (c) "are highest this month?" or (d)"are high this month?" Participant is shown a scrollable list of four different views of the distribution, shown in randomized order. Participant must choose top preference.
- Free text response for reason for preference.

Details are available in the supplementary materials.

4.5 Results

Data for 274 participants was gathered; nearly evenly spread across shapes (91, 94, and 89) and question types (91, 91, and 92). Overall, participants preferred to see information highlighted within bar charts rather than filtered to the top few (see Figure 3). Only 20% preferred filtering in either format. This suggests that highlighting results in bar charts is a better interface strategy than filtering.

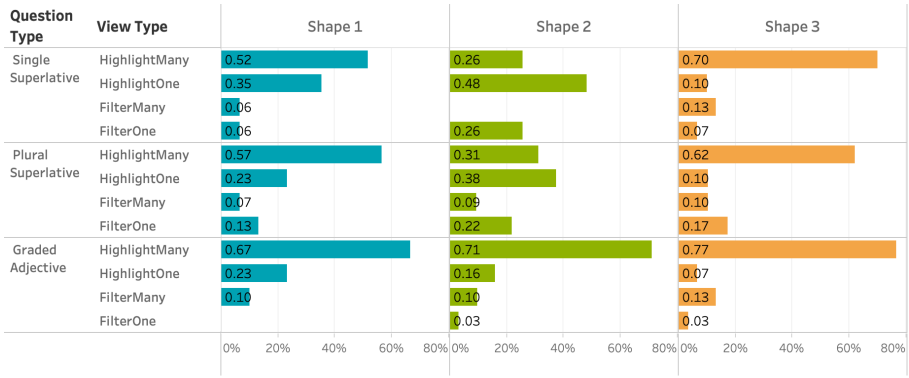


Figure 2: Participant preferences, in percentages, by shape, question, and view. As hypothesized, Shape 2 tended not to have lower agreement among views and Shape 3 had higher agreement than Shape 1.

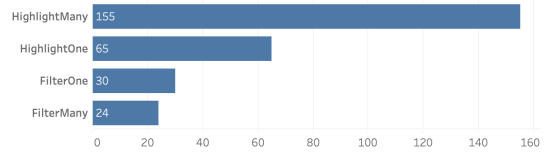


Figure 3: View preferences, across participants, question types, and shapes. Highlight Many is most preferred, by 56% of participants. Highlight One is second most preferred, by 24%.

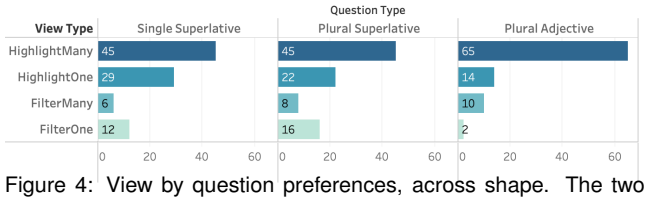


Figure 4: View by question preferences, across shape. The two superlative question types behave similarly to one another overall. Filter One is chosen 13% and 17% of the time for Single Superlative and Plural Superlative, respectively, but only 2% for Plural Adjective.

Figure 4 shows preferences across question types. Highlight Many was preferred in both of the superlative cases, and even more frequently in the graded adjective case. Contrary to our hypothesis, the two superlative cases behave more similarly to one another overall, suggesting perhaps that some crowd workers are interpreting the plural form of the question as asking for the single highest expense.

Highlight One is the second most common choice, occurring slightly more often in the singular superlative question than in the plural superlative. This is the direction that we hypothesized, but the difference between the two is not as large as we anticipated. Filtering to just one item was rarely (2%) chosen for the graded adjective case, but was chosen about 13% and 17% of the time for the superlative cases. Filter Many was chosen nearly equally often (between 7 and 11%) across the three question types.

Figure 2 shows the results by shape, question type, and view type. As noted above, overall we see strong preferences for Highlight Many, with a few exceptions. As hypothesized, Shape 2 was especially problematic in terms of meeting more than 50% agreement, and Shape 3 had high agreement overall.

In more detail, for Shape 1, we hypothesized a preference for Highlight One for singular superlative, versus Highlight Many for plural superlative and graded adjective. This was partially born out. Participants did not choose Highlight One as hypothesized; instead, more context is desired. We hypothesized little agreement for Shape 2, and a majority was not reached for single superlative nor plural superlative. Interestingly, there was high agreement for graded adjective, showing the complex interactions that occur between the distributions of the data and the nuances of language.

As hypothesized, there is more than 60% agreement on the response for Shape 3 for plural superlative and for graded adjective. Interestingly, this is also the option chosen by a strong majority for the singular superlative, which we did not anticipate. However, this is the best answer if context is wanted. As one participant wrote: “Gives more information and its a five-way tie.”

