

Compare Clouds: Visualizing Text Corpora to Compare Media Frames

Nicholas Diakopoulos¹, Dag Elgesem², Andrew Salway³, Amy Zhang⁴, Knut Hofland³
¹University of Maryland, ²University of Bergen, ³Uni Research, ⁴MIT CSAIL

ABSTRACT

Media frames represent distinct ways of communicating about issues that are reflected in choices of key words and phrases. In this paper we develop a visualization technique and visual analytic system that enables the study of media frames across text corpora. In particular our system allows scholars or other analysts to compare media frames in a visualization called the Compare Cloud, which explicitly maps word *prevalence* and *context* information between two corpora. We assess the error profile of the visualization layout and demonstrate the utility of the system by comparing the media discussion between mainstream media and blogs on the topic of surveillance. We report salient observations that the visualization made possible and discuss future challenges related to scalability and effective filtering to support visual frame analysis.

Author Keywords

Media framing, text visualization, surveillance

INTRODUCTION

On June 5th, 2013 a trove of classified U.S. government documents began to become public and change the way the world saw the surveillance state both in the U.S. and internationally. Edward Snowden had acquired the documents while working for Booz Allen Hamilton at an NSA center and decided to disclose them to journalists who then published a series of investigative news stories informed by the documents [20]. Is Edward Snowden a “dissident” or “traitor”, a “whistleblower” or “leaker”? The words that are chosen to label actors in the news affect how they are perceived, the connotations that are evoked, and ultimately the public opinion that may arise [4,29]. In this case the Associated Press (AP) decided that Snowden was more of a “leaker” than a “whistleblower”¹. And whereas The Guardian framed Snowden more as a whistleblower, The Washington Post preferred the term “leaker” [36].

This vignette is a salient example of *media framing*, the result of linguistic choices that affect how an issue is presented in the media, including the various viewpoints and notions that are communicated with respect to that issue. Media frames manifest linguistically through the presence or absence of particular keywords, stock phrases, patterns of word use, or thematic clusters [17] and are thought to be influenced by social norms, organizational pressures, journalistic routines, and ideological orientations

[29]. Identifying frames, persuading audiences to notice and reflect on them [2], and comparing frames across different news outlets or facets of media sources [13,21] are all important tasks to the study of media frames.

In this work we focus on *enabling the comparison of media frames* by developing a visual analytic system incorporating what we call a “Compare Cloud”. We argue that existing text visualizations are not adequate for the task of text corpora comparison, which is essential for the study of media frames. Thus the Compare Cloud is designed to explicitly encode comparisons in a layout that enables exploratory analysis and comparison of two corpora. We then use the system to compare media frames in mainstream news media and online blogs related to the Snowden leaks. Our contributions include the design of the Compare Clouds layouts as informed by the task demands of media frame comparison, error profiling of the layouts, and a set of analytic observations exposed by the tool that highlight its utility. We discuss the ongoing opportunities for further research based on our results and highlight the challenges that arise when designing scalable visual analytics for frame comparison.

RELATED WORK

Text visualization is now used in a wide variety of different domains and for different purposes, from the communicative [30], to the exploratory analysis of topic models [5], to single document visualizations [23]. In their early work Wise et al. [35] identified three levels of textual data that benefit from visualization: *word frequencies or rates*, *word contexts*, and *semantics*. Here we focus on related work for each of these in turn, with a specific emphasis on techniques potentially useful for comparing framing effects in corpora.

Word Frequencies

One of the most prevalent visualizations that addresses the communication or analysis of word frequencies is the Word Cloud [1,11,31], which encodes information in the font size, font weight, color, or intensity of the words. Another, variant of the Word Cloud is the Wordle, which is a more compact and space efficient layout of terms [30]. WordBridge is a further visualization in the Word Cloud family which enhances the idea of a Wordle by creating a deterministic layout that produces reliable and repeatable outputs given the same inputs [24].

