Quantity Estimation in Visualizations of Tagged Text

Michael A. Correll
University of Wisconsin-Madison
mcorrell@cs.wisc.edu

Eric C. Alexander
University of Wisconsin-Madison
ealexand@cs.wisc.edu

Michael Gleicher
University of Wisconsin-Madison
gleicher@cs.wisc.edu

ABSTRACT

A valuable task in text visualization is to have viewers make judgments about text that has been annotated (either by hand or by some algorithm such as text clustering or entity extraction). In this work we look at the ability of viewers to make judgments about the relative quantities of tags in annotated text (specifically text tagged with one of a set of qualitatively distinct colors), and examine design choices that can improve performance at extracting statistical information from these texts. We find that viewers can efficiently and accurately estimate the proportions of tag levels over a range of situations; however accuracy can be improved through color choice and area adjustments.

Author Keywords

Text visualization; text analytics; information visualization; perceptual study;

ACM Classification Keywords

H.5.0. Information Interfaces and Presentation: General

INTRODUCTION

Text analysis determines properties of collections of texts using techniques ranging from statistical processing to manual annotation. Some text visualization tools attempt to convey patterns and trends across the entire (potentially large) corpus. In contrast, *tagged text* can also be used in visualizations: individual words are marked (with e.g. a color, glyph, or other token) to indicate their associated properties. Showing specific words can inform the viewer as to what textual details contribute to the overall pattern and can help them localize patterns in the larger text. However, for such tagged text visualizations to be useful, the viewer must still be able to infer the larger trends from the lower level details in specific words. In this paper, we empirically evaluate this ability of viewers to determine aggregate properties from displays using tagged text, both confirming the viewer's capability to estimate efficiently but also presenting and validating design ideas that address sources of inaccuracy.

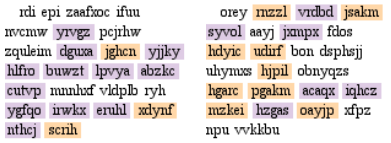
CHI 2013, April 27–May 2, 2013, Paris, France.

The ability to connect between aggregate patterns and specific words is desirable in a number of applications. For example, a sentiment analysis application may analyze a collection of articles to identify positive and negative remarks about a product. While one could display the final aggregate totals of this analysis, linking directly to the original text preserves content, context, allows detection of noise and outliers, and affords analysis at multiple scales. A visualization using tagged text can show what words contribute to the sentiments and allow an analyst to identify where in the texts sentiment inducing remarks occur. However, these details are only useful if an analyst can still determine the aggregate pattern: are there more positive than negative features?

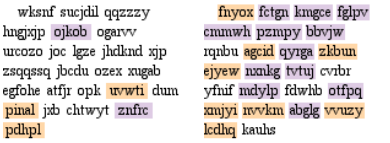
This paper considers the specific task of estimating the proportion of words that are tagged with a particular class of tag. We consider the common case where each word is associated with at most one tag, and there are few enough tag categories that each can be associated with a color. In this case the task would be to estimate the approximate number (the *numerosity*) of a particular color of tag with respect to all the other colors of tag. Our experiments seek to determine whether these numerosity judgments can be done efficiently and accurately.

The ability of the visual system to efficiently estimate aggregate properties has been shown repeatedly by the perception literature. Reports of the ability of people to make approximate judgments of numerosity suggest that tagged text displays may be efficiently and accurately interpreted. However, it is unclear if these results apply to tagged text applications, as the experiments usually consider small scale, artificial stimuli and do not consider other factors that may affect text displays (e.g. biases due to relative area, relative density, reading order, or text spacing). Therefore, we need to understand the performance of viewers at estimating the aggregate proportions of tagged text, and the potential biases that can affect accuracy. Only with such understanding can we design effective tagged text displays.

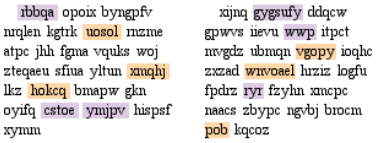
In this paper we provide evidence for the ability of viewers to make accurate judgments about numerosity in tagged text, and offer design choices which further improve this ability. Our work conducted a series of five experiments (summarized in Figure 1) that confirm that viewers can make efficient estimations using displays of tagged text, expose biases in those displays, and validate designs that address those biases. Our experiments show that viewers are able to make numerosity judgments efficiently and accurately for a wide range of stimuli. However, they also indicate there are certain properties of stimuli that can create biases in estimation. In particular color (perceptual illusions brought on by specific color choices) and word length (whether the certain categories of



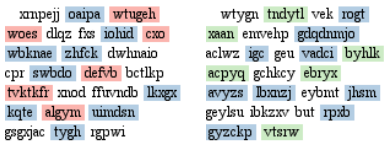
(a) Viewers can accurately estimate the proportion of tags (e.g. whether the text is 20% or 60% orange) to within about 5%



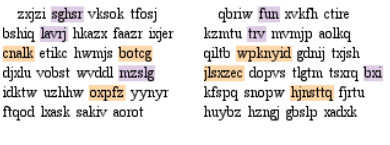
(b) Estimation accuracy is robust to changes in density (e.g. whether 20% or 80% of all words are tagged).



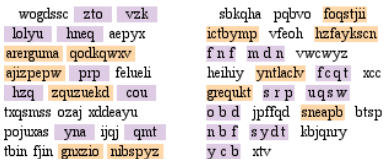
(c) Estimation accuracy is robust to variance in word length (e.g. whether words can be 5, 4-6, or 3-7 letters long).



(d) Estimation of tag number is biased by color. Certain colors may systematically be selected more than others (e.g. commonly selected red versus commonly avoided green).



(e) Estimation of tag number are area biased. Larger areas may seem to have higher counts. (e.g. on the right, where orange words are systematically longer than purple words)



(f) Area bias can be mitigated by adjusting the padding (extra space around words) or tracking (inter-word spacing) of tagged words. Here area is adjusted to match proportion.

