Eurographics Conference on Visualization (EuroVis) 2016
K.-L. Ma, G. Santucci, and J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

TextDNA: Visualizing Word Usage with Configurable Colorfields

D. A. Szafir[1,2], D. Stuffer[2], Y. Sohail[2], and M. Gleicher[2]

[1]University of Colorado Boulder
[2]University of Wisconsin–Madison

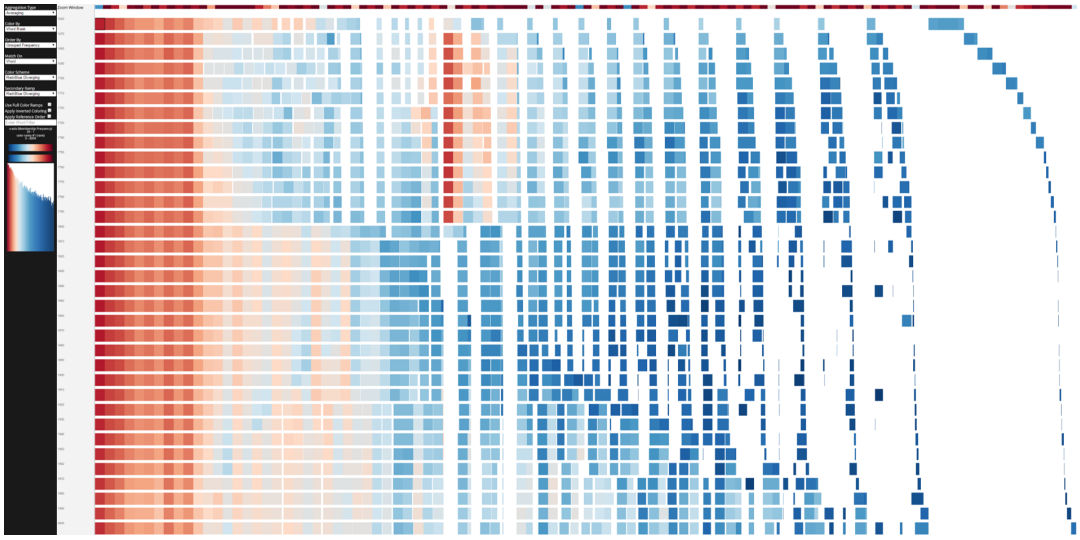


Figure 1: *TextDNA allows people to compare word usage patterns across large text corpora. Here, configurable colorfields provide an overview of the 5,000 most commonly used words per decade over the last 350 years (one decade per row). Ordering words according to the sets of decades in which they are common (left are common in all, right in one) and coloring by a word's commonality within each decade reveals temporal correspondences between decades (c.f. §6.1.2)*

Abstract

Patterns of words used in different text collections can characterize interesting properties of a corpus. However, these patterns are challenging to explore as they often involve complex relationships across many words and collections in a large space of words. In this paper, we propose a configurable colorfield *design to aid this exploration. Our approach uses a dense colorfield overview to present large amounts of data in ways that make patterns perceptible. It allows flexible configuration of both data mappings and aggregations to expose different kinds of patterns, and provides interactions to help connect detailed patterns to the corpus overview. TextDNA, our prototype implementation, leverages the GPU to provide interactivity in the web browser even on large corpora. We present five case studies showing how the tool supports inquiry in corpora ranging in size from single document to millions of books. Our work shows how to make a configurable colorfield approach practical for a range of analytic tasks.*

Categories and Subject Descriptors (according to ACM CCS):
H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical User Interfaces; I.7.m [Document and Text Processing]: Micellaneous—Text Analysis; J.5 [Computer Applications]: Arts and Humanities—Literature

1. Introduction

Visualizations of word usage, such as the Google N-Gram Viewer [MSA*11], allow users to explore how patterns in key words change over different documents or over time. These tools allow users to carefully track small collections of important terms. However, such tools are limited in that they do not afford exploration over large numbers of words, either to find words for closer examination or to identify patterns involving many words. Nor do they necessarily support localizing these patterns within individual documents.