The Parallel Tag Clouds (PTCs) technique [8] combines layout techniques from parallel coordinates with word

¹ <http://blog.ap.org/2013/06/10/whistle-blower-or-leaker/>

clouds consisting of a single column of terms, one column for each facet of a text collection being compared. Clark created word association spectrums for comparing keyword frequency and associations between two corpora [6]. Cui et al. [10] have adapted word cloud visualizations for comparing distinct points in time in a temporal document corpus. More recently the Vennclouds technique was presented for comparing two corpora using juxtaposed word clouds [9]. However, none of these techniques enable adequate comparison of word *contexts*, i.e. the words that surround instances of a keyword in a corpus. Unlike these other techniques, the Compare Cloud visualization is expressly designed to *show word contexts and align with the task of frame comparison*.

Word Contexts

The visualization of word context has been enabled through a number of techniques. Context generally refers to a fixed range on either side of a keyword, or to the sentences or paragraphs containing a keyword. A standard corpus analysis technique for showing word contexts is KWIC (Key Word in Context) and straightforwardly lists out concordance lines to show a certain number of terms to the left and right of all instances for a given search term [11]. Our system incorporates a KWIC display for seeing and comparing specific sentences between corpora, but our Compare Cloud provides an overview comparison of aggregate context that a user can use to navigate to specific examples where they want to examine more detail.

Word Trees attempt to provide a more manageable view of concordance data by using suffix trees to allow for the interactive exploration of visual trees built from word sequences [34]. Word Trees are a powerful way to understand context around a term, however they are limited in that they only represent context in terms of strict adjacency of terms. Other more subtle framing effects might be elucidated from a less strict treatment of context. The Double Tree extends the Word Tree and shows both the prefix and suffix tree context around a word, thus supporting some form of comparison between the two and providing a more robust overview of word use context [12]. Collocation networks have also been employed to show word contexts by creating network visualizations of word collocations [26]. The FeatureLens system [16] allows for the visual exploration of frequent textual patterns and is augmented by analytics algorithms which help to surface those patterns.

The Lingscope system seeks to surface words with different usage patterns and contexts in different parts of a discourse [13] by using simple small multiples visualizations of word usage frequency in the context of an anchor term. Our work builds on Lingscope's approach towards the study of word frequency in context, however we provide a wider overview of corpora as well as a spatial layout technique to enhance comparison.

Semantics

The final level of text visualization that Wise et al. mention is that of semantics, which has also garnered some attention in the literature. The DocuBurst system utilizes semantic information from WordNet to create a hierarchical representation of word relationships [7]. Other work in the digital humanities has developed techniques that visualize the Part of Speech (POS) relationships in a text [32]. Phrase Nets are a technique that creates a graph out of word relations like possessives, conjunctions, prepositions or other orthographic connections between words [22].

The WordSeer platform is a full-featured exploratory text analysis system allowing for slicing corpora in different ways, exploring related words, annotating, and close reading of the original content [27]. It also extracts some semantic information such as parts of speech and named entities and allows scholars to search for word relations. However the creators of WordSeer still identify the need for further work in developing effective visualizations to *compare* corpora of texts, which is something that is at the core of this present work.

More recently the Reflext system was developed to directly address the use-case of visualization to support reflection on media frames [2]. Reflext uses a modified version of the Wordle algorithm and a connected radial tree view to show selectional preferences of terms to key nouns in a corpus, thus indicating terms that exhibit certain grammatical relationships with the selected noun. While semantic information in general can be useful for certain linguistic analytic tasks and may be incorporated in our future work, here we instead chose to simplify our system and focus on visualizing word contexts, with direct access to original source texts that can be used to read text excerpts in order to understand the meaning and use of a term.

COMPARE CLOUDS

Here we describe the overall design goals and task, layout algorithm, interactions and filters, data processing, and implementation of the Compare Cloud. Then we present results that benchmark and characterize error in the layouts.

The Analytic Task of Frame Comparison

A key challenge to the development of effective visualizations is in aligning visual abstractions with real world tasks [5]. Previous work, including some of our own [14,15] and others [21,25] has elucidated some of the design requirements such as the need to drill into original source material to do deep reading of original text content, and the need to compare corpora across sources or other facets, which motivates this paper. Key to the task definition is an ability to *compare the contextual use of key words* between corpora to better understand how words are used in relation to an anchor key term.