Figure 1: A summary of the results of the five experiments reported in this paper. Together, they suggest that tagged text displays can be a useful presentation of data that accurately conveys the overall proportions of tags while allowing the reader to see the individual words, providing some design guidelines are followed.

tags have characteristically longer or shorter words) have strong effects on performance. By avoiding color schemes with known perceptual issues and using small manipulations in inter-character spacing (tracking and padding) we can mitigate these biases and promote good viewer performance even for a wide range of text stimuli.

Summary of Results

In this paper we present the results of a series of experiments using crowd-sourced participants to make judgments about tagged text. The experiments lead to the following conclusions (summarized visually in Figure 1):

- Participants were able to make accurate (to within a few percent) and efficient (faster than counting) judgments about the relative numbers of tagged words in paragraphs of text for a wide variety of stimuli.
- These judgments are robust to tag density and uniform variation in word length (experiments 1 and 2)
- Semantically or perceptually problematic color choices can skew these judgments (experiment 3).
- Area of tags systematically bias estimates of the number of tags: large differences in area can perceptually inflate small differences in number, and vice versa (experiment 4).
- By artificially adjusting comparative area this confound is reduced (experiment 5).

Contributions

In this paper we present results in the relatively unexplored domain of text annotation (specifically, applications where text is tagged with semantically meaningful colors). We present empirical validation of several factors related to performance at numerosity estimation tasks for tagged text, and further validate design choices that improve this ability. For example, we provide a novel design which adjusts text area to facilitate accurate perception. Our findings suggest that tagged text displays can be an effective visualization.

BACKGROUND AND MOTIVATION

Our work is part of a general trend of translating results from perceptual psychology to specific problems in information visualization. Therefore we draw upon prior work in both the specific visualization task we consider (namely the visualization of tagged text) as well as the general perceptual research into perception of numerosity and general aggregation.

Tagged Text Visualization

Large text corpora are often represented abstractly by high-dimensional vectors which are then visualized in network graphs or in lower-dimensional spaces via dimensionality reduction [30]. The dimensions of these vectors are often specific words or tokens in the corpus, creating a one to one link between individual words and high level statistical patterns (such as topic membership, rhetorical difference, or entity possession). Other text visualization work, especially work dealing with streaming text data, has also incorporated annotated raw text or annotated word clouds into their visualizations [7, 17]. Since the analyses behind these visualizations operate at the level of individual tokens, tagging can be used to connect higher level properties to specific passages of text.

Our recent research has highlighted the need for high level abstractions to be connected with the original text [9]. This connection to specific texts becomes even more important when dealing with applications for the humanities, where patterns of word usage and rhetorical style (rather than semantic content) are the subjects of analysis [5, 12]. Aggregate statistics can identify different patterns of word usage , but they cannot help *explain* these differences: this requires close analysis of the texts. Past tools we have created for humanities collaborators rely on tagged text to link high level trends in data

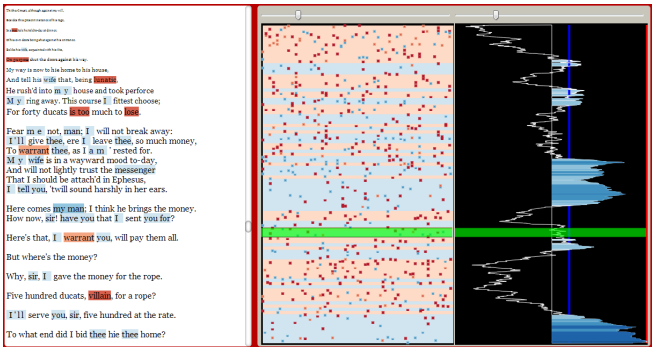


Figure 2: An example of a visualization that combines direct display of specific text tags with other aggregate displays. This visualizations is used to connect patterns of usage to high level concepts. Words are tagged based on rhetorical function and then used to classify different types of texts (here red words are common in Shakespeare's comedies and blue words are common in his tragedies). The right-hand side shows some overall statistical information, but only by considering particular passages of text in context were users able to craft arguments about the meaning of these overall patterns. Users may have questions at potentially arbitrary levels of aggregation (which line has the most of a certain type of word? which paragraph? which quotation?) so we must rely on the ability of users to provide rough comparative estimates of tag counts at many scales. These tasks may not require *exact* calculations, but comparative *estimates* of relative tag counts.

with low level structures in text [10]. Our collaborators have successfully used these tools to identify specific passages that exemplified larger patterns [20]. Figure 2 shows an example of a tagged text visualization currently in use: in all cases the high level patterns are important, but require grounding in low level word usage. Only presenting aggregate statistics in a complementary display is not sufficient to examine the effects of high level patterns at the level of individual text.

The aggregate statistics of word tags can be computed easily and presented visually. For example a wordle could be computed for each paragraph, or a graph could plot occurrence density over the length of the text. While such displays may scale to longer texts, they also do not serve all needs. For example, they do not link back to specific instances of words in the text, require advance knowledge of the desired unit of aggregation (e.g. to look for passages or paragraphs), give little sense of the distribution (e.g. to look for noise and outliers), and provide little way to compare the mixture of tags within a region. In practice, displays of aggregate statistics offer complementary advantages to direct display of tags, and are often used together with them (an example is shown in Figure 2).

Methods of presenting and annotating raw text for visualization applications is understudied, partially because the affordances of annotated text are not fully understood. This work hopes to provide more insight in this area with the idea of motivating further investigation and applications.

Numerosity Estimation and Aggregation

The visual system is capable of performing a number of tasks efficiently. Some of these tasks, such as searching and grouping, are well-known and well-studied by the visualization community. Knowledge of other tasks, including efficient judgements of aggregates, is more recent and is just beginning to influence visualization design.

