# Guidelines for Effective Usage of Text Highlighting Techniques

Hendrik Strobelt, Daniela Oelke, Bum Chul Kwon, Tobias Schreck, Hanspeter Pfister

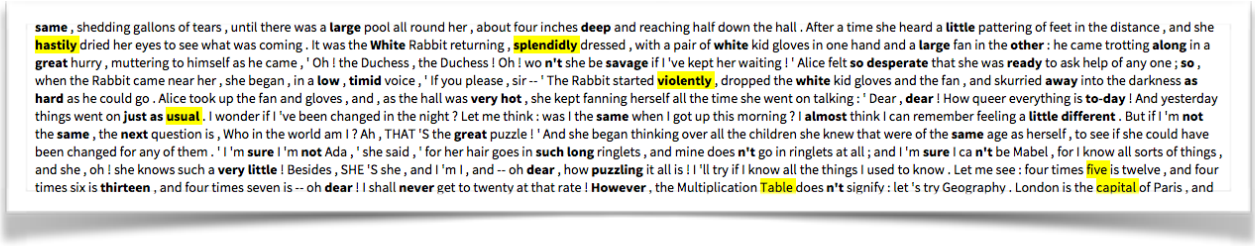


Fig. 1: Text highlighting techniques are commonly used to mark text features in documents. In this excerpt of "Alice in wonderland" all occurrences of adjectives and adverbs derived from part-of-speech tagging are highlighted in bold font, while words with typical adjective/adverb endings are highlighted with yellow background.

**Abstract**— Semi-automatic text analysis involves manual inspection of text. Often, different text annotations (like part-of-speech or named entities) are indicated by using distinctive text highlighting techniques. In typesetting there exist well-known formatting conventions, such as bold typeface, italics, or background coloring, that are useful for highlighting certain parts of a given text. Also, many advanced techniques for visualization and highlighting of text exist; yet, standard typesetting is common, and the effects of standard typesetting on the perception of text are not fully understood. As such, we surveyed and tested the effectiveness of common text highlighting techniques, both individually and in combination, to discover how to maximize pop-out effects while minimizing visual interference between techniques. To validate our findings, we conducted a series of crowdsourced experiments to determine: i) a ranking of nine commonly-used text highlighting techniques; ii) the degree of visual interference between pairs of text highlighting techniques; iii) the effectiveness of techniques for visual conjunctive search. Our results show that increasing font size works best as a single highlighting technique, and that there are significant visual interferences between some pairs of highlighting techniques. We discuss the pros and cons of different combinations as a design guideline to choose text highlighting techniques for text viewers.

**Index Terms**—Text highlighting techniques, visual document analytics, text annotation, crowdsourced study



## 1 Introduction

Automatic text processing is an important research area in data analytics, because a large part of all data occurs as natural language text. Computational linguists and text mining experts strive to train computers to process text in a semantically meaningful way and have been able to report impressive advances within the last decade. Visual document analysis brings the expert into the loop and provides means to process text data when fully automatic processing is not (yet) possible. Furthermore, interactive visual interfaces allow users to browse and explore document collections.

Within the visualization community, previous work has mainly presented abstract visualizations that summarize documents according to certain properties of interest. In contrast to this, we focus on the effective usage of visual means to highlight certain words or phrases directly in a text. Popular highlighting techniques from text typesetting include background coloring, changing the font weight (bold face), and underlining words and phrases of interest.

Text highlighting is important in any scenario where close reading (sequential word-by-word reading) is required and text annotations exist, that should be made accessible to the reader. Imagine a smart writing assistance tool. To provide feedback to the user, spelling errors, stylistically-inappropriate terms, redundancies, and difficult vocabulary have to be marked. The identification of such text properties is what we call an *annotation*. Text annotations can for instance be stored in an XML document. Thus, for our work it does not make a difference if the annotation was added manually or computationally. The single *annotation type* (e.g., spelling errors, difficult vocabulary) is also called a *text feature*. We speak of *text highlighting techniques* if we refer to the visual markup (e.g., bold typeface, background coloring, etc.) that is used to make the annotation visible for the user.

If only a single text feature (e.g., all important phrases) is to be highlighted, then solutions such as applying background coloring will create the desired pop-out effect. However, the problem becomes more challenging if multiple annotations have to be made in a document, or if the highlighting techniques are intended to convey information about an underlying categorical or quantitative variable. In addition, text may also include author-intended highlights in underlines and bold typefaces, which act as design constraints.

To choose proper highlighting techniques, it is necessary to assess how strong the pop-out effect is for each annotation, and how effectively annotations can be used in combination with one another. From perception theory it is known that visual low-level features can interfere with each other, and this must be considered to avoid masking information to the low-level visual system [14, 33]. Furthermore, not all highlighting techniques can be used in combination if overlaps exist, e.g., when a word has two or more annotations and multiple highlighting techniques need to be applied. The contributions of this paper are:

1. A detailed requirement analysis and classification of the most common visual markups for text highlighting with respect to those requirements (see Section 3).
2. A ranking of the visual markups with respect to their effectiveness for highlighting text, determined by a perception study (see Section 5).
3. A study examining the degree of visual interference of different

*Hendrik Strobelt and Hanspeter Pfister are with Harvard University, email: [hstrobelt,pfister]@seas.harvard.edu*

*Daniela Oelke is with Siemens AG (affiliation when the paper was written: German Institute for International Educational Research, DIPF)*

*BC Kwon, the corresponding author, is with University of Konstanz, email: bumchul.kwon@uni-konstanz.de*

*Tobias Schreck is with TU Graz, email: tobias.schreck@cgv.tugraz.at*

text highlighting techniques (see Section 6).
4. A study examining the effectiveness of the combination of two techniques for visual conjunctive search (see Section 7).
5. Application examples and guidelines which show how the results can be employed in practice in various scenarios (see Section 8).

## 2 Related Work

This section discusses related work about visualization of annotations in documents - the task to which our study results can be applied. This is followed by a review of other, more general works on assessing the perception of visual properties and means for visual boosting.

### 2.1 Document annotation viewers

The research area of Visual Document Analysis deals with supporting the analysis of single documents or document collections by means of a tight integration between automatic natural language processing algorithms and effective visualization methods. This usually involves summarization and abstraction of the data to provide an overview regarding some text property of interest. These works deal with questions like how the 'black box' of automatic text processing can be opened [16, 31], or how higher-level representations of a document can be created, e.g., showing the development of text properties within a text [9, 17] or a summary of the content of a document [30]. Other works consider how whole document collections can be inspected (see document landscapes [34] or techniques that show the development of topics within a collection over time [23]). For a summary of visual document analysis techniques, see [21] or [2].

In contrast, our goal is to visualize document annotations directly in the text to allow close reading. This is a requirement in many text analysis tasks, such as traditional text analysis methods within Humanities or Social Sciences, among others. Viewing annotations directly in the text is also necessary when working with more sophisticated visual document analysis systems that abstract from the data to verify findings of interest ([7, 9, 12, 22] all contain a document viewer).

Related work of tools that support close-reading of documents include the VarifocalReader [19]. Koch et al. suggest an approach that aims to provide access to a document at multiple levels of detail from higher aggregation to the text level directly. A key feature of the tool is that the different levels of detail can be navigated smoothly in parallel. Similarly, Correl and Gleicher developed a visual tool for literacy scholars to annotate phrases with multiple definitions and to explore these phrases [7].