Yet the majority of the aforementioned works, with some exceptions (e.g. [8,13]) visualize text frequency, context, or semantics but do so in ways that make it difficult to

compare these between corpora. Gleicher et al. [19] provide a theoretical framework for thinking about visual comparison in information visualization including methods such as juxtaposition, superposition, and explicit encoding. Of the work cited above that directly addresses comparison it primarily focuses on the use of juxtaposition to do so [7,10,12,27,31], or calls for future work on enabling visual comparison [22,27,32].

We argue that the use of juxtaposition does not support the task of frame comparison well. Text visualizations oriented towards comparative tasks might make better use of explicit encodings mapped to salient visual features such as spatial layout, intensity, or color [33] in order to improve their ability to pre-attentively signal interesting contrasts. Compare Clouds represents an effort towards developing a text visualization with explicit encoding that aligns with the text comparison task so that communication scholars can explore differences in word contexts and framing.

Visualization Design and Interface

The design goal of Compare Clouds was to provide an overview and detail interface for a scholar to be able to compare word frequencies (rates) and word contexts between two corpora. We sought to create a visualization

that would make it easier to study media frames and their comparisons by using explicit encodings of usage rates of words. We designed it as a tool to aid exploratory analysis and hypothesis generation rather than with any overt efficiency goal. The focus was on visualizing aggregate statistics to find overall patterns, but with access to readable text fragments so that users could go deeper into the texts that they were analyzing.

As shown in Figure 1 the Compare Cloud visualization encodes information in multiple visual channels. Words to the left of the layout are more often found in one corpus (Mainstream Media or MSM in this example) and words on the right of the layout are more often found in a second corpus (blogs in this case), whereas in the middle are words that are used at similar rates in both corpora. Words are alphabetically ordered along the vertical axis. There is a dual encoding of the ratio of use between corpora in the color channel assigned to each word, (red maps to more use by MSM, blue maps to more use by blogs). The size of terms in the layout reflects the overall prevalence of the term as used across both corpora.

When initially loaded the cloud will map the overall rate of use of terms between the two corpora, giving the user an