Psychophysical experiments have confirmed that people are capable of efficiently perceiving the counts of small numbers of objects [22]. Even when the number of items is large there is still evidence for an approximate number system capable of estimating a range of numbers [3, 25] which is distinct from verbal numerical associations [26]. The accuracy, precision, and extent of this system is still a matter of conjecture, but it is known that there are several confounding variables that can hurt the performance of people asked to estimate numerosity. In particular, artificial "calibrations" (e.g. stating that "this stimulus has 30 dots," whether true or not) can systematically bias estimates [25], as can the relative areas of the stimuli [21], the density of the samples in the stimulus, the convex hull of the stimulus [15], and the gestalt segmentation of objects within the sample [13]. Multiple (occasionally conflicting) models of the effects of these other factors on the approximate number system have been proposed [11, 1, 33], but there is still some ambiguity as to the relative importance of different visual information on resulting comparative or estimation numerosity tasks.

Stimuli for these studies of numerosity estimation are intentionally simplistic or artificial, to better control for certain channels of visual information [14]. Recent work extends these results to more complex stimuli for applications in information visualization [8]. This paper further extends this line of work to the more natural case of text visualization while at the same time identifying factors that influence task performance to inform the design of visual displays.

GENERAL EXPERIMENT DESIGN

We conducted a series of five experiments in which participants were exposed to paragraphs of text where certain words had been given a colored background (tagged). We asked the participants to make judgments about the relative counts of tagged words.

The relative proportion of tagged words of a certain color as a percentage of total tagged words (hereafter the "mixture") was used as a proxy for a general class of aggregate judgments. Using this metric afforded two classes of experimental task: either a forced binary choice task ("which color tag is the most common tag") or an estimation task ("what percentage of the tagged text is a particular color"). This task is a common one in current tagged text applications, and is also a primitive task in a larger set of quantitative and statistical tasks (such as per text comparison or corpus distribution estimation). In practice, precise quantification is not always necessary, but sufficiently good estimation is required to make comparative judgments.

The basic experimental design was similar across our five experiments. After giving consent, participants were shown a brief tutorial explaining the experimental task and emphasizing that the judgments would be about the count of words rather than the length of words. They were then presented with a set of stimuli in random order. In order to discourage subjects from explicitly counting the tags, each stimulus

vanished after 20 seconds, though the subject could still take longer to give their answer.

Stimulus Design

All the words in the stimuli were made up of random, lowercase letters in 12-point Times New Roman font. We chose to use random text to both remove possible confounds based on the semantic meanings of words or text, and also to discourage participants from spending too long reading through each paragraph. This choice of random text also afforded us more control over the exact visual and textual properties of the stimuli, at the expense of task realism. Our use of random text reserves study of confounds with comprehension for future work. However, we feel that this tradeoff is important in these initial experiments: the control allows us to explore a larger space to understand a variety of conditions, including distributions both simpler and more extreme than natural (although the notion of "natural" is quite varied and application dependent). Tags were represented using filled-color boxes surrounding the words. Stimuli were 400 pixels wide, left-justified, and consisted of between 200-300 words depending on the experiment.

We considered two orthogonal dimensions of tagged text: the ratio of tagged words (of any color) to the number of words in the paragraph, which we call "tag density", and the ratio of tags of a particular color to the total number of tagged words, or the "mixture level" as discussed above. For the orange-or-purple color schemes, mixture level is defined as the ratio of specifically orange tags to the number of tagged words in the paragraph. Figure 3 shows four points in this parameter space. An important feature of both of these two factors is that they are unaffected by word length, as they deal solely with the numerosity of the tagged words rather than their lengths. This is not true of a third factor we considered, which we will call "tag area," which we define as the percentage of physical area (at the pixel level) encompassed by the tag boxes that corresponds to a particular color of tag. If words that are systematically longer than the average are grouped in a class their perceived area will be larger, while classes with shorter words will take up comparatively less space.

Participants

All 210 participants for the five experiments in this study were recruited using Amazon's Mechanical Turk infrastructure, specifically those in North America with at least a 95% approval rating. Ishihara plates were used as a pretest to exclude potential participants with Color Vision Deficiency (e.g. color blindness) [18]. The demographics data roughly conforms to the general distribution of North American Turk users [29]: age of participants ranged from 18-65 ($\mu = 34.7$, $\sigma = 11.1$), with 101 male and 109 female participants. We followed acknowledged best practices to improve the reliability of our experimental data, including validation questions, randomizing questions, requiring mandatory questions of different input types, and checking for "click-through" behavior [23, 27]. Despite these measures, we expect responses from crowd-sourced participants to have higher variance than in-person results. We take positive results from crowd-sourcing as indication that the relevant effects are robust.

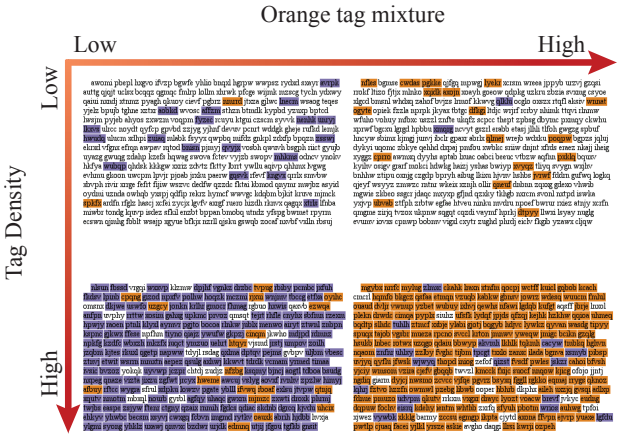


Figure 3: Example stimuli from our parameter space, varying tag density (how many words are tagged with *any* color) and tag mixture (how many of the tagged words have a specific color, in this case orange rather than purple words). In real world examples both of these variables are presumably independent of the display choice but an underlying property of the data.

Note

We present an overview of results in the following section. Further details, including ANOVA tables, sample stimuli, and additional charts for all experiments are available in the supplemental appendix.