Our goal is to enable people to compare word usage between *collections* of words. These collections may be chapters within a book, plays written by an author, or a set of books published in a given decade. For each word in a collection, we know how often it occurs, and, in some cases, where it occurs. The difficulty in exploring this data comes from scale and pattern complexity: there are often too many words to visualize simultaneously, and interesting patterns can be complex and difficult to define *a priori.* The range of potential questions asked of these overviews is diverse and often not known before patterns reveal themselves. Even at the smallest scales (e.g., chapters), addressing these questions effectively requires broad overviews due to the number and diversity of words within each collection. The main challenge for comparing these collections is, therefore, how to present such large and diverse collections in ways that allow a breadth of interesting patterns to emerge.

In this paper, we present TextDNA, a web-based visualization system that uses a *configurable colorfield* design combined with tailored interaction techniques to enable multiscale exploration of word usage data. Our design centers around a dense colorfield display that enables aggregate judgments at multiple scales. Users can interactively reconfigure how data is mapped to color and position within these colorfields to answer a variety of questions.

This configurable colorfield approach has been part of prior systems, notably the work of Keim [Kei96, OJS*11] and the Sequence Surveyor system for multiple sequence alignment comparison [ADG11]. In applying this approach to word usage, we addressed a number of critical challenges unsolved by prior systems. In particular, TextDNA (1) defines a new, flexible data model that supports a range of word usage analyses, (2) uses configurable colorfields to generate overviews that visualize multiple properties of a collection simultaneously, (3) supports examining specific words in context through a variety of interaction techniques, and (4) achieves interactive performance with large datasets in a web browser to ease accessibility, reproducibility of findings, and remote collaboration.

This paper explores how TextDNA extends the configurable colorfield approach to explore word usage patterns in text corpora of various sizes. We first outline a pair of task-driven data models to support frequency- and location-based analyses (§3). We then describe how flexible mappings can expose different kinds of patterns, for example, by reordering sequences based on set membership or relationship to a particular reference (§4.1). We show how interaction can highlight specific patterns in context (§4.2) and use a GPU implementation to provide interactive performance in the browser (§5). We conclude with five case studies generated by researchers in the humanities showing the utility of our approach at scales ranging from a single novel to millions of texts.

2. Related Work

Visualizations for comparing text collections often rely heavily on statistical processes such as topic modeling [AKV*14, CMH12, WLS*10] or dimensionality reduction [CWG11] to find specific commonalities between documents. However, these models do not allow users to directly compare word usage across documents.

Word usage visualizations often use word counts to characterize texts. For example, tag cloud techniques (see [VW08] for examples) map how frequently a word appears in a document collection to its font size. FeatureLens [DZG*07] visualizes the most frequently used words within a corpus at both an overview and detail level. Many word usage visualizations communicate how word counts change over time for a set of specific words [MSA*11, KBK*11, LRKC10]. Others show frequency in the context of different descriptive metadata, such as frequently co-occurring terms [WV08, JKM12] or contemporaneous events [WJS*15]. However, these techniques only support questions about predefined subsets of dozens to hundreds of words. Instead, we consider how visualization can support comparison across hundreds of thousands of words simultaneously.

Alternative word usage techniques focus on visualizing the frequency of all words within a single document. For example, DocuBursts [CCP09] use radial layouts to visualize frequency and semantic relationships between words in a document. TextArc [Pal02] encodes the frequency and average position of words within a document. While these techniques allow users to explore word usage broadly across a document, they do not readily facilitate comparison between multiple collections. Parallel tag clouds [CVW09] address this issue by explicitly encoding word frequency relations across different facets of a corpus, but compare only a subset of words within each facet. Alternatively, literature fingerprinting [KO*07] and pixel boosting [OJS*11] provide summary views of different word usage properties in individual documents using color. These techniques allow users to perform specific kinds of comparison, yet they do not scale to the needs of very large text corpora. They have limited configurability, and have limited support for grounding high-level comparisons in individual examples.