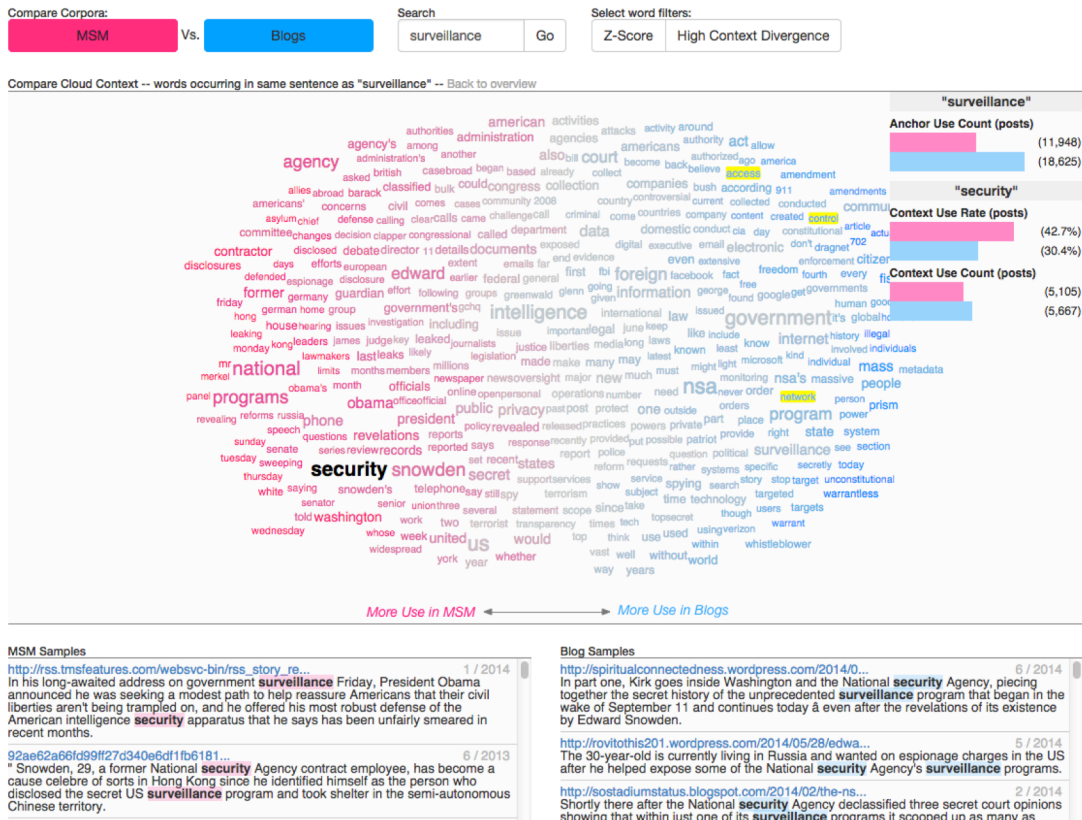


Figure 1. A Compare Cloud showing words used in the same sentence context as “surveillance” between MSM and Blogs. Details are shown for the selected word “security” which is used in sentences with “security” more prevalently by MSM than blogs. Highlighted in yellow are words found in the same context as “security” including “control” and “access” that are clearly more in use by the blogs than MSM.

overview and starting point for browsing terms and comparing the prevalence of use in either corpus. If a specific term is searched for, the layout will show a comparison of the rate of use of *words in context* (i.e. words used in the same sentence) of that searched term. Thus we provide both (1) an initial overview, and (2) a word context oriented view for examining specific frames around an anchor search term.

As with any visual analytic interface, interaction is essential to the use of Compare Clouds. We use the progressive disclosure design pattern [5] to allow for transitioning from an overview down to the raw textual data itself. Hovering over a term in the layout shows a small stacked bar chart indicating the ratio of use of the term in each corpus; an information scent for the user. Clicking a term will show the details panel as seen in the upper right corner of Figure 1. For the context view shown it indicates the overall usage of the anchor term, “surveillance” in this case, as well as the rate of use of the context term that was selected, “security” in this case.

Clicking a term in the overview layout or in a context layout will bold the selected term and cause other words to be highlighted with a yellow background if they are often found in similar contexts according to the Wordnik API². This highlighting function can be used to investigate words to see variations on meaning that may be used in either corpus more or less.

Clicking a term in a context layout will also show up to 100 randomly selected sentences where the anchor and context terms occurred together in each of the two corpora. These are shown to the lower left and lower right of the layout in Figure 1 as a KWIC display and indicate the sentence, highlighting the terms of interest, with a link back to the original article, and a date stamp for temporal context.

Other terms can be searched by double-clicking a term in any layout, or by typing directly into the search box at the top, which has type-ahead for all of the terms pre-computed in the system. We also provide filters in the UI based on a z-score test of the ratio of use of a term, and a “high context divergence” score filter described below. The filters are an important component of Compare Clouds as they enable the visualization of smaller and more manageable subsets of words. Enabling a filter triggers a redraw of the display that hides any filtered words and lays out only unfiltered words.

Layout Algorithm

The layout of the Compare Cloud is built on the force-directed graph layout implementation in D3 [3] with additional soft constraints and an update step. The y-position of each term in the layout is initialized based on a simple partitioning of the vertical space in the layout according to alphabetical ordering which allows for easy

scanability of the layouts. The x-position of each term is initialized based on the ratio of the use of the term in corpus one (C1) and corpus two (C2). Ratios of less than 1 are linearly mapped to the left-most 30% of the space, and ratios greater than 1 are linearly mapped to the right-most 30% of space. Terms are connected via a link force to their ideal layout position, so as to prevent them from straying too far from that ideal position. Then the gravity force (initialized to 0.015) of the layout serves to pull all terms toward the center of the layout and create a rounded effect. We found that this initialization strategy served to better partition the terms that should clearly be more toward the left or right and avoid most muddying of terms in the middle of the layout.

On each tick of the layout we do intersection tests between all terms in the layout. Two terms whose bounding boxes intersect are repulsed along the vector connecting their center points. There is an additional charge force in the layout (set to -1.3) which also serves to repulse nodes / terms. When the cooling parameter, alpha, for the force directed layout hits a threshold we again assess the number of term overlaps. If there are few overlaps we stop the layout algorithm and finish, otherwise we run a refinement step to nudge terms away from the central vertical axis so as to provide more horizontal spread of terms and avoid word overlaps. Alpha is then reset so that the force directed layout algorithm continues, and the gravity is reduced slightly (to 0.011).