SPECIFIC DESIGNS AND RESULTS

Experiment 1

Our first question was to measure the general task performance and to confirm that this holds over a range of situations. Our hypothesis, based on previous experiments with the perception of aggregate statistics, is that people can accurately estimate the relative numerosity of tagged text. For our initial experiments with robustness, we chose mixture level and tag density. We had no hypotheses about the effect of mixture or tag density on performance, but we still wished to confirm the null hypothesis in the general case to afford grounded manipulations of text stimuli in later experiments. To test these three hypotheses, we performed an experiment that measured performance over a range of mixture levels and tag densities.

Our stimuli were 200 random words consisting of five random characters each. 20 participants were each presented with 36 paragraphs of text at different levels of tag density (20%-100% dense; we excluded 10% as this allowed participants to simply count the number of tags and make precise rather than estimated judgments). For each level of density participants were presented with paragraphs with two colors of tags (a purple and orange drawn from the Colorbrewer scale of qualitative colors [19]), and asked to estimate the relative counts of each color with a slider (e.g. a paragraph could be "20% purple, 80% orange"). Performance was measured by the absolute difference of the participant's guess from the true distribution. We also included four "validation stimuli" with mixtures of either 100% orange or 100% purple for a

total of 40 stimuli presented in random order. These validation stimuli were used as engagement checks and to verify that participants had properly understood the instructions, and were not considered in further analysis. No participants needed to be excluded based on the validation stimuli.

To assess our first hypothesis that people can accurately estimate the relative numerosity of tagged text, we consider the error level across all conditions. Here, we take the absolute difference between the presented mixture level and the participant reported level. We find that absolute error was low across all conditions ($M = 0.063, SD = 0.088$, lower than the fidelity of the slider.

To test our hypotheses about robustness across conditions, we performed a two-way analysis of variance (ANOVA) to test whether tag density and mixture level affected absolute error. Our results found no significant effect of tag density on performance (F(8,8) = 1.40, $p = 0.20$) but did find an effect of mixture on performance (F(8,8) = 5.66, $p < .0001$): in particular a post hoc comparison of Tukey's Honest Significant Difference (HSD) found that the absolute error was significantly lower for mixtures of where one color was only 10% of the total number of tags, and where there were equal counts of both colors of tag. For example, at 10% mixture, the average error was approximately 1%, whereas the error across all conditions was closer to 6%. Some of this difference may be due to the natural bias that arise when scales with midpoints are used in human subjects research [28]. A closer analysis of the patterns of response using confidence intervals at the $\alpha = 0.05$ level of significance shows that, when given a slider input to choose the appropriate mixture level, participants were statistically more likely to choose the extremes and the midpoints of the scale than other responses (thus creating a lower average error than when the mixture actually *was* on of these values). This effect of mixture was consistent across all experiments with slider inputs (specifically experiments 1, 2, and 5).

Experiment 2

Having seen evidence of robustness for two potential factors, our next experiment explored another potentially relevant factor for performance, namely variance in word length. Our initial hypothesis was that in the case where there is a large variance in word length, the noisier patterns in stimuli might make judgments difficult. In real corpora, the lengths of tagged words can vary wildly (for instance, the distribution of word lengths of code tagged with syntax highlighting would be very different from prose). It was infeasible to test every possible distribution of word lengths in texts. For this experiment we wanted simply to determine if variation per se in word length would have an effect on performance: in the case of a positive result follow-on experiments would investigate the situations where degradation would occur.

The stimuli for this experiment were 300-word paragraphs with three different conditions for word length. In the first condition, all words were five letters long, as in experiment 1. In the second condition, word lengths were uniformly distributed across lengths of four, five, and six letters. In the third condition, word lengths were uniformly distributed across each length from three to seven letters. We generated stimuli at two density levels, 30% and 70% dense, as a check for possible interaction effects. We again recruited 20 subjects and asked them to use a slider to estimate the mixture levels of 36 stimuli (6 examples of each word-length condition cross tag density). We used similar validation stimuli to those in the first experiment (100% orange and 100% purple) and again did not need to exclude any participants.

Participants again had low error across conditions ($M = 0.051, SD = 0.081$). We performed a two-way ANOVA to test whether word length condition and mixture level affected absolute error. Our results found no significant effect of variance in word level on performance (F(2,2) = 1.97, $p = 0.14$), though we again found a significant main effect for mixture level (F(8,8) = 4.42, $p < .0001$). As with the previous experiment, post-hoc analysis with Tukey's HSD test shows that the significant outliers were the mixtures at the edge cases of 0.1 and 0.9. This disconfirms our initial hypothesis, but provides evidence of the generalizability of the good performance seen in these two experiments when applied to real text and real world applications, where one would expect variable word lengths but not necessarily systematic per-tag class skews in word length (experiments 4 and 5 test performance when this assumption of no per-tag class bias in word length is broken).

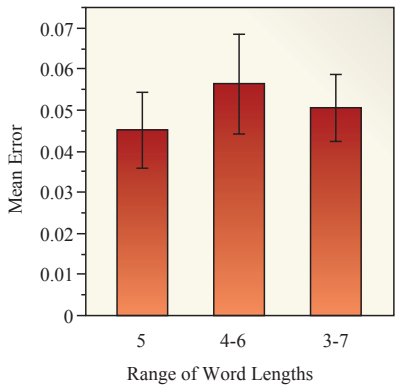


Figure 4: In Experiment 2, there was no significant effect of variance in word length on performance, and overall errors were small. Participants were only able to answer in increments of 10%, and yet average error was significantly less than 10% across all conditions.

Experiment 3

We wanted to extend our results to multi-category situations. We hypothesized that performance would be robust to small numbers of categories. However, indicating multiple categories requires choosing sets of colors to indicate the various tag classes. We hypothesized that if we followed best practices in choosing sets of colors, the specific colors used would not affect performance. The need to examine multi-way comparisons lead us to a different experimental design. We used a forced choice design where the participant was asked to choose the most commonly occurring color from five choices.