In addition to tools developed by the visualization community, text viewers in other domains exist. The Brat rapid annotation tool [28] provides support for manually adding structured annotations and labels, and can also deal with relations between word phrases. The text mining tool GATE [8] uses background coloring of words together with an annotation stack view. QDAMiner [26] is a tool for computer assisted qualitative analysis which allows the user to annotate subsections of documents, where annotations are then shown next to the document. In Egas [10], concept names are colored with rectangular boxes that can be nested. Relations are shown as directional lines.

To the best of our knowledge, no in-depth study of the effectiveness of typeset text highlighting techniques has been conducted. Instead, most document viewers within text analysis systems solely employ coloring as a highlighting technique (mostly background coloring), which falls short when multiple overlapping annotations or non-binary annotation types are to be shown.

Tools that are not based on word coloring include the Ink Blot technique [1]. This overlays text with colored circles which visually encode the weights of key features assigned by a text classification system. Stoffel et al. [29] apply a distortion algorithm to highlight (boost) text passages of interest. Both techniques cannot be easily applied to text documents in standard text editors and are therefore excluded from the study.

### 2.2 Assessing perception of visual properties

Our work extends previous studies on the perception of visual variables in general. In 1984, Cleveland and McGill [5] researched the accuracy by which different visual variables can be perceived. In [6], they presented a ranking of visual features to provide guidance for designing graphics with well-perceivable features. Healey and Enns [14] researched how textures and color interfere with each other. Mackinlay et al. [24] show how understanding the effectiveness and interference between visual features can feed into effective automation of optimal presentation of visual results. Those general studies on perception are a good basis for our work in which we specialize on perception of text highlighting techniques. A good summary of perception studies and their results can be found in [13] and in [33].

Crowdsourced experiments have recently become an effective method for evaluating perception in Information Visualization [15, 20]. Prior studies confirm that results from crowdsourced perception studies are comparable to lab-based studies [3, 15]. Despite their effectiveness, care must be taken to remove noise in the data from non-serious study participants (random clickers) [18]. In addition, the task design should be straightforward and easy, to ensure participants are well-prepared for the given tasks [11]. We follow the advice in these studies for our own study design.

## 3 Analysis of requirements and exploring the design space

Here, we first shed light on requirements for text highlights, before framing the design space for highlighting techniques.

### 3.1 Requirement analysis

Natural language processing (NLP) researchers are (among others) concerned with the automatic annotation of documents with respect to certain text properties. We therefore decided to informally interview five researchers working in different natural language processing (NLP) projects to further understand the requirements for text highlighting. We interviewed each of them separately, starting by asking them to explain their project and to provide us with the context in which they would use text highlighting. Furthermore, we asked them to name as many text features as they could that are important for their task. The interviews and paper reviews led us to the following insights:

- Many text features exist: statistical text features (word length, sentence length, verb/noun ratio, number of smileys, term frequencies, n-grams etc.), syntactic features (parse tree, sentence structure, active vs. passive voice, co-references, etc.), semantic features (sentiment signal words, term-topic associations, etc.), and structural features (font size of header, width of page margin, etc.). Often domain-dependent semantic annotations are added (highlighting of proper names) as well.
- Text features can be boolean (negation words), categorical (part-of-speech (POS) tags), or quantitative (word length). The number of different categories can be high (the tag set of the Penn Treebank contains 36 POS tags for the English language). Sometimes, the quantities and different categories are not of interest, which means that text features can be treated as boolean.
- Text annotations may be at the level of characters (word endings), tokens (word length), sentences (exclamations), whole paragraphs (co-reference chains), or the whole document (text genre).
- If a text is annotated with multiple text features, the annotations can overlap in the text.
- In some cases, relations between words have to be shown, e.g., co-reference chains, dependency parsing results, etc. In this paper, we ignore this requirement because these types of annotations cannot be displayed with common highlighting techniques and need special visualizations, e.g., link relationships [27]. The study of links between text portions remains subject of future work.

### 3.2 Design space

The design space spans the variety of highlighting techniques and their usage for different kinds of data. We elaborate on the techniques first (Section 3.2.1) and discuss annotations especially for categorical and quantitative data (Section 3.2.2).

Table 1: Common text highlighting techniques with typical parametrization. The last column indicates which variations were used in our study. The 'Use' column indicates if the technique should be used for categorical (c) or quantitative (q) data. Trivially, all highlights can encode binary data by using absence or presence of the technique.

| Technique | Use | Typical variations | Used in our studies |
|---|---|---|---|
| Font color | c q | Saturation, luminance, hue | Red color (`rgb(227,26,28)`) |
| Background color | c q | Saturation, luminance, hue | Bright yellow (`rgb(255,255,50)`) |
| Underlined | c q | Styles, thicknesses | Single underline |
| Font size | - q | % increase | 150% increase |
| *Font style* | - - | Italics, subscript,... | Italics |
| **Font weight** | - - | Font weight | bold font |
| Rectangular border | c q | Styles of border, lines, thickness | Single border |
| S p a c e d o u t f o n t | - q | Letter spacing | 5px spacing |
| Text shadow | - - | Offset, intensity,... | CSS: `text-shadow: 4px 4px 3px rgb(50, 50, 50);` |
| Font family | (c) - | Sans-serif, Times, Helvetica,.. | — |
| CAPITALIZATION | - - | Small caps, large caps | — |
| ~~Strike-through~~ | - - | True, false | — |
| * Blinking * | - - | True, false | — |

#### 3.2.1 Text highlighting techniques

The number of possible text highlighting techniques is large, and so we restrict ourselves to a set of common techniques that can be easily realized in a web browser with HTML, or in common text processing environments such as Word or LaTeX. A list of common highlighting techniques is given in Table 1.

In the study presented in Section 5, we tested the visual saliency of all highlighting techniques in Table 1 except for:

- Blinking: Motion attracts attention, but it is also known to be disturbing or intrusive [33]. Furthermore, blinking cannot be used in a static environment like paper.
- Font family: Changing the font family also implicitly changes other font attributes like letter spacing, the degree of tilting of letters, or the boldness of letters. Therefore, it significantly interferes with other highlighting techniques if used in combination with them.
- Capitalization / small caps: Often this technique cannot be used because the original text already contains capitalized words or capitalized abbreviations.
- Strike-through: This highlighting technique comes with inherent semantics that are not appropriate in many cases, e.g., its text may be interpreted as being wrong or unwanted.
- Color choices for font and background: We cannot test all colors in the scope of this project; instead, we choose to use red text and yellow background in this study.

#### 3.2.2 Categorical and quantitative data

Highlighting annotations with underlying categorical or quantitative data needs special consideration. Perception theory teaches that "for the pop-out effect to occur, it is not enough that low-level feature differences simply exist, they must also be sufficiently large" (C. Ware in [33], page 31). This can conflict with the requirement that the different values of a categorical variable should be perceived as a group, and therefore be more similar to each other than to all other highlighting techniques used.

For categorical data, highlighting techniques include:

- Different hues of text or background colors, e.g., red, green,
- Different underline styles, e.g., solid, dotted, dashed, double,
- Different borderline styles, e.g., solid, dotted, dashed, double,
- Different font families (though discouraged as mentioned).

Care must be taken to choose variations of the highlighting techniques in a way that maintains similar perceptual saliency to avoid visual boosting of certain categories. If the categorical variable is the only one displayed, then the requirement that the highlighting techniques used should visually group becomes unnecessary. This means that different categories can be treated as boolean variables and be encoded with any of the other available techniques.

In a real-world scenario, certain text features may have a large number of categories (see Section 3.1). This conflicts with the limitation of the number of distinguishable variants of highlighting techniques. Even considering colors whose variations are theoretically unlimited, it is known that only a certain number of different shades can be distinguished effectively [25].