Recent systems, such as Serendip [AKV*14], ShakerVis [GCL*13] and JigSaw [GLK*13] show how overview+detail approaches can be applied to connect statistical text visualizations to word-level details. These approaches enable users to engage in both *close* (reading passages directly) and *distant* (exploring corpora at a glance) reading for more holistic analysis [JFCS15]. In TextDNA, we have designed such an approach for word usage visualizations with a focus on multiscale pattern finding.

Beyond text visualization, visualizing set membership is a general visualization challenge [AMA*14], as the binary inclusion combinations may yield an exponential number of groupings. While several recent systems (e.g., Onset [SMDS14] and Upset [LGS*14]) address this in a more complete manner, our Sequence Co-Occurrence ordering allows viewing these relations in color-

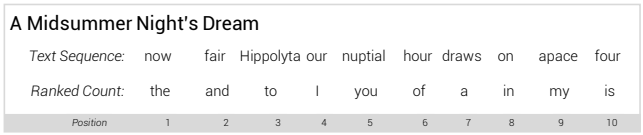


Figure 2: *The first ten words of our two data models applied to* A Midsummer Night's Dream. *The text sequence model emphasizes word location whereas the ranked count model orders words according to frequency patterns. TextDNA's configurable colorfield approach allows researchers to explore text data using these models to address a breadth of analyses.*

field sequence views to explore set relations across different text collections.

3. Task-Driven Data Models

Comparing collections of words within a corpus characterizes structural, stylistic, thematic, and cultural properties of writing and helps people explore how these patterns change between groups of texts. Existing tools visualize these patterns for small sets of words [MSA*11, LRKC10], but require users to define interesting words ahead of time. We held informal discussions with researchers across various Digital Humanities groups to understand the inquiries corpus-level comparison might enable. Due to the unprecedented scalability of our system, these discussions did not identify specific tasks the system should support. However, researchers identified a need to explore data at levels of detail ranging from global comparisons across the entire corpus (e.g., how similar is the language in Shakespeare's comedies and tragedies?), to identifying relations across sets of words or collections (e.g., what words are common in the 1700s and no longer used today? What decades or documents have similar frequent words?), and to mining patterns across individual words (e.g., how does the use of "love" and "hate" vary across different plots?).

In developing TextDNA, we regularly consulted with three literature scholars to observe how they used the tool and collect feedback for refining the prototype. Through this collaboration, we identified three important data attributes used to compare collections of words: *word frequencies*, *word co-occurrences* and *word locations*. Word frequency analyses model how frequently each word appears in a collection or corpus. Word co-occurrence analyses measure which words appear in the same collections. Word location analyses compare where each word appears in different texts. Comparisons often involve some combination of these three properties.

Exploring all three properties for large datasets requires not only flexible encodings, but also flexible data models designed to order and aggregate words to prioritize important patterns. In TextDNA, we use two different models to facilitate these analyses: linear text sequences and ranked word count data (Fig. 2).

Linear text sequences support word localization tasks by modeling the reading order of each text collection. We do so by concatenating all of the words in a particular collection (e.g., a play or a chapter) into a single sequence. The ways in which we divide a corpus into collections defines the kinds of patterns and co-occurrences that users can explore.

On the other hand, **ranked word counts** support word frequency tasks. Word frequency distributions tend to be highly skewed [Li92]: a small number of words occur with high frequency and are often less interesting. To address this challenge, we use a "bag of words" representation for each collection and rank each word within the collection according to its count. Because this model represents a word collection as a ranked bag of words, we can compute word counts over multiple documents to create a single collection and support scalable analysis without requiring data preprocessing. For example, we can use document metadata, such as publication year (§6.1), to represent millions of documents in a single overview visualization. Rank data is also insensitive to magnitude, allowing for comparisons between collections of different sizes.

4. Visualization Design

Our data models frame word usage as a collection of sequences, represented as individual rows of data. Each row has large numbers of words and therefore needs to be expressed compactly. We can use color to express these words using as little space as a single pixel. Flexibility in how