The Biasing Effect of Word Length in Font Size Encodings

Eric Alexander*
University of Wisconsin-Madison

Chih-Ching Chang†
University of Wisconsin-Madison

Mariana Shimabukuro‡
University of Ontario Institute of Technology

Steven Franconeri§
Northwestern University

Christopher Collins¶
University of Ontario Institute of Technology

Michael Gleicher‖
University of Wisconsin-Madison

Figure 1: To test whether word attributes that should be irrelevant might still affect the perception of their font size, we highlighted words within word cloud visualizations and asked participants to choose the larger font. On the left, "zoo" has the larger font, but the length of "moreover" can bias participants toward choosing it as larger. On the right, "source" has the larger font, but the taller ascending and descending parts of "begged" can bias participants toward choosing it as larger.

Abstract

From word clouds to cartographic labels to word trees, many visualizations encode data within the sizes of fonts. While font size can be an intuitive dimension for the viewer, it may also bias the perception of the underlying values. Viewers might conflate the size of a word's font with a word's width, with the number of letters it contains, or with the larger or smaller heights of particular characters ('o' vs. 'p' vs. 'b'). In an ongoing set of experiments, we have found that such factors—which are irrelevant to the encoded values—can indeed influence comparative judgements of font size. For this poster, we present one such experiment showing the biasing effect of word length.

Keywords: Text and document data, cognitive and perceptual skill, quantitative evaluation.

1 Introduction

As the amount of textual data available continues to grow, new methods for analyzing these data are of increasing importance. Text visualizations support analysts in many tasks, including forming a gist of a collection, seeing temporal trends, and finding important documents to read in detail. One common method for encoding data using text rendering is to vary the font size. The importance and impact of font size as an encoding go beyond visualizations of text collections, as words of varying sizes are embedded as labels in many other types of visualizations. Font size encodings can be seen in word cloud applications [8], cartographic labeling [7], and hierarchical visualization tools [3].

Despite their ubiquity, however, there is some question of how effective people are at interpreting font size encodings [4]. Concerns about these encodings arise in part because there are many ways in which words vary with one another outside of font size. In particular, a word's *shape* can vary tremendously. Longer words with more letters take up more area on the screen. The glyphs for some letters are inherently taller or wider than others. Kerning and tracking can create diverse spacing between characters. Differences in font would exacerbate these problems, but even the same font is often rendered differently depending on the platform. Other potential factors that could skew perception include color, font weight, and of course a word's semantic meaning [1, 5, 6].

In an ongoing set of experiments, we are evaluating the degree to which a word's shape can affect comparative impressions of its font size. Here, we present results highlighting the biasing effect of word length on perception of font size in word clouds.

2 Task and Factors

For this experiment, we wanted to investigate the effects of **word length**—the number of characters contained within a word—on perception of font size. Longer words take up more space and often have a larger *area* than shorter words of the same or larger font size. We predicted that these differences in area could interfere with the ability to perceptually distinguish words by pure font size alone.

In particular, we looked at *comparative* judgements of size rather than exact ones, as while reading exact values is not the typical use case for font size encodings, a designer should still have confidence that their users' relative impressions of words are grounded in the data. Specifically, we focused on the use of word clouds. Word clouds are both a common medium for font size encodings as well

*e-mail: ealexand@cs.wisc.edu
†e-mail:chih-ching@cs.wisc.edu
‡e-mail:marianaakemi.shimabukuro@uoit.ca
§e-mail:franconeri@northwestern.edu
¶e-mail:christopher.collins@uoit.ca
‖e-mail:gleicher@cs.wisc.edu

as a challenging context for reading values, given the dense proximity of distracting words and lack of alignment between words.

Upon being shown a word cloud in which two words were highlighted using a darker color (see Figure 1), participants were asked to click on the highlighted word that had been given a larger font size. To check for bias, we looked at **word length agreement**: whether or not the difference in word length *reinforces* or *opposes* the difference in font size. For example, if the word within a given pair with the larger font size also contains more letters, we would say that word length **agrees** with font size. However, if the word with the larger font size contains fewer letters, we would say that word length **disagrees** with font size. If both words are the same length, then the agreement factor is **neutral**.

3 Experimental Design and Results

We showed participants word clouds created using the D3 visualization library [2]. For greater control in stimulus generation, we used words of random characters, excluding characters with ascenders or descenders (e.g., "h" or "g") as well as characters of abnormal width (e.g., "w" or "i"). We enforced a minimum distance between the two highlighted words, and ensured that they shared no common horizontal or vertical baselines that would aid in comparison.