For quantitative data, highlighting techniques include:

- Font size,
- (Luminance / Saturation of) font or background colors,
- Thickness of underlines,
- Thickness of borderlines of frames,
- Degree of letter spacing.

An increase in size results in a more distinctive highlighting technique than a decrease in size ([33], page 35). Furthermore, though theoretically an unlimited number of intermediate steps from the larger / thicker / darker state are possible, in practice only a limited number of steps can be distinguished.

We concentrate on the scenario of highlighting boolean text features (annotated vs. not annotated), and leave the in-depth analysis of highlighting categorical and quantitative variables for future work.

## 4 Study Design

Having framed the design space, we focus our user study on text highlighting of boolean text features. We conducted three user studies via Amazon Mechanical Turk.

**Study 1** analyzes each highlighting technique in isolation for its performance for identifying boolean highlights in a given text. It results in a ranking of highlighting techniques with respect to the strength of their pop-out effect, and orders them on a scale from strong to weak. For a strong highlighting technique, the probability is higher that users can accurately detect the highlighted texts. For weak techniques, the probability is lower. The labels **weak** and **strong** for each technique are used as reference in the remainder of the paper. The insights of this study can be applied to scenarios where only one text feature is highlighted, e.g., to highlight text search result on a webpage text in a browser, but also for building combinations of multiple highlighting techniques. The study is designed so that each user is faced with a continuous text which mixes two text variants: standard text and highlighted text. The task is to find as many highlights as possible in a given time. Section 5 discusses setup and results in detail.

**Study 2** analyzes each highlighting technique for its performance for identifying boolean highlights in a given text, when the user is distracted by terms also being highlighted with a second technique, either

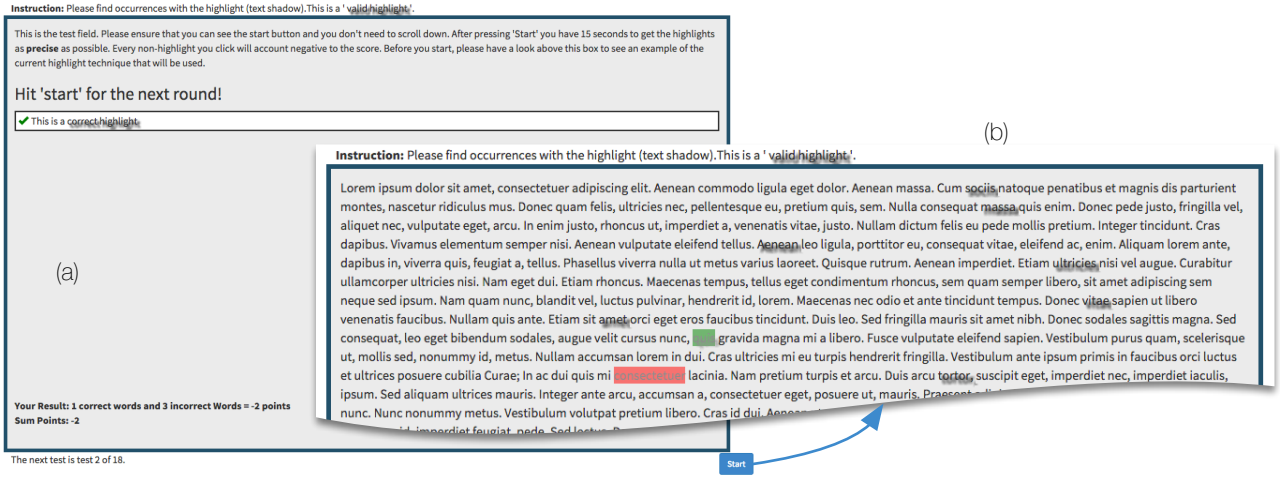


Fig. 2: Screenshots of the developed evaluation tool (used for all three studies). The target highlighting technique here is *shadow*. a) Start page introducing highlighting technique of next trial, and showing results for previous trial. b) A test page. Terms that were clicked on are marked to provide visual feedback. See the supplementary material for detailed figures of the test system.

alone or in combination with the studied technique. The driving questions are: How do weak and strong techniques interfere when being used in the same text? How about two strong techniques? Does a weak distractor result in a smaller decrease in performance compared to a strong distractor when searching for the other highlighting technique, or vice versa? The insights from this study are applicable to scenarios where two highlighting techniques operate on the same source, e.g., in collaborative annotation of text between two proof readers. For the study, each user had to identify highlights of type A while being distracted with technique B. The continuous text is now assembled from four text variants: highlighted texts of classes A, B, A+B, and plain text. The task is to find all highlights of type A in a given amount of time. An overlap of both techniques does count as incorrect because technique A is mixed with the distracting technique B. Section 6 describes details on setup and discusses results for techniques acting as target (A) or distractor (B).

**Study 3** analyzes visual conjunctive search when using combinations from our set of techniques. The goal is to find out how two highlighting techniques used together can be spotted, when being distracted by each contributing technique alone. These results allow us to check whether a combination of techniques generates more pop-out than its individual parts; or, if the combination performs equal or more poorly. A typical scenario is a situation where spotting the overlap of highlighting techniques is the primary goal and single highlights act only as secondary information. In the study, the user is faced with the same text configuration as described for Study 2, but this time the user must find only the overlap of highlights. More details are given in Section 7.

The following three sections provide details on each study, while Section 8 discusses practical implications and applications.

## 5 Study 1: Ranking of Text Highlighting Techniques

The goal of Study 1 was to establish a ranking of text highlighting techniques with respect to the strength of their pop-out effect.

### 5.1 Setup

In total, we recruited 63 participants from Amazon Mechanical Turk for this study, with the following recruiting specification: Compensation: $1.50, Turker requirement: 10,000 HITs or more approved, 99% HIT Approval Rate. Among 63 participants, 18 participants were excluded for the following reasons: did not complete all trials (n = 14); used a tablet (n = 2); failed in a color blind test (n = 2). We excluded tablet users because the touch on a screen to complete the task is significantly different from clicking on a target with a mouse. We excluded people who failed the color blind test to make sure all participants in our study can differentiate highlighted text from plain text. Thus, 45 participants were included for analysis (Gender: 23 males, 22 females; Age: 19 in 20-30 years old, 23 in 30-60 years old, 3 in 60+ years old).

Figure 2 shows a screenshot of the evaluation tool. In each trial, the user is presented a text in which 20 randomly chosen terms are highlighted with one of the highlighting techniques. We used artificial text without any semantics (*Lorem ipsum - Text*) to make sure that participants concentrated only on the text highlighting. A freely available sans-serif font ('Source Sans Pro'[1]) of size 14px was used with regular line spacing. The *lorem ipsum* text had a length of 673 words (4633 characters) and was presented in a text box of size 1000px × 600px. Each participant was tested three times with each highlighting technique, i.e., 27 trials in total per participant. Given approximately 60 participants, our pilot study determined two repetitions per technique would detect the significant effect, and we added an additional one to make sure that we could avoid any learning curves or fatigue effect. The selection of words within the text in each trial was randomized. The order in which the highlighting techniques were presented was randomized. The experiment required a minimum screen resolution of 1070 x 700, which was enforced by a start button placed at the lower right corner. It could only be reached if the screen resolution was sufficient (page scrolling was blocked).