Data Processing

Text data is processed through a fairly standard NLP pipeline. Texts are first cleaned of any spurious HTML markup then segmented into sentences, the main contextual unit of analysis. Text is lower-cased so that words at the head of a sentence are not aggregated separately. However, text is not stemmed as we wanted to retain different word morphologies for frame analysis. Then frequencies are generated for single words (unigrams): specifically, the number of unique sentences and the number of unique posts/articles in which a word appears. These values are saved only for unigrams that surpass a set threshold of usage (5%) in either of the two corpora. This reduces the presence of terms that only occur rarely in both corpora and allows us to focus on the analysis of more prevalent terms. For the words that remain, context data is generated: for each word, counts are made for all other words that occur in its sentences. The context words are thresholded at 1% for use in a particular context for either of the two corpora. For each anchor term and context term combination 100 exemplar sentences are randomly sampled from each corpus, which allows both for a fast response to the client, as well as limiting the space needed to store all of the pre-compute examples.

Context divergence scores are used to enable a filter that shows a subset of words in the Compare Cloud that have contexts that are particularly different. For each anchor

² <http://developer.wordnik.com/>

term in our dataset, we look at the context terms that have been computed in each of the two corpora. The score is calculated as the average L1 norm distance between the usage rate of each of the terms in the context vectors. The score helps measure the gross difference between contexts and identify terms for which the contexts are more varied between the two corpora.

Visualization Layout Critique and Evaluation

In designing Compare Clouds we struggled to push their scalability. Given a standard laptop display (e.g. 1150 x 550 pixels in our tests) the Compare Cloud can scale to visualize approximately 500 terms. However, even with the initial thresholds we described above, we still had over 1600 terms in the overview.

In order to be compact and space filling while avoiding overlapping terms, each term cannot necessarily inhabit its ideal position in the layout. Term positions are thus adjusted in order to accommodate other term positions. Because of this the layout distorts the actual data in an effort to produce a more legible and readable layout that still communicates the overall “gist” of relationships between terms and still supports comparison in the layout. Next we quantify the degree of distortion and evaluate how visual comparisons may be affected by the layout.

Aggregate Ordering Error

To evaluate the frequency and magnitude of layout errors we want to know: (1) when considering every pairwise comparison between terms in the layout, how many ordering errors there are, and (2) for errors that do exist, how severe are they in terms of relative x-positions of the terms involved. To assess these measures we ran an evaluation that tabulated errors for layouts with a size of 300 randomly sampled terms in a layout area of 1150 x 550 pixels. We computed errors averaged over 100 runs of the layout. Amongst words that had an ordering error (i.e. they were left of a term they should have been right of or vice versa), the average horizontal error with respect to other mis-ordered terms was 39 pixels. Thus mis-ordering of terms was not too severe on average. Across all 100 runs of the layout, the average number of ordering errors was 3,393. For a 300 term layout there are $\binom{300}{2}$ possible ordering errors, thus 3,393 represents an error rate of 7.57%. The highest pairwise error rate was 8.2% and the lowest was 6.8%. We re-ran the process with layouts of 200 terms and found similar error rates.

Ordering Error on an Individual Layout

To give a more precise evaluation of the errors that might occur in a layout we tabulated horizontal pixel error for one of the more important layouts in our application: the overview. The evaluation is done using a 359 term layout (the number of terms in the overview of the corpora which is described below), and the default size of 1150 x 550 pixels. In this particular layout we find a pairwise ordering error rate of 7.1%, and an average horizontal error rate

amongst mis-ordered terms of 40 pixels. A majority of mis-ordered terms (85.2%) have an error in the range of 20-50 pixels, with only seven mis-ordered terms (1.9%) having an error greater than 70 pixels. Thus even where there are mis-ordered terms, the mis-orderings are largely not too severe in terms of x-value difference in pixels.