5 Discussion and Future Work

These results show a strong signal in favor of highlighting the distribution with the k top values, where k is a function of the distribution of the data. We have found evidence that for distributions with disagreement as to which values apply to the modifier, the decision about the best view to show becomes muddled, which we predicted from results from the cognitive linguistics literature.

Further research must be done to determine how robust this result is across distributions, and if it applies to very long distributions that require scrolling. Nonetheless, there was quite a strong signal across different variations of expression of vague modifiers and three quite different distribution types, which suggest that the standard way of responding to these questions (showing only the top item for a superlative question, or top few items, for a vague modifier) is not the best approach no matter what the distribution is.

This study has limitations; it only looked at presentation of static bar charts. Still to be investigated are interactive visualizations and other forms of presentation, such as maps for adjectives like “near” and line charts for temporal information such as “recent.” The study only examined one adjective, and although research in cognitive linguistics suggest that these results transfer across words, future work should test these ideas against other words. Future work should also include investigations into more complex sentence structure and assessing the effects of modifiers with complex components, such as “best.” Finally, future work should test these ideas in a live system rather than in the artificial setting of a study.

6 Conclusions

This work has described issues surrounding and occurrences of vague modifiers in natural language interfaces to visual analytics systems. It has presented empirical results based on a crowdsourcing study with 274 participants that compared different ways to respond to a natural language question containing a vague modifier such as “high” or a superlative such as “highest.” The findings are that highlighting the items that correspond to the meaning of the modifier is generally preferred over filtering the results to the top item or top few items, and should serve as a good default interface guideline.

REFERENCES

[1] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Natural language interfaces to databases–an introduction. *Natural language engineering*, 1(1):29–81, 1995.

[2] G. De Melo and M. Bansal. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290, 2013.

[3] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 489–500. ACM, 2015.

[4] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212. ACM, 2010.

[5] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):309–318, 2018.

[6] J. Hullman, E. Adar, and P. Shah. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470. ACM, 2011.

[7] C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45, 2007.

[8] C. Kennedy and L. McNally. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381, 2005.

[9] W. Kessler and J. Kuhn. A corpus of comparisons in product reviews. In *LREC*, pp. 2242–2248. Citeseer, 2014.

[10] M. Lamm, A. T. Chaganty, C. D. Manning, D. Jurafsky, and P. Liang. Textual analogy parsing: What's shared and what's compared among analogous facts. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2018.

[11] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[12] M. Morzycki. *Modification*. Cambridge University Press, 2015.

[13] D. H. Park and C. Blake. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the workshop on detecting structure in scholarly discourse*, pp. 1–9. Association for Computational Linguistics, 2012.

[14] R. A. Pazos R, J. J. González B, M. A. Aguirre L, J. A. Martínez F, and H. J. Fraire H. Natural language interfaces to databases: an analysis of the state of the art. *Recent Advances on Hybrid Intelligent Systems*, pp. 463–480, 2013.

[15] A.-M. Popescu, A. Armanasu, O. Etzioni, D. Ko, and A. Yates. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 141. Association for Computational Linguistics, 2004.

[16] C. Qing and M. Franke. Meaning and use of gradable adjectives: Formal modeling meets empirical data. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, 2014.

[17] L. A. Schmidt, N. D. Goodman, D. Barner, and J. B. Tenenbaum. How tall is tall? compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual conference of the cognitive science society*, pp. 2759–2764. Citeseer, 2009.

[18] G. Scontras, J. Degen, and N. D. Goodman. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66, 2017.

[19] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 365–377. ACM, 2016.

[20] V. Setlur, M. Tory, and A. Djalali. Inferencing underspecified natural language utterances in visual analysis. In *IUI*, 2019.

[21] C. Shivade, M.-C. de Marneffe, E. Fosler-Lussier, and A. M. Lai. Identification, characterization, and grounding of gradable terms in clinical text. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 17–26, 2016.

[22] S. Solt and N. Gotzner. Expensive, not expensive or cheap. In *An experimental investigation of vague predicates, Slides presented at the 11th Szklarska Poreba Workshop, Poland*, 2010.

[23] S. Solt and N. Gotzner. Experimenting with degree. In *Proceedings of SALT*, vol. 22, p. 353364, 2012.

[24] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics*, 24(1):511–521, 2018.

[25] K. van Deemter. *Not exactly: In praise of vagueness.* Oxford University Press, 2010.

[26] S. Vegnaduzzo. Acquisition of subjective adjectives with limited resources. In *Proceedings of the AAAI Spring Symposium on exploring attitude and affect in text: Theories and applications*, 2004.

[27] B. Weijters, E. Cabooter, and H. Baumgartner. When cheap isn't the same as not expensive: Generic price terms and their negations. *Journal of Consumer Psychology*, 28(4):543–559, 2018.