Each trial consisted of a) a start page in which the highlighting technique that needed to be searched for was introduced (see Figure 2, left) and b) the actual test page that was shown after the participant pressed the start button (see Figure 2, right). In each trial, we recorded the number of highlighted terms (words) that were correctly clicked by a user. In addition, we also recorded the number of incorrectly identified terms. This permitted us to filter out random clickers or robots that presumably would have had a high number of incorrect hits. The task for each participant was to click on as many of the highlighted terms as possible within 13 seconds. The duration for each trial was chosen so that the timespan was too short to click on all 20 terms (even for highlighting techniques with a strong pop-out effect), but large enough not to bias participants with a short attention span. Clicked terms were marked to provide visual feedback to the participant. Participants were given a break as long as they wanted between trials. In total, participants took 35 minutes on average to complete all trials.

[1] https://www.google.com/fonts/specimen/Source+Sans+Pro

### 5.2 Results

Due to the mechanical aspect of the task (clicking multiple items within a time period), we observed two types of unwanted variation in the results. *Individual Difference:* The results show significantly different performances between individuals. We expected some differences between their perceptual abilities and clicking speeds, which are inherent to individuals. Thus, we normalized responses with respect to their performance range. In this way, we maintained the performance effect of highlighting techniques while mitigating unwanted variation from individuals. *Learning Curve:* The results showed that participant performance in the first trials was significantly lower than in the following two. Thus, we excluded the first trials of all participants from the analysis. We did not observe any fatigue effects.

We analyzed the normalized correctness, i.e. the score, using an analysis of variance (ANOVA). We found a significant effect between techniques, $F(8, 801) = 171.5897$, $p < .0001$. Post hoc analysis using Tukey HSD showed the differences between individual techniques (Figure 3). Font size was higher than the rest except for border. The bottom three (the weakest) techniques were underlined, letter spacing, and italic typeface. In particular, italic type face had a very low mean score, significantly lower than the rest of the techniques.

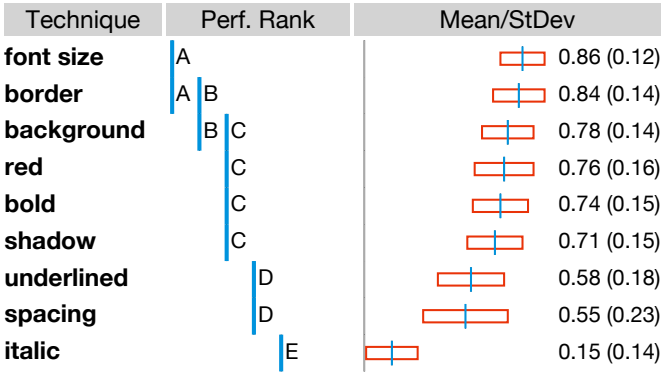


Fig. 3: Performance rank of nine text highlighting techniques (Study 1). The *Perf. Rank* Groups were defined based on results of pairwise comparison between all techniques using the Tukey HSD test ($p < .05$). How to read: Performance was significantly higher for techniques with earlier alphabetical ranks, e.g., A > B, C∼D > E. Performance had no significant difference for techniques sharing alphabetical ranks, e.g., A ≅ A∼B, A∼B ≅ B∼C.

### 5.3 Discussion

Researching in-depth details of *why* highlights perform in the presented order is beyond the scope of this paper. Instead, we provide a set of hypotheses that are subject to verification (or falsification) in future work.

For the top four ranked highlighting techniques (font size, border, yellow background, red text) we think the following hypotheses can point to answers why they perform well:

- Text features that are encoded with increased **font size** stick out from the cap line of the surrounding text. They also fill the white space between lines and could therefore be perceived as an interruption.
- A word surrounded by a box (**border**) stands out more than a word underlined because the box might make the target appear bigger, thereby making it easier to detect. The size can be an important feature that makes words easier to detect because the task becomes detecting something bigger than normal.
- Color is known to have a strong pop-out effect (provided that contrasting colors are used). If the background does not vary with respect to color (which is the case when black font is printed on white background), then the additional color can be considered as a new visual characteristic that can be effectively biased for. **Background coloring** may have received a higher ranking than coloring the **font in red** because the colored area is much larger (and therefore more prominent).

For the two lowest ranked highlighting techniques (letter spacing, italics) our hypotheses are:

- The characteristic feature of **letter spacing** is that additional empty space is introduced between the characters. However, empty space is a normal feature within a text (it exists between every two words) and is not exclusive to letter spacing. Therefore, the feature-level contrast to the background is rather low.
- The characteristic feature of **italic typeface** is that the characters are all slanted. But the resulting new angles of the lines are not a unique feature that would effectively discriminate terms in italic typeface from the ones in normal typeface. Instead many characters also contain slanted lines without being printed in italic typeface, e.g., “X”, “Y”, “Z”, “A”, “R”, “V” etc.

## 6 Study 2: Search with Distractor

Healey and Enns note that “Certain combinations of visual features cause interference patterns that mask information in the low-level visual system” (page 150, [14]). The goal of Study 2 was to determine how much the different techniques interfere with each other when used in the same text. Study 2 investigated how easy or difficult it is to search for terms that were only highlighted with one of the two techniques. Note that visual interference is asymmetric [13, 14, 32] and therefore has to be tested with each technique as a target.

### 6.1 Setup

In Study 2 the participants were instructed to choose text highlighted by a target highlighting technique (A), where there exists a distracting highlight (B), a combination of target and distracting highlights (A+B), and plain text. In each trial, we provided twenty highlights each for A, A+B and B, to maintain a consistent number of correct highlights. This task aimed to test the pop-out effects of a highlighting technique A in presence of another text highlighting technique. In total, a participant was given the entire permutation of pairs of the nine highlighting techniques (72 trials). We used the same setup as Study 1 for this study.

We recruited 38 participants from Amazon Mechanical Turk for this study, with following specification: Compensation: $3.00; Turker requirement: 10,000 HITs or more approved, 99% HIT Approval Rate. Among the participants, 8 were excluded for the following reasons: did not complete all trials (n = 7); failed in a color blind test (n = 1). Thus, 30 participants were included for analysis (Gender: 14 males, 16 females; Age: 1 in -20 years old, 7 in 20-30 years old, 21 in 30-60 years old, 1 in 60+ years old). We used the same procedure and web platform as in Study 1. Instead of having additional repetitions as in Study 1 (Study 1 only had 27 trials in total), which makes the entire tasks for crowdsourced participants appear time-consuming and effortful, we added ten trials as a training session at the beginning with randomly selected combinations to avoid learning curves. After the experiment, we confirmed that there was no learning curve or fatigue effect. Furthermore, we closely inspected individual cases because there might have been some participants who were misinformed about the task, e.g., choosing A *and* A+B instead of only A, but no participants showed any evidence of having been misinformed. Participants took on average 73 minutes to complete all trials.