Summary of Layout Evaluation

The results of these evaluations have indicated that even for mis-ordered terms the errors are relatively small (in the 20-50 pixel range), suggesting that the overall gist of the distribution is intact. If a term in the layout were more than 50 pixels to the left or right of another term the user could reasonably expect that spatial ordering to reflect a valid comparison of the underlying data. For cases where the error is larger and more problematic we have a fallback, which is the dual-encoding of the x axis ideal position as a color for the term itself. Thus for terms that do have higher error rates, color can signal if the term is out of place, since its color will be less similar to those of the terms it is adjacent to. The color mapping thus becomes an indicator of spatial mapping error.

ANALYSIS OF SURVEILLANCE CORPORA

Here we present the results of an analysis of the discourse around surveillance in both the MSM and in blogs since the Snowden files were leaked in June 2013. We first describe the corpora that were used to produce the Compare Cloud. Then we describe an exploratory analysis aimed at understanding the different attitudes toward Snowden and different facets of the surveillance issue as discussed in the MSM and blogs.

Corpora

Two corpora were collected, one to sample the media around the issue written across three blog platforms (26,364 posts), and another from fifteen mainstream news sources (19,303 articles). Both corpora were limited to the time period after the Snowden leaks, from June 2013 until the end of July 2014. The news corpus is comprised of mostly U.S. publications with the exception of The Guardian. The blog posts were all English language, but it was not possible to restrict them by country.

Blogs

A corpus of English-language blog posts related to the topic of surveillance was gathered by querying three search engine APIs daily from April 11th 2014 to August 15th, 2014. Repeated querying captured both new posts (after the start date) and older posts that later appeared in the results sets (presumably after re-indexing of the search engines). The search engines used were Google, Bing and Yahoo. Twenty-one query terms were chosen based on the coverage of the issue in MSM and blogs, in particular with respect to the EU directive on data retention, and with respect to Snowden. Domain expertise and inspection of frequent n-grams in some relevant blog posts also informed the selection of query terms. Any redundancies in posts returned by multiple different queries were removed. The

query terms were: “data retention”, “edward snowden”, “electronic surveillance”, “fisa court”, “government surveillance”, “intelligence agencies”, “intelligence surveillance”, “internet surveillance”, “mass data collection program”, “mass surveillance”, “nsa program”, “nsa spying”, “nsa surveillance”, “prism program”, “retention directive”, “retention policy”, “security agency”, “surveillance act”, “surveillance program”, “warrantless surveillance”, and “warrantless wiretapping”.

In order to make the subsequent processing of blog posts feasible, the queries were restricted to three blog platforms – WordPress, Blogspot and Typepad. In previous web crawling we had ascertained that the vast majority of relevant posts would appear on one of these platforms. Each day a query was made to each search engine API for every combination of twenty-one search terms and three blog platforms. The collected posts were processed with JusText³ to extract the main text content, which was stored along with the date (month and year) that was extracted from the URL.

Mainstream Media

Data was provided by NewsCred which aggregates full text news content from thousands of news sources, and makes it available via an API. We chose fourteen sources from NewsCred which ranged from large wires (AP, Reuters), to broadcast (CNN), national newspapers (USA Today, Washington Post, News York Times), regional or local newspapers (LA Times, Chicago Tribune, NewsDay, Houston Chronicle, Philly Inquirer, Minneapolis Star Tribune, Honolulu Star-Advertiser) to online only (Huffington Post); we also included articles collected directly from the Guardian API. Articles were collected using the same set of query terms as used for blogs and spanned the same time period.

Exploratory Analysis

In examining the overview Compare Cloud with the high context divergence filter applied, one of the aspects that was apparent is the focus of the MSM on terms relating to the international reactions to the story. Terms like “russia”, “london”, “international”, “britain”, “france”, “berlin”, “iran”, “syria” are used at a higher rate in the MSM corpus. Looking at the context of “European” is illuminating. It’s clear for instance that blogs are talking much more about the European data “retention” “directive” as well as the “convention” on human rights in Europe. The MSM instead frames Europe as “allies” e.g. “European allies” and reports on “talks” with Europe moreso than blogs.