This design enabled a different measure of performance. The difference between the value of the "winning" class and the next highest class provides a measure of task hardness. As

this difference decreases, the ambiguity between winner and runner-up increases.

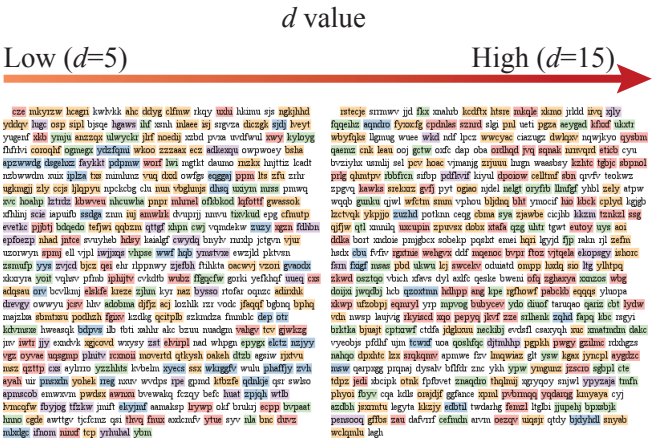


Figure 5: The d parameter for the multicolor experiments (left $d = 5$, right $d = 15$). In both cases orange tags are 35% of the text, but on the left red accounts for 30% of the area, versus 20% on the right. The larger the value of d (and so the bigger the gap between the winner and the second most common color), the better the performance.

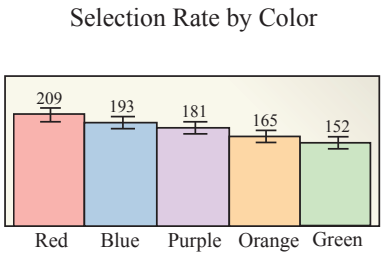


Figure 6: Counts of color choices for the multicolor experiments. Even though participants saw equal numbers of all stimuli where each color was the correct answer, participants picked red significantly more often than green. Expected number of guesses per color is 180.

For this experiment stimuli were tagged with five different colors again drawn from the Colorbrewer qualitative color sets. Each stimulus had a "winner" color with the highest proportion of tags and a clear "runner-up" with the second highest proportion. While the winner always accounted for 35% of the tags in a given stimuli, we varied the difference between the winner and the runner-up (hereafter the parameter d) across three different conditions: 5%, 10%, and 15%. (Thus for the 5% case the winner was 35% of the total count of tags and the runner up was 30% of the total count of tags). The remaining three colors were given proportions at least five percent less common than the runner-up. Figure 5 shows the effect of d on a sample stimulus. Word lengths were evenly distributed from three to seven letters across all colors. Each of our 20 subjects was shown three stimuli for each combination of d and winning color for a total of 45 stimuli. We again included validation stimuli in which a single color had 100% mixture level, and again did not need to exclude any participants.

A two-way ANOVA tested the effects of d and the winning color on accuracy (the likelihood of selecting the correct winning color): our results show that the parameter d was an effective proxy for task difficulty ($F(2,2) = 30.97$, $p < .0001$). A post hoc Tukey HSD test confirmed that task performance rose monotonically with d from 71% accuracy when d=5% to 95% when d=15%. The results show that viewers can make judgements across multiple classes, but the performance degrades as the stimuli become more ambiguous.

A main effect was present for color ($F(4,4) = 6.29$, $p < .0001$). A post hoc analysis of color choice using confidence intervals at the $\alpha = 0.05$ level of significance shows that, while we would expect participants to evenly guess colors (and so the confidence intervals for probability of guessing a particular color should include 20%), participants guess red (the modal guessed color) significantly more than green (the least guessed color). Figure 6 shows this effect with respect to the number of times each color was guessed. Since we had adjusted the colors to be isoluminate, we attribute this effect to known perceptual artifacts present for particular choice of colors. Color biases of these sort (and red-green biases specifically) have occurred in other information visualization settings where area is perceived as larger when certain choices of color encoding are made [6, 32]. Since the effect of these illusions is usually to artificially inflate the perceived area of regions, this supports the hypothesis that area calculations play a role in numerosity judgments.

Experiment 4

Figure 7: An example of a stimulus with an extreme mismatch between area and mixture: although there are more purple words than orange words in the paragraph, since the orange words are longer they both account for a larger amount of visual area on screen. Even in this extreme case (where 40% of the words are orange but 60% of the area is orange), participants were still fooled (they correctly chose the dominant color only 37.5% of the time, versus an accuracy of 86.25% when there was no mismatch).

The color bias exposed in Experiment 3 suggested that we should investigate other potential sources of bias. Previous research into numerosity estimation indicates that area, density, and other gestalt groupings can bias or otherwise confound the approximate number sense. Area is particularly relevant in text, as different tag classes and texts contain different words that may have different skews in their word length distributions. To be effective, tagged text displays must be robust to skews in word length (and therefore the visual area of tags) relative to numerosity. Long words should not count for more than short words. To that end we investigated cases

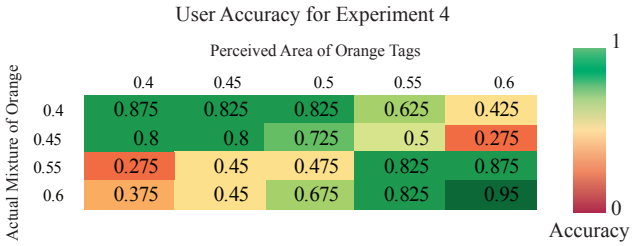


Figure 8: Participant accuracy at the forced choice estimation task ("which class of tag is more common?") was significantly higher for stimuli in which the mixture levels and tag areas corresponded than in cases where they conflicted. Performance is best in the upper left and lower right corners, where the mismatch between area and number is away from the decision boundary in the correct direction.

where there are mismatches between relative area and relative numerosity in text. It was our hypothesis that large mismatches would systematically bias results in the direction of the area. However, some perception results suggest that number could dominate area for some types of stimuli.