### 6.2 Results

We analyzed normalized correctness using an analysis of variance (ANOVA). We found significant effects for highlighting techniques, $F(8, 2114) = 236.61$, $p < .0001$, and distractors, $F(8, 2114) = 65.60$, $p < .0001$. Post hoc analysis using Tukey HSD shows the differences between individual techniques (see Figure 5). In general, all techniques (except for italic typeface) decreased in performance from Study 1, which was expected since a distractor had been added. Besides italic typeface, underlined was affected the least by the presence of a distraction (-12%), which made its performance rank significantly higher

| | distractor technique --> | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **fs** | **bo** | **bg** | **red** | **bold** | **sha** | **und** | **spa** | **it** |
| font size | | **-13.3** | **-9.2** | -4.4 | **-42.8** | **-11.1** | **-25.1** | **-48.2** | **-38.3** |
| **bo**rder | **-21.3** | | -6.0 | -5.5 | **-8.1** | **-9.6** | **-39.9** | **-30.0** | **-37.5** |
| **b**ackground | **-11.9** | **-13.8** | | **-14.9** | -6.4 | **-12.7** | **-20.7** | **-28.6** | **-33.3** |
| **red** | **-14.7** | -8.8 | 2.7 | | **-14.2** | **-16.6** | **-23.4** | **-28.3** | **-32.8** |
| **bold** | **-40.7** | **-13.4** | 0.3 | 3.4 | | **-13.1** | **-17.4** | **-23.0** | **-30.2** |
| **sha**dow | **-16.7** | -9.4 | -1.7 | -1.3 | **-11.9** | | **-39.5** | **-19.2** | **-42.3** |
| **und**erlined | **-18.6** | **-20.3** | 3.1 | 7.9 | -6.4 | -9.6 | | **-27.2** | **-28.8** |
| **spa**cing | **-35.9** | **-31.2** | -6.0 | -4.3 | **-23.3** | **-17.9** | **-30.8** | | **-49.3** |
| **it**alic | 30.2 | **55.2** | **93.3** | **60.3** | **46.9** | **40.5** | 18.5 | 2.6 | |

Fig. 4: Percentage changes in performance of target highlighting techniques in Study 2 as compared to Study 1. All reported performance gains / losses are relative to the technique in the rows. **Bold** cells show significance at 0.05; **bold and underlined** cells show significance at 0.01.

than letter spacing even though they were equal in Study 1. The rank order was almost identical, except for the switch between font size and border. Scores of font size and border were no longer significantly higher than yellow background and red text.

For detailed analysis, Figure 4 shows how much each technique (row) gained or lost from the existence of the second technique (column) in comparison to the results from Study 1. For example, the cell value (-13.3%) of the first row and the second column shows that when font size is used as a main target with border as a distractor, the performance decreases by 13% from when font size is used without any distractors. Red color shows decrease, while blue shows increase. Bold font shows significance. Toward the top right corner of the table, we see significant percentage change decreases, especially for red, bold, and underline. When the four techniques, text shadow, underlined, letter spacing, and italic are used as distractors on techniques ranked higher than the distractors, we can expect a significant decrease in pop-out effects. When we take a look at the table columns for background and red color used as distractors, we see that these do not have a statistically significant influence, except for two combinations of techniques (font size and italic for background; background and italic for red color).

| Technique | Perf. Rank | Mean/StDev | Deviation |
|---|---|---|---|
| **border** | A | 0.67 (0.22) | -0.17 (-20%) |
| **font size** | A B | 0.65 (0.25) | -0.21 (-24%) |
| **background** | A B | 0.64 (0.19) | -0.14 (-18%) |
| **red** | A B | 0.63 (0.20) | -0.13 (-17%) |
| **bold** | B C | 0.62 (0.19) | -0.12 (-16%) |
| **shadow** | C | 0.58 (0.22) | -0.13 (-18%) |
| **underlined** | D | 0.51 (0.20) | -0.07 (-12%) |
| **spacing** | E | 0.41 (0.20) | -0.14 (-25%) |
| **italic** | F | 0.22 (0.14) | +0.07 (+47%) |

Fig. 5: Performance rank of target highlighting with a distractor (Study 2). The column *Deviation* reports the Deviation of the Mean Score from Study 1 (Percentage Change of Mean Score from Study 1). See caption of Figure 3 for how to read the *Perf. Rank* column.

### 6.3 Discussion of the results

Healey and Enns reported that "background variation in non-target attributes produced small, but statistically significant, interference effects. These effects tended to be largest when target detectability was lowest" (page 153, [14]). Due to this and similar statements in related work, our assumption for Study 2 was that techniques with a stronger (individual) pop-out effect would also be stronger distractors than techniques with a weaker pop-out effect. However, as described in Section 6.2, strong indications of the opposite effect were observed: weak techniques negatively influenced strong ones. One explanation for this might be that our task forced the participants to distinguish A from A+B, which means that all terms highlighted with the strong technique A have to be checked for the existence of an additional highlighting with technique B. This is easier if technique B has a strong pop-out effect itself. This assumption is supported by the fact that far more often A+B was wrongly selected than terms highlighted only with technique B (see Table 2).

Table 2: Error Analysis for using $B$ as distracting technique. All samples after the 10th trial. Insights: $AB > B$ and $AB > else$ for weak techniques, whilst $AB < else$ for strong techniques, but always $AB, B, else << correct(A)$.

| Distracting Technique (B) | source of error | | | correct (A) |
|---|---|---|---|---|
| | AB | B | else | |
| Letter spacing | 175 | 14 | 29 | 2523 |
| Italic typeface | 150 | 2 | 29 | 2469 |
| Underlined | 80 | 7 | 25 | 2601 |
| Bold typeface | 56 | 13 | 42 | 2927 |
| Font size | 55 | 30 | 39 | 2656 |
| Border | 29 | 12 | 34 | 2823 |
| Yellow background | 28 | 18 | 39 | 3241 |
| Text shadow | 22 | 8 | 40 | 3128 |
| Red text | 13 | 3 | 35 | 3208 |

These results likely would have been different if we had asked the participants to pick all terms which had been highlighted with A, whether they are additionally highlighted with B or not. The tedious differentiation between terms highlighted only with the technique that has a strong pop-out effect and terms that are additionally highlighted with a weaker technique is then not necessary anymore. We can assume that in this case the observations of Healey and others would apply, and stronger techniques would be less affected by distractors than weak ones.

Another finding was that both background coloring and term coloring only rarely interfered with other techniques. We assume that this can be attributed to the fact that coloring is visually orthogonal to the other techniques that all directly influence the type face or work with visual features that are an intrinsic part of the typeface, e.g., horizontal and vertical lines as in border or underline. This interpretation is in line with Ware's observation that "to minimize this kind of visual interference (it cannot be entirely eliminated), one must maximize feature-level differences between patterns of information" and in line with his guideline that "as a general rule, like interferes with like" (page 51, [33]).

A surprising result from our study was that one highlighting — *italic* — apparently profited from the distractors. In general, it is known that the more noisy a background is, the more difficult it is to concentrate on a single visual feature ([33]). Thus, our expectation was that no technique would profit from the addition of a distractor.

## 7 Study 3: Visual Conjunctive Search

The task of finding a target composed of two visual features is called a visual conjunctive search (page 31, [33]). The goal of Study 3 was to determine how well the different combinations of highlights perform, compared to their use as single targets.

### 7.1 Setup

The task of Study 3 was to choose A+B against A, B, and plain text, where A and B again denoted two different highlighting techniques. In each trial, we provided twenty highlights each for A, A+B and B, to maintain a consistent number of correct highlights with Study 1 and 2. This task aimed to test the pop-out effect of the combination of two techniques to support visual conjunctive search. In total, a participant was given the entire combination of pairs of the nine highlighting techniques (36 trials). We used the same setup as for Study 1 and 2.

| | fs | bo | bg | red | bold | sha | und | spa | it |
|---|---|---|---|---|---|---|---|---|---|
| font size | | -14.1 | -8.8 | -8.4 | -24.5 | -13.3 | -15.0 | -36.0 | -41.7 |
| border | -12.1 | | -17.9 | -5.3 | -12.5 | -24.5 | -43.5 | -12.2 | -42.8 |
| background | 0.5 | -11.6 | | -25.7 | -7.5 | -27.3 | -29.9 | -19.3 | -39.3 |
| red | 3.6 | 4.7 | -23.8 | | -20.9 | -28.1 | -25.5 | -23.4 | -41.0 |
| bold | -12.2 | -0.7 | -2.5 | -18.8 | | -22.0 | -25.8 | -26.0 | -39.2 |
| shadow | 5.0 | -10.7 | -20.1 | -23.1 | -18.7 | | -43.9 | -30.2 | -51.2 |
| underlined | 26.0 | -18.2 | -5.8 | -2.4 | -5.3 | -31.3 | | 9.8 | -49.4 |
| spacing | 0.0 | 34.1 | 14.5 | 5.8 | -0.4 | -9.9 | 15.7 | | -42.5 |
| italic | 234.4 | 220.2 | 215.7 | 199.0 | 200.0 | 130.8 | 95.8 | 110.8 | |

Fig. 6: Percentage changes of combinations of target highlighting techniques in Study 3 from that in Study 1. All reported performance gains / losses are relative to the technique in the rows. **Bold** cell shows significance at 0.05; **bold and underlined** cell shows significance at 0.01.