Also dominant on the overview visualization is the term “surveillance” which makes sense as that was meant to be the topic of the two corpora. In searching for “surveillance” to inspect the context (See Figure 1) we find several interesting features. Blogs employ critical language in the

context of surveillance, using terms like “warrantless”, “unconstitutional”, “illegal”, and “dragnet”. Illustrative of the story presented at the start of this paper, we found that blogs use the term “whistleblower” in the context of “surveillance” more often. The term “leaker” dominated the context for MSM. The Compare Cloud thus clearly articulated the style guidance of the AP indicating the preference of “leaker” by the MSM. Another observation about the context of “surveillance” is that the MSM talks much more about a “contractor”. They frame Snowden as someone in a temporary position, not as an employee, which would presumably have different implications for the company who had hired him. This speaks to how Snowden’s credibility was framed by the MSM.

To follow-up on the idea of credibility, we then searched for “expert” and found that the blogs do frame Snowden more as an expert (8.8% of posts) versus the 6.5% of MSM articles, but by reading individual text snippets closely found that the term “expert” is actually used in different ways. In the blogs Snowden is often called an “expert” e.g. “what the young computer *expert* has made public“, or “The revelations brought to us by IT *expert* Edward Snowden”. But in MSM “snowden” occurs in the context of “expert” mostly because the article cites an expert saying something about Snowden. There are examples in the snippets where Snowden is called an expert, e.g. that he was “through contract or direct hire as a technical *expert* for the United States National Security Agency”. But this is used as a description of his background and not to give credibility to his statements. He is not framed as an expert source like the other people who are called “experts”. For example: “One legal *expert* was puzzled as to why *Snowden* fled to Hong Kong, because it has an extradition treaty with the United States while mainland China does not.” This seems to be an interesting difference in the way Snowden’s expertise is invoked in the assessment of the credibility of his claims. Our tool’s support for doing deep and detailed reading of excerpts was essential to uncovering these observations.

The utility of highlighting words that are used in similar contexts was also apparent in our analysis of the context comparison of “expert” (shown in Figure 2). The figure



Figure 2. Compare Cloud context for the anchor term “expert”, with “privacy” selected and words related to “privacy” via Wordnik highlighted in yellow.

³ <https://code.google.com/p/justext/>

CONCLUSIONS

In this work we have described a visual analytic system that is designed to enable the exploratory analysis and comparison of media frames. Unlike previous text visualizations the Compare Cloud explicitly encodes comparative information while maintaining readability of and access to underlying raw texts, and a set of filters and highlights that can draw attention to interesting terms. We characterize the error profile of the layout and evaluate its ability to faithfully represent comparisons through the spatial ordering of terms. Through an exploratory case study we found that the layout and features of the system helped to lead to insights about media framing differences between MSM and blogs discussing the surveillance issue.