We tested two main factors: font size and word length. Both factors were examined using within-subject comparisons. Font size for the first target word was either 20px, 21px, or 22px, while font size for the second word was either 20px or 22px. Length for both target words alternated between 5 characters and 8 characters. The full combination of these factors created 24 conditions, of which 16 had a "correct answer" (i.e., one of the words had a larger font size), and 8 of which did not (i.e., the words were the same font size). This allowed us to observe both instances of factor agreement and disagreement, as well as see which way people leaned at the extreme marginal case where the sizes were, in fact, equal.

We tested 31 participants recruited from Amazon's Mechanical Turk framework—all native English speakers residing in North America with at least a 95% approval rating. Each saw 150 stimuli (6 per each of the 24 conditions described above, as well as 6 engagement tests). We analyzed answers to questions with a correct answer and without a correct answer separately.

For data where there was a correct answer, we calculated the font size difference (1 or 2 px) and word length agreement ("agree," "neutral," or "disagree") for each stimulus. We then ran a two-way analysis of variance (ANOVA) to test for the effect of the font size difference and word length agreement. We saw main effects for both font size difference ($F(1,174) = 24.68$, $p < 0.0001$) and word length agreement ($F(2,174) = 6.09$, $p = 0.003$). Specifically, participant performance decreased when the difference in word length *disagreed* with the difference in font size, as well as when the difference in font size was smaller (see Table 1). A post hoc test using Tukey's HSD showed that accuracy the "disagree" condition ($M = .83$, 95% CI $[.79, .87]$) was significantly different from both the "neutral" condition ($M = .91$, 95% CI $[.88, .95]$, Cohen's D of 0.47) and the "agree" condition ($M = .91$, 95% CI $[.88, .95]$, Cohen's D of 0.46), though the "neutral" and "agree" conditions were not statistically distinguishable from one another (see Table 1).

For data where there was no correct answer, we tested to see if the rate at which participants picked the *longer* of the two words was significantly different from chance. Specifically, we calculated the rate at which each participant picked the longer of the two words when the font sizes were the same ($M = 0.59$, $SD = 0.17$) and ran a two-tailed, paired Student's t-test to compare these values against an equally sized collection of values of 50%. We found that participants were significantly more likely to pick the longer of the two words ($t(30) = 2.99$, $p = 0.005$), indicating the same direction of bias as seen with the data with correct answers.

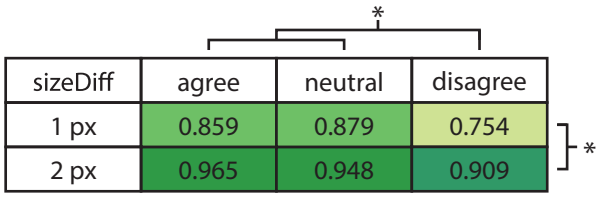

Table 1: This table shows the average participant accuracy for each combination of factors. A two-way ANOVA showed significant main effects for both size difference and length agreement. A post hoc Tukey's HSD test showed that the "disagree" condition (i.e., when the longer of the two words had the smaller font size) was significantly different from the "agree" and "neutral" cases, though the latter two were not distinguishable from one another.

4 Discussion

In this experiment, we saw a consistent bias towards longer words. Word length, it appears, does affect user perception of font size. However, user accuracy was higher than we had been anticipating, with each combination of factors having a mean accuracy of greater than 75% (see Table 1). Even at very close font sizes, participants did notably better than chance. Therefore, while a bias does seem to exist, there may be cause to *trust* user perceptions of font size encodings.

However, the number of letters is just one of many features that factors into the diversity of shapes words can make. We are already analyzing data from experiments exploring these other features, including: differences in the height of individual letters; the different effects of width, number of letters, and area; and adjustments to the encoding that counteract this perceptual bias.

Acknowledgements

This work was supported in part by NSF award IIS-1162037, a grant from the Andrew W. Mellon Foundation, and funding from NSERC and the Canada Research Chairs program.

References

[1] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, pages 193–202. ACM, 2008.

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[3] R. Brath and E. Banissi. Evaluating lossiness and fidelity in information visualization. In *IS&T/SPIE Electronic Imaging*, pages 93970H–93970H. International Society for Optics and Photonics, 2015.

[4] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 160–160. IEEE, jan 2008.

[5] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-Computer Interaction–INTERACT 2009*, pages 392–404. Springer, 2009.

[6] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 995–998. ACM, 2007.

[7] A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5274–5278, 2004.

[8] C. Trattner, D. Helic, and M. Strohmaier. Tag clouds. In *Encyclopedia of Social Network Analysis and Mining*, pages 2103–2107. Springer, 2014.