In total, we recruited 34 participants for this study, with following specification: Compensation: $2.50; Turker requirement: 10,000 HITs or more approved, 99% HIT Approval Rate. Then, we excluded ten participants for the following reasons: did not complete all trials (n = 7); failed the color blind test (n = 3). Thus, 24 participants were included for analysis (Gender: 15 males, 9 females; Age: 1 in -20 years old, 11 in 20-30 years old, 23 in 30-60 years old). We used the same procedure as before to confirm that there is no learning curve, fatigue effect, or wrongly instructed cases in our data. Participants took on average 63 minutes to complete all trials.

### 7.2 Study Results

Figure 6 shows how much the combination of two techniques gained or lost in performance compared to Study 1. This comparison can be made in two directions: a) Percent of performance gain or loss of technique A compared to the score for A+B, and b) percent of performance gain or loss of technique B compared to the score for A+B. Although the matrix of absolute scores is symmetric (because the score for A+B = score for B+A), the matrix with the percentage changes is not. The reported percentage changes are always relative to the technique reported in the rows. For example, a value of -12.1 in row "border" and column "font size" means that the score for border+font size is 12.1% lower than the score for border without any distractor as determined in Study 1. Conversely, the value of -14.1 in row "font size" and column "border" means that the score for border+font size is -14.1% lower than the score for "font size" without a distractor.

In accordance with Study 2, the upper right triangle shows that when the bottom (weakest) four techniques were combined with the higher ranked techniques, the performance was consistently lower than when just using the higher techniques alone, with the exception of one case: underlined+spacing. In contrast, the lower left triangle shows less significant changes, with a few combinations gaining in performance for underlined and spacing. An exception is the generally weakly performing italic technique, which gains when combined with any other of the studied techniques. We also see that there is only one combination for which both techniques have a gain in score, when combined with the other one: underlined and spacing (although not significantly).

### 7.3 Discussion

On the one hand, Ward et al. state: "If we want to search rapidly for combinations of data values, care must be taken to ensure that the resulting combinations contain at least one unique feature for the visual system to cue on" (page 104, [32]). On the other hand Ware finds that "most visual conjunctions are hard to see" (page 31, [33]). If a combination itself does not have a pop-out effect, the task results in a serial search, focusing first on one technique and then filtering those candidate terms for visual conjunctions with the second technique [33].

In our experiment, only the combination underlined+spacing achieved a performance gain relative to both techniques. We can assume that this combination results in a new unique visual feature that can be biased in the visual conjunctive search. Likely, the unique visual feature is the empty underlined space in the terms (Figure 7).

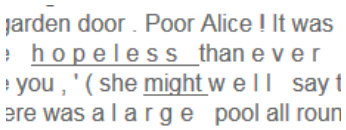


Fig. 7: Example for underlined and letter spacing.

All other combinations are asymmetric or result in a loss of performance for both techniques. The significant losses, especially when a technique with a strong pop-out effect is combined with one with a low rank in Study 1, can be explained by the fact that biasing for the strong technique is fast, but the slower the subsequent filtering step to restrain the result to those terms highlighted with both techniques, the weaker the pop-out effect of the second technique.

On the other hand, the gain of weak techniques by being combined with high ranking techniques can be explained by the fact that the first step of the sequential search, by biasing for the strong technique, reduces the number of candidates considerably compared to Study 1.

## 8 Putting Things into Practice

Section 8.1 provides guidelines and recommended combinations of highlighting techniques for common scenarios. Following this, Section 8.2 explains how we derived the recommendations from the study results and illustrates how the detailed matrices can be used to take project specific requirements into account when selecting appropriate combinations. Section 8.3 demonstrates the usefulness of the results in two concrete application scenarios. Finally, we conclude with a discussion of limitations.

### 8.1 Recommended highlighting techniques

In the following we provide guidelines for which techniques to use in the most common annotation scenarios.

| | |
|---|---|
| **Scenario 1** | Only one feature must be highlighted. |
| Guideline | Choose a highlighting technique with a strong pop-out effect. In our test font size, borders, and yellow background scored best with some others following closely (see Figure 3 for details). |
| **Scenario 2** | Both features should have the same visibility, visual conjunctive search is not important. |
| Guideline | Choose highlighting techniques that do not interfere much with each other and have a strong pop-out effect of similar strength. This is for example the case for bold+ yellow background, border+red, font size+yellow background, font size+border. |
| **Scenario 3** | The conjunction of the two features is more important than their single occurrence. |
| Guideline | Choose two techniques that scored high in the visual conjunction test of Study 3. This is for example the case for border+red, font size+red, and font size+yellow background. |
| **Scenario 4** | One feature is significantly more important than the other and should stick out. |
| Guideline | Choose the two techniques in a way that one of them has a significantly higher pop-out effect than the other. Try for instance yellow background+spacing, font size+underlined, border+italic. |

| | |
|---|---|
| **Scenario 5** | Both features should have the same visibility and the conjunction of the two should be easy to see. |
| Guideline | Choose highlighting techniques that do not interfere much with each other and have a strong pop-out effect of similar strength. Additionally, their visual conjunction should be easy to detect. Good candidates are border+red, font size+yellow background, and yellow background+bold. |

### 8.2 How to make use of the detailed matrices

Although the percentage changes (see Sections 6 and 7) were very helpful for understanding the results of the study, matrices with absolute performance values are more informative. We therefore provide matrices with the absolute values of Study 2 in Figure 8 (referenced as M2) and for Study 3 in Figure 9 (referenced as M3). The absolute values for the results of Study 1 are included in Figure 3. In the following we explain how we derived the recommendations in Section 8.1 from the study results and illustrate how the detailed matrices can be used to take project specific requirements into account.

While single good performing highlighting techniques can be directly read from Figure 3 (as needed for Scenario 1), finding good *combinations* of highlighting techniques (as in Scenarios 2-5) requires a deeper analysis of the study results. To derive good highlighting techniques for Scenario 2 (both features should have the same visibility), we first calculated the delta of the matrix M2 (absolute values for study 2) and its transpose (see Figure 10). The lower the delta between two techniques is, the more similar their perceptual strength is. Given only this criteria, also weak combinations like underlined+spacing would be ranked high. Therefore, we additionally have to take the performance of the techniques when used in combination into account. This can be read directly from matrix M2 (Figure 8). In this case the values for both, A vs. B and B vs. A should be as high as possible to ensure that one technique does not dominate the other. The four recommendations for good combinations mentioned above were derived by requiring the delta value to be below or equal to 0.1 and the absolute performance values to be above 0.65. Note that those values are to a certain degree arbitrary and were selected in a way that a set of 3-4 high scoring combinations could be found.