REFERENCES

1. Bateman, S., Gutwin, C., and Nacenta, M. Seeing things in the clouds: the effect of visual features on tag cloud selections. *Proceedings of conference on Hypertext and hypermedia*, (2008), 193–202.
2. Baumer, E.P.S., Cipriani, C., Davis, M., et al. Broadening Exposure, Questioning Opinions, and Reading Patterns with Reflex: a Computational Support for Frame Reflection. *Journal of Info. Technology & Politics* 11, 1 (2014), 45–63.
3. Bostock, M., Ogievetsky, V., and Heer, J. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics*, (2011).
4. Chong, D. and Druckman, J.N. Framing Theory. *Annual Review of Political Science* 10, 2007, 103–126.
5. Chuang, J., Ramage, D., Manning, C., and Heer, J. Interpretation and trust: Designing model-driven visualizations for text analysis. *Proc. CHI*, (2012), 443–452.
6. Clark, J. News Spectrum. 2008. <http://www.neoformix.com/Projects/NewsSpectrum/NewsSpectrum.html>.
7. Collins, C., Carpendale, S., and Penn, G. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum* 28, (2009), 1039–1046.
8. Collins, C., Viégas, F.B., and Wattenberg, M. Parallel tag clouds to explore and analyze faceted text corpora. *Proc. VAST*, (2009), 91–98.
9. Coppersmith, G. and Kelly, E. Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis. *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces*, (2014).
10. Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M., and Qu, H. Context-preserving, dynamic word cloud visualization. *IEEE PacificVis*, (2010), 121–128.
11. Culy, C. and Lyding, V. Corpus Clouds - Facilitating Text Analysis by Means of Visualizations. *Proceedings of the 4th conference on Human language technology*, (2009).
12. Culy, C. and Lyding, V. Double tree: An advanced KWIC visualization for expert users. *Proceedings of the International Conference on Information Visualisation*, (2010), 98–103.
13. Diakopoulos, N., Zhang, A., Elgesem, D., and Salway, A. Identifying and Analyzing Moral Evaluation Frames in Climate Change Blog Discourse. *Proc. ICWSM*, (2014).
14. Diakopoulos, N., Zhang, A., and Salway, A. Visual Analytics of Media Frames in Online News and Blogs. *IEEE InfoVis Workshop on Text Visualization*, (2013).
15. Diakopoulos, N. Visual Analytics of Blog Corpora for Communication Scholars. *IEEE InfoVis workshop on Interactive Visual Text Analytics*, (2012).
16. Don, A., Zheleva, E., Gregory, M., et al. Discovering interesting usage patterns in text collections: integrating text mining with visualization. *Proc. CIKM*, (2007), 213–221.
17. Entman, R.M. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, (1993), 51–58.
18. Fisher, D., Popov, I., Drucker, S.M., and Schraefel, M. Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. *Proc. CHI*, (2012).
19. Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C.D., and Roberts, J.C. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309.
20. Greenwald, G. *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State*. Metropolitan Books, 2014.
21. Grundmann, R. and Krishnamurthy, R. The Discourse of Climate Change: A Corpus-based Approach. *Critical Approaches to Discourse Analysis across Disciplines* 4, 2 (2010), 113–133.
22. Van Ham, F., Wattenberg, M., and Viégas, F.B. Mapping text with phrase nets. *IEEE Transactions on Visualization and Computer Graphics*, (2009), 1169–1176.
23. Hearst, M.A. TileBars: Visualization of Term Distribution Information in Full Text Information Access. *Proc. CHI*, (1995).
24. Kim, K., Ko, S., Elmqvist, N., and Ebert, D.S. WordBridge: Using Composite Tag Clouds in Node-Link Diagrams for Visualizing Content and Relations in Text Corpora. *Proc. HICSS*, (2011).
25. Liu, S., Wang, X., Chen, J., Zhu, J., and Guo, B. Topic Panorama: A Full picture of relevant topics. *Proc. VAST*, (2014).
26. Magnusson, C. and Hannu Vanharanta. Visualizing Sequences of Texts Using Collocational Networks. In *LNCS Machine Learning and Data Mining in Pattern Recognition*. 2003, 276–283.
27. Muralidharan, A. and Hearst, M.A. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing* 28, (2012), 283–295.
28. Oelke, D., Strobelt, H., Rohrdantz, C., Gurevych, I., and Deussen, O. Comparative Exploration of Document Collections: a Visual Analytics Approach. *Proc. EuroVis*, (2014).
29. Scheufele, D. Framing as a theory of media effects. *Journal of communication*, (1999), 103–122.
30. Viegas, F.B., Wattenberg, M., and Feinberg, J. Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137 – 1144.
31. Viégas, F.B., Wattenberg, M., Ham, F. van, Kriss, J., and McKeon, M. Many Eyes: A Site for Visualization at Internet Scale. *IEEE Trans. on Vis. and Comp. Graphics*, (2007), 1121–1128.
32. Vuillemot, R., Clement, T., Plaisant, C., and Kumar, A. What's being said near “Martha”? Exploring name entities in literary text collections. *Proc. VAST*, (2009), 107–114.
33. Ware, C. *Visual Thinking for Design*. Morgan Kaufmann, 2008.
34. Wattenberg, M. and Viégas, F. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008).
35. Wise, J.A., Thomas, J.J., Pennock, K., et al. Visualizing the non-visual: spatial analysis and interaction with information from text documents. *Proc. InfoVis*, (1995), 51–58.
36. Wu, A. “Whistleblower or Leaker? Examining the Portrayal and Characterization of Edward Snowden in US, UK and HK Posts. *Hong Kong Assoc. for Educ. Comm. and Tech. Conf.*, (2014).