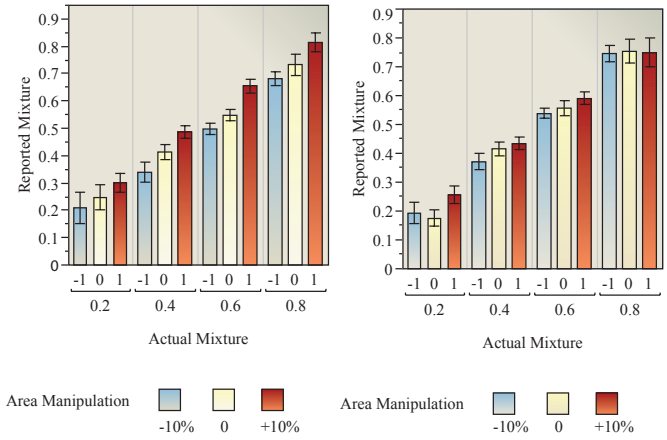
In this experiment we tested our subjects' ability to differentiate between the numerical proportions of tags (mixture level) and the proportions of the physical space they take up (tag area). The task was a forced choice decision between two tag classes (orange and purple) to determine which tag was more common in a paragraph of text. The winning color accounted for 55% or 60% of the tags in a given paragraph, but accounted for 40%, 45%, 50%, 55%, or 60% of the *area* of the tagged words. Letters were randomly added to words to form words of 3-10 characters in order to generate these area discrepancies (e.g. when the winner was 60% of the *count* of tags but only 40% of the *area* then it would mostly be made up of very short words; in the opposite case it would be mostly long words). Each of our 20 subjects was shown two stimuli from every condition of mixture level cross tag area, for a total of 40 stimuli. Figure 7 shows an extreme example stimulus with a large area/count mismatch.

Since the distinction between tag area and mixture level was so crucial to this experiment, we included extra stimuli (not included in the analysis of our results) to make sure participants understood these definitions. We presented each participant with four validation stimuli with area/mixture mismatches in which there were few enough tagged words that it was easy to simply count tags (e.g. eight purple words of three letters each and four orange words of twelve letters each). We were not forced to exclude any participants based on performance on these validation stimuli.

We conducted a two-way ANOVA to determine the effects of area and mixture on accuracy. We found no significant effect of area by itself ($F(4,4) = 1.63$, $p = 0.17$). We would expect that in some cases area mismatches would be beneficial (where the skew is away from the decision boundary), but harmful in others (where the skew is in the opposite direction). There was a significant main effect of mixture ($F(3,3) = 3.46$, $p = 0.016$). This effect is similar to the mixture effect previously seen, except that with the simplified task of a binary choice, participants do predictably better the more mixture levels are skewed towards one color or the other.

There was a significant interaction effect between area and mixture ($F(12,12) = 14.5$, $p < 0.0001$). Post-hoc analysis using Tukey's HSD test shows that while accuracy is generally high when area and mixture level are aligned, there is a significant drop in performance when they disagree (see figure 8). For example, accuracy may drop to as low as 37.5% in an extreme (and admittedly contrived) situation (Figure 7). While performance in realistic settings is unlikely to suffer to this extreme degree, biases in less extreme cases will arise in practice. Therefore, it is clear that this problem must be addressed in the design of tagged text displays.

Experiment 5



(a) No Correction **(b)** Tracking Correction

Figure 10: A per condition breakdown of the effect of area manipulations on response. For each level of mixture (20,40,60, or 80% of the tagged words were orange) we manipulated the area either by making the orange words significantly shorter than the other tags (thus under-representing orange in terms of area) or significantly longer than the other tags (thus over-representing orange). Without the use of any measures to correct for this bias, participants would conflate the area manipulation with the tag mixture, allowing confusion between different levels of stimuli. When extra area is added to words and the inter-word tracking is adjusted, these confusions are reduced.

Experiment 4 revealed that area was a significant confound for numerosity. In order to translate this result into design guidelines, we decided to investigate this effect at a finer level of granularity, as well as analyze the design space for overcoming the bias it introduces. One potential strategy is to artificially correct the tag areas to match mixture levels. In such a strategy, whenever a given color's area was lower than its mixture level, we could add extra colored pixels (padding) to the beginnings and endings of underrepresented words until the area and mixture were equal. Pilot tests, documented in the supplemental appendix, suggested the hypothesis that these corrections would partially (but not entirely) mitigate the effects of the area bias. We attempted to evaluate these corrections more thoroughly in our final experiment.

We conducted a between-subjects experiment with three conditions: no area correction, area correction by padding

(a) No Correction **(b)** Padding Correction **(c)** Tracking Correction

Figure 9: Three levels of our area manipulation factor. 60% of the words in the paragraph are orange, but systematic biases in word length have made 70% of the tagged area orange. On the left the mismatch between area and word count is unaltered. In the middle case extra padding is added to the purple words to compensate for the bias. In the last case inter-word tracking is adjusted to fill the extra buffer space while still maintaining legibility.

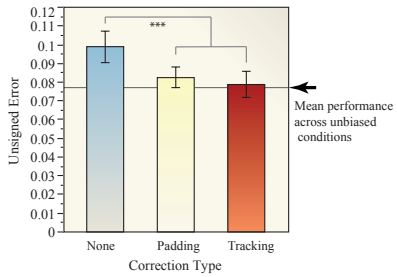


Figure 11: The effect of our different area manipulations on accuracy at determining the mixture (in terms of count) in a paragraph of tagged text. Significances are at the $\alpha = 0.05$ level. The gray line represents participant's average performance when there is no mismatch between numerosity and area. When systematic biases to area are introduced this accuracy suffers. By adding extra space to under-represented tags this error is reduced. By altering the inter-word spacing (tracking) of under-represented tags the error is reduced, but not significantly more so than in the previous case.

the colored area of underrepresented words, and "tracking-adjusted" area correction. By tracking-adjusted, we mean that as opposed to effectively resizing underrepresented tags and centering words inside them, we instead adjusted the tracking (space between characters within a word) so that each word fully filled (or as close as was possible) the horizontal width of its tag. While both of these techniques corrected the area/mixture discrepancy, it was our belief that the tracking-adjusted method represented a more natural solution to the area problem that would preserve legibility. Figure 9 shows example stimuli for each of these conditions.