To derive good combinations for Scenario 3 (visual conjunction is most important), only matrix M3 (Figure 9) must be consulted for high performance values.

Scenario 4 (one feature is significantly more important than the other) requires the selection of the top ranked techniques and one of the low ranked techniques (what is top or low ranked can be read from Figure 3). Note that matrix M2 cannot be used for choosing appropriate techniques in this case. In Study 2 we asked the participants to select only terms that are highlighted with highlighting technique A, not the ones highlighted with A+B. In contrast to this, in Scenario 4 both would be hits. As described in Section 6.3 we hypothesize that the combination with weak techniques slowed down some otherwise highly performant techniques, because extra time was needed to check if the weaker markup is present, too. This, however, would not have happened if the task was to select both, A and A+B and therefore the results are not informative for this Scenario.

Scenario 5 requires combinations of highlighting techniques that are suitable for both Scenario 2 and 3.

Knowing how to read the detailed matrices is especially important in applications that pose restrictions on the choice of highlighting techniques. For example drawing shadows might not be possible in all scenarios and common techniques like underlining or bold typeface might already have been used in the editor's version of the text. In this case, the matrices can be inspected to find good "second best" solutions.

Furthermore, when multiple requirements must be satisfied, a trade-off might become necessary if no ideal candidate exists. Scenario 5 for instance needs the requirements of both Scenario 2 and 3 to be fulfilled. In practice, one of the two might be more important than the other.

distractor technique -->

| | fs | bo | bg | red | bold | sha | und | spa | it |
|---|---|---|---|---|---|---|---|---|---|
| font size | | 0.75 | 0.78 | 0.82 | 0.49 | 0.76 | 0.64 | 0.45 | 0.53 |
| border | 0.66 | | 0.79 | 0.79 | 0.77 | 0.76 | 0.50 | 0.59 | 0.53 |
| background | 0.69 | 0.67 | | 0.66 | 0.73 | 0.68 | 0.62 | 0.56 | 0.52 |
| red | 0.65 | 0.69 | 0.78 | | 0.65 | 0.63 | 0.58 | 0.55 | 0.51 |
| bold | 0.44 | 0.64 | 0.74 | 0.77 | | 0.64 | 0.61 | 0.57 | 0.52 |
| shadow | 0.59 | 0.64 | 0.70 | 0.70 | 0.63 | | 0.43 | 0.57 | 0.41 |
| underlined | 0.47 | 0.46 | 0.60 | 0.63 | 0.54 | 0.52 | | 0.42 | 0.41 |
| spacing | 0.35 | 0.38 | 0.52 | 0.53 | 0.42 | 0.45 | 0.38 | | 0.28 |
| italic | 0.20 | 0.23 | 0.29 | 0.24 | 0.22 | 0.21 | 0.18 | 0.15 | |

Fig. 8: Absolute performance values of Study 2 (referenced as Matrix M2).

| | fs | bo | bg | red | bold | sha | und | spa | it |
|---|---|---|---|---|---|---|---|---|---|
| font size | | 0.74 | 0.78 | 0.79 | 0.65 | 0.75 | 0.73 | 0.55 | 0.50 |
| border | 0.74 | | 0.69 | 0.80 | 0.73 | 0.63 | 0.47 | 0.74 | 0.48 |
| background | 0.78 | 0.69 | | 0.58 | 0.72 | 0.57 | 0.55 | 0.63 | 0.47 |
| red | 0.79 | 0.80 | 0.58 | | 0.60 | 0.55 | 0.57 | 0.58 | 0.45 |
| bold | 0.65 | 0.73 | 0.72 | 0.60 | | 0.58 | 0.55 | 0.55 | 0.45 |
| shadow | 0.75 | 0.63 | 0.57 | 0.55 | 0.58 | | 0.40 | 0.50 | 0.35 |
| underlined | 0.73 | 0.47 | 0.55 | 0.57 | 0.55 | 0.40 | | 0.64 | 0.29 |
| spacing | 0.55 | 0.74 | 0.63 | 0.58 | 0.55 | 0.50 | 0.64 | | 0.32 |
| italic | 0.50 | 0.48 | 0.47 | 0.45 | 0.45 | 0.35 | 0.29 | 0.32 | |

Fig. 9: Absolute performance values of Study 3 (referenced as Matrix M3).

| | bo | bg | red | bold | sha | und | spa | it |
|---|---|---|---|---|---|---|---|---|
| font size | 0.08 | 0.09 | 0.17 | 0.05 | 0.17 | 0.17 | 0.09 | 0.34 |
| border | | 0.12 | 0.10 | 0.13 | 0.12 | 0.04 | 0.21 | 0.29 |
| background | | | 0.12 | 0.01 | 0.02 | 0.02 | 0.04 | 0.23 |
| red | | | | 0.11 | 0.07 | 0.04 | 0.02 | 0.27 |
| bold | | | | | 0.02 | 0.07 | 0.15 | 0.30 |
| shadow | | | | | | 0.09 | 0.12 | 0.20 |
| underlined | | | | | | | 0.04 | 0.24 |
| spacing | | | | | | | | 0.12 |

Fig. 10: Delta of the matrix M2 (absolute values for Study 2) and its transpose.

### 8.3 Example application scenarios

Our first example stems from language analysis. Part-of-speech (POS) tagging allows automatic identification of the word class to which a term belongs. In our scenario, we focus on adjectives and adverbs that were identified with a POS tagger that is based on the Penn treebank annotation. In addition to those advanced language analysis techniques, simple heuristics also exist, such as identifying adjectives and adverbs by typical word endings such as "-able", "-ly", and "-ive." Our task is now to highlight all adjectives and adverbs as defined by the POS tagger and all words with typical adjective endings. Our goal is to gain a visual impression of how many adjectives / adverbs we would miss if we used just the simple heuristic. Additionally, we want to find examples of false positives, i.e., words with an adjective / adverb ending which are not adjectives / adverbs.

Figure 1 shows an excerpt from "Alice in Wonderland" in which the adjectives / adverbs that were identified with a POS tagger are highlighted in bold typeface, and at the same time the background of the words with adjective endings is colored in yellow. We can see in

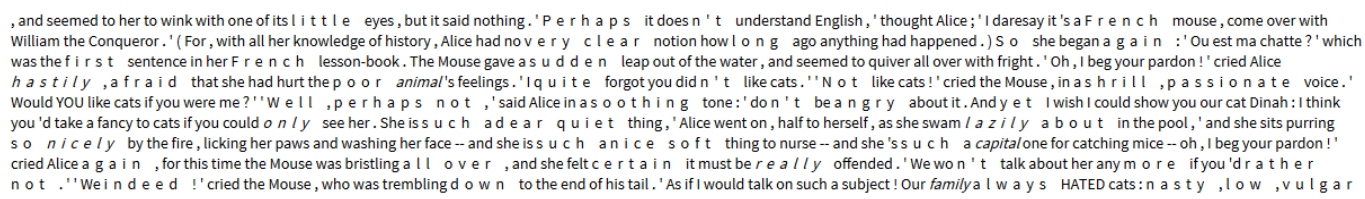


Fig. 11: Example of combining techniques letter spacing and italics – according to our analysis this is not an effective combination for highlighting two equally important text features.

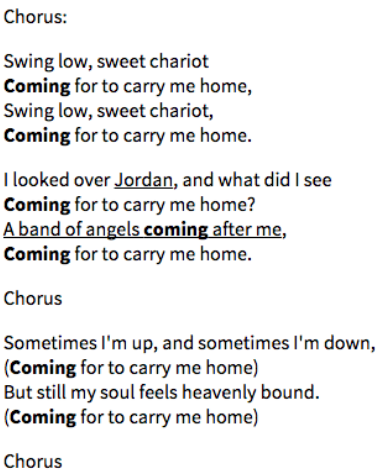


Fig. 12: Example of a dominant technique (bold) to highlight search results for "Coming" and a more recessive technique (underlined) for singing emphasis. Lyrics for "Swing Low, Sweet Chariot".

the text that there are three examples for words that end in an adjective / adverb ending, but that belong to other word classes (see words with yellow background color but not highlighted in bold typeface). The three terms are "table", "capital", and "five".

Figure 11 shows an alternative visualization of the same data and text. This time, a bad combination of highlighting techniques has been deliberately used for comparison (letter spacing for POS tagged words and italics for adjective / adverb endings). Again, three words with adjective / adverb endings are in the text, but with the bad choice of highlighting techniques, it is now much more difficult to find them.

As an example for dominant vs. informative text feature, we imagine an application that highlights text search results in song lyrics for an e-book device. These lyrics can include some singing-intended typefaces for several text segment cases. For instance, a song interpreter underlines a text passage which she thinks must be emphasized. This introduces a design constraint. When searching for a specific keyword, a second highlighting technique has to be added. We do not want to remove the first highlight, but we would like for the second search highlight to be dominant as this is the active task. Using font size together with underlined text would be a good combination, because font size is more dominant than underlined text (see Sections 8.1 and 8.2). However, if varying the font size within a text is not possible in an e-book reader, we have to search for an alternative. We decided to use bold typeface, which has also been determined as significantly more dominant than underline, as a second technique to pair up with underline. Figure 12 gives an example for an excerpt from a famous spiritual.

### 8.4 Discussion

We assume that the results can also be applied to find combinations of more than two techniques. For example, when searching for a combination of three techniques A, B, and C, one might consider finding good performance for all combinations AB, BC, and AC.

In a controlled study, inevitably choices between several study design alternatives have to be made that influence the results of the study. While applying the results to applications, readers need to keep the following restrictions and limitations in mind: First, in our study, we fixed some experimental settings, such as font type, font size, and interline spacing. We expect that for different settings, the highlighting techniques may perform differently. Furthermore, our study was conducted on the web with participants recruited from Amazon Mechanical Turk. Since our results reflect such environmental influences, future researchers need to keep this in mind while using other settings. Thus, testing our results with different settings will be a promising future work. Second, we designed our task, finding and clicking on a target word with highlights, to test our hypotheses. In practice, users may need to also read text context around highlights and in general, pursue high-level analysis. It would be an interesting experiment to test for effects of the highlighting techniques regarding cognition and analysis processes. Thus, care must be taken while following the results and guidelines, especially for other types of text analysis tasks. Third, there are numerous techniques that are not tested, for instance, different color combinations. Testing these combinations will be another interesting future direction.

## 9 Conclusions

We have empirically investigated the effective use of highlighting techniques for visualization applications for text data. Based on a literature analysis and survey among text analysis researchers, we have identified a set of candidate text highlighting typesettings which informed our crowdsourced user study. Our results provide design guidelines for the effectiveness of nine web-friendly text highlighting techniques in multi-annotation cases. The resulting matrices from evaluation studies as well as application scenarios will help information visualization application designers examine the effects of those techniques easily.

The study also identifies future work in visualization applications for text. The studied typesetting options can highlight individual terms within a text. However, it would also be interesting to study a combination of typesetting highlighting with overlay visualizations, e.g., to visualize relations, other boosting techniques, or even glyph visualizations embedded within text like Gestaltlines [4]. Last but not least, different colors for font and background could be tested.

The study results, the test system and the used source code are available at `http://textanno.hs8.de`.

## Acknowledgments

The authors thank James Tompkin and Sebastian Mittelstaedt. This work is supported by DARPA grant FA8750-12-C-0300 and the EU project Visual Analytics for Sense-making in Criminal Intelligence Analysis (VALCRI) FP7-SEC-2013-608142.

## References

[1] A. Abbasi and H. Chen. Categorization and Analysis of Text in Computer Mediated Communication Archives Using Visualization. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 11–18. ACM, 2007.

[2] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich. Seeing Beyond Reading: A Survey on Visual Text Analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):476–492, 2012.

[3] O. Amir, D. G. Rand, and Y. K. Gal. Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE*, 7(2):e31461, 2012.

[4] U. Brandes, B. Nick, B. Rockstroh, and A. Steffen. Gestaltlines. *Computer Graphics Forum*, 32(3):171–180, 2013.

[5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[6] W. S. Cleveland and R. McGill. Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, 229(4716):828–833, 1985.

[7] M. Correll and M. Gleicher. 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*, 2012.

[8] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, 9(2):e1002854, 2013.

[9] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pages 213–222, 2007.

[10] Egas by BMD Software Ltd. https://demo.bmd-software.com/egas/index.html; last accessed June 2015.

[11] A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino. Keep It Simple: Reward and Task Design in Crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 14:1–14:4. ACM, 2013.

[12] C. Gorg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, 2013.

[13] C. G. Healey and J. Enns. Attention and Visual Memory in Visualization and Computer Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012.

[14] C. G. Healey and J. T. Enns. Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 5(2):145–167, 1999.

[15] 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*, pages 203–212. ACM, 2010.

[16] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.

[17] D. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *IEEE Conference on Visual Analytics Science and Technology*, pages 115–122, 2007.

[18] S.-H. Kim, H. Yun, and J. S. Yi. How to Filter out Random Clickers in a Crowdsourcing-based Study? In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, pages 15:1–15:7. ACM, 2012.

[19] S. Koch, M. John, M. Worner, A. Muller, and T. Ertl. VarifocalReader - In-Depth Visual Analysis of Large Text Documents. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1723–1732, 2014.

[20] R. Kosara and C. Ziemkiewicz. Do Mechanical Turks Dream of Square Pie Charts? In *Proceedings of the 2010 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '10, pages 63–70, New York, NY, USA, 2010. ACM.

[21] K. Kucher and A. Kerren. Text Visualization Browser: A Visual Survey of Text Visualization Techniques, Poster Paper at IEEE VIS 2014 and webpage: http://textvis.lnu.se; last accessed June 2015.

[22] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, 2012.

[23] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. EventRiver: Visually Exploring Text Collections with Temporal References. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.

[24] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007.

[25] G. Miler. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2), 1956.

[26] QDA Miner by Provalis Research. http://provalisresearch.com/products/qualitative-data-analysis-software/; last accessed June 2015.

[27] M. Steinberger, M. Waldner, M. Streit, A. Lex, and D. Schmalstieg. Context-Preserving Visual Links. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2249–2258, 2011.

[28] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. BRAT: A Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[29] A. Stoffel, H. Strobelt, O. Deussen, and D. A. Keim. Document Thumbnails with Variable Text Scaling. *Computer Graphics Forum*, 31(3):1165–1173, 2012.

[30] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. Keim, and O. Deussen. Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1145–1152, 2009.

[31] S. van den Elzen and J. van Wijk. BaobabView: Interactive construction and analysis of decision trees. In *IEEE Conference on Visual Analytics Science and Technology*, pages 151–160, 2011.

[32] M. O. Ward, G. G. Grinstein, and D. A. Keim. *Interactive Data Visualization - Foundations, Techniques, and Applications*. A K Peters, 2010.

[33] C. Ware. *Visual Thinking for Design*. Morgan Kaufmann Publishers Inc., 2008.

[34] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *IEEE Symposium on Information Visualization*, pages 51–58, 1995.