We presented participants with an estimation task in which they were presented with a paragraph of text with two tag classes (orange and purple) and asked to estimate, with intervals of 5%, the percentage of tags that were orange rather than purple. We generated mixtures of 20%, 40%, 60%, and 80% orange. For each mixture, we generated cases where the area matched the mixture, the area was 10% greater than the mixture, or the area was 10% less than the mixture. We recruited 60 total subjects (20 for each area correction condition), each of whom saw three stimuli for each mixture level cross tag area difference for a total of 36 stimuli. We also included validation stimuli in which area and mixture level did not match, but there were few enough tags that it was easy to count (e.g. three orange tags of three letter each, one purple tag of twelve letter). The validation stimuli were not area corrected. We did not need to exclude any participants based on validation performance.

We performed a one-way ANOVA to test whether area correction method had an effect on subject accuracy. We found that there was a significant main effect of area correction method ($F(2,2) = 8.96$, $p < .0001$). While a post hoc Tukey HSD test confirmed that our two area correction method were not significantly different from each other, both significantly increased overall subject performance over the non-corrected stimuli (see Figure 11). A Student's t-test shows no statistically significant difference between corrected stimuli and stimuli where no area bias was present ($p > 0.71$). This confirms that our manipulations were able to mitigate area/mixture mismatches.

DISCUSSION

Our experiments have shown that viewers can make accurate estimations of numerosity in tagged text for a wide range of stimuli. While there are some factors that introduce bias, these can be mitigated through design. Figure 12 shows one such design: a presentation that accounts for specific biases in human perception of numerosity but takes into account concerns of legibility and known aesthetic principles for text display drawn from the HCI literature.

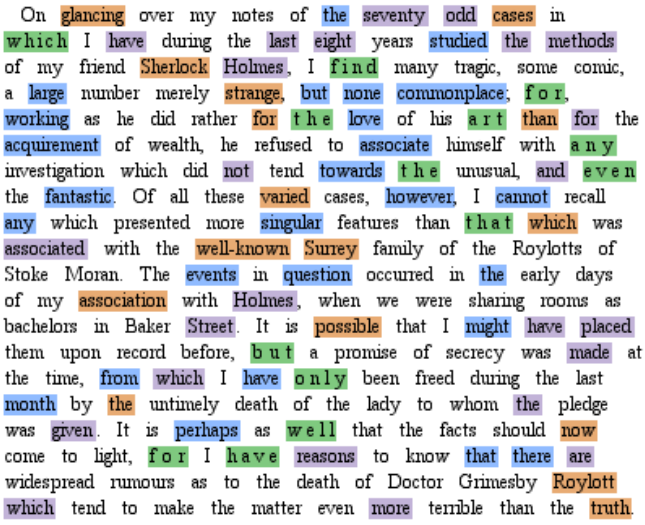


Figure 12: An example of inter-word tracking changes on real text. Since the green tagged words are on average shorter than the other colors, there is a mismatch between perceived area and perceived numerosity. Modifying the inter-word tracking space attenuates this mismatch and produces better accuracy. The example is an extreme case of length mismatch: in practice, the required spacing changes are more subtle.

Experiment five showed two different designs for addressing the area bias problem. In terms of measured performance the two had no significant difference. However, they may have different impacts on aesthetics and legibility. Adding space, either between words or letters, does impact text appearance. This may create possible concerns over aesthetics and legibility. However, the literature suggests that increased tracking for individual words may actually improve legibility [2, 4]. Therefore, we feel that adjusting area by tracking provides a plausible mechanism for countering area bias, but should be more extensively tested in real-world applications where legibility and aesthetics must be considered.

Similarly, the choice of color sets has impacts on estimation performance, aesthetics, and legibility. Guidelines for text backgrounds in the literature are mainly concerned with contrast, as this is a key element in legibility [16]. However, other effects suggest that certain colors be avoided. In Experiment 3, we observed that red and green may be problematic, as shown in prior studies. Without a better understanding of the underlying mechanism, it is hard to make stronger design suggestions. However, we believe that standard practices in visualization for choosing distinguishable colors should be applied (e.g. using known Colorbrewer colorsets [19]). Also, in real world applications, care must also be taken to avoid the Stroop effect [31] (best known by examples where a word denotes a particular color but is colored a separate color, i.e. the word "red" colored green): there should not be a confusion between the appearance of the word and the semantic content of the word. The more conflict between visual appearance and semantics, the more difficult the associated text-related or color-related tasks [24].

Limitations

Our experimental models were focused more on the lower level psychophysical features of the task. As such we did not present stimuli using real text (which would require reflection, reading time, and checks for comprehension) which would make it difficult to limit exposure time in such a way as to prevent participants from explicitly counting the numbers of tagged words. We think this choice improves the generalizability of the results at the expense of artificiality of the task, although we feel our stimuli are closer to what might be seen in an information visualization than previous lower-level results in this area.

Our work examines only short (single paragraph) texts. As the scale of the task increases (both in terms of number of paragraphs, length of paragraphs, number of tag classes, and different possible values), performance may degrade. Our future work will examine larger scales that require aggregate or statistical judgments in order to analyze the impact of summarization tools and techniques for quickly juxtaposing multiple short sections of text (such as focus+context displays, or multi-window views). We also plan to investigate more sophisticated queries (such as different aggregate statistics e.g. skew, kurtosis, and variance).

CONCLUSION

In this paper we have examined the ability of viewers to make judgments about estimated values in tagged text. We have shown viewers can accurately and efficiently make these judgements across a large set of stimuli. However, we have shown that certain factors such as relative area or choice of color can degrade performance. We have proposed and empirically validated a design which accounts for these factors.

Our work has implications for the design of visualizations. First, it shows that designers can use tagged text displays with some confidence that the aggregate statistics will be conveyed accurately. Second, it shows that these designs are robust across a number of factors. Third, it shows that while there are some potentially problematic biases, these may be mitigated by considering them in the display design.

ACKNOWLEDGMENTS

This work was supported by NSF awards CMMI-0941013 and IIS-1162037. Work in specific domains are supported by NIH award R01 AU974787 and a Mellon Foundation Grant.

REFERENCES

1. Allik, J., and Tuulmets, T. Occupancy model of perceived numerosity. *Perception & Psychophysics 49*, 4 (Apr. 1991), 303–14.
2. Arditi, A., Knoblauch, K., and Grunwald, I. Reading with fixed and variable character pitch. *JOSA A 7*, 10 (1990), 2011–2015.
3. Burr, D., and Ross, J. A visual sense of number. *Current Biology : CB 18*, 6 (Mar. 2008), 425–8.
4. Chaparro, B., Baker, J., Shaikh, A., Hull, S., and Brady, L. Reading online text: A comparison of four whitespace layouts. *Usability News 6*, 2 (2004), 1–7.

5. Clement, T., Plaisant, C., and Vuillemot, R. The Story of One: Humanity scholarship with visualization and text analysis. *Relation 10*, 1.43 (2009), 8485.
6. Cleveland, W., and McGill, R. A color-caused optical illusion on a statistical graph. *The American Statistician 37*, 2 (1983), 101–105.
7. Collins, C., Viegas, F., and Wattenberg, M. Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, IEEE (2009), 91–98.
8. Correll, M., Albers, D., Franconeri, S., and Gleicher, M. Comparing averages in time series data. In *Proceedings of ACM CHI* (2012).
9. Correll, M., and Gleicher, M. What shakespeare taught us about text visualization. In *IEEE Visualization Workshop Proceedings*, The 2nd Workshop on Interactive Visual Text Analytics: Task-Driven Analysis of Social Media Content (oct 2012).
10. Correll, M., Witmore, M., and Gleicher, M. Exploring collections of tagged text for literary scholarship. *Computer Graphics Forum 30*, 3 (Jun. 2011), 731–740.
11. Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., and Morgan, M. J. A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences 108*, 49 (2011), 19552–19557.
12. Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, ACM (2007), 213–222.
13. Franconeri, S. L., Bemis, D. K., and Alvarez, G. a. Number estimation relies on a set of segmented objects. *Cognition 113*, 1 (Oct. 2009), 1–13.
14. Gebuis, T., and Reynvoet, B. Generating nonsymbolic number stimuli. *Behavior research methods 43*, 4 (Dec. 2011), 981–6.
15. Gebuis, T., and Reynvoet, B. The role of visual information in numerosity estimation. *PloS one 7*, 5 (Jan. 2012), e37426.
16. Greco, M., Stucchi, N., Zavagno, D., and Marino, B. On the portability of computer-generated presentations: The effect of text-background color combinations on text legibility. *Human Factors: The Journal of the Human Factors and Ergonomics Society 50*, 5 (2008), 821–833.
17. Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D., Haug, L., and Hsu, M. Visual sentiment analysis on twitter data streams. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, IEEE (2011), 277–278.
18. Hardy, L., Rand, G., and Rittler, M. Tests for the detection and analysis of color-blindness. *JOSA 35*, 4 (1945), 268–271.
19. Harrower, M., and Brewer, C. Colorbrewer. org: an online tool for selecting colour schemes for maps. *Cartographic Journal, The 40*, 1 (2003), 27–37.
20. Hope, J., and Witmore, M. The hundredth psalm to the tune of "green sleeves": Digital approaches to the language of genre. *Shakespeare Quarterly 61*, 3 (2010), 357–390.
21. Hurewitz, F., Gelman, R., and Schnitzer, B. Sometimes area counts more than number. *Proceedings of the National Academy of Sciences of the United States of America 103*, 51 (Dec. 2006), 19599–604.
22. Kaufman, E. L., Lord, M. W., Reese, T. W., and Volkmann, J. The discrimination of visual number. *The American Journal of Psychology 62*, 4 (1949), pp. 498–525.
23. Kittur, A., Chi, E., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceedings of ACM CHI*, ACM (2008), 453–456.
24. Klein, G. S. Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology 77*, 4 (1964), pp. 576–588.
25. Lemer, C., Dehaene, S., Spelke, E., and Cohen, L. Approximate quantities and exact number words: dissociable systems. *Neuropsychologia 41*, 14 (2003), 1942 – 1958.
26. Lyons, I. M., and Beilock, S. Symbolic Estrangement: Evidence against a strong association between number sense and numerical symbols. In *Annual Meeting of the Cognitive Science Society* (2011), 1515–1520.
27. Mason, W., and Suri, S. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* (2011), 1–23.
28. Matell, M., and Jacoby, J. Is there an optimal number of alternatives for likert scale items? study i: Reliability and validity. *Educational and Psychological Measurement* (1971).
29. Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the International Conference on Human Factors in Computing Systems*, ACM (2010), 2863–2872.
30. Šilić, A., and Bašić, B. Visualization of text streams: a survey. *Knowledge-Based and Intelligent Information and Engineering Systems* (2010), 31–43.
31. Stroop, J. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General 121*, 1 (1992), 15.
32. Tedford Jr, W., Bergquist, S., and Flynn, W. The size-color illusion. *The Journal of General Psychology 97*, 1 (1977), 145–149.
33. van Oeffelen, M. P., and Vos, P. G. A probabilistic model for the discrimination of visual number. *Perception & Psychophysics 32*, 2 (Aug. 1982), 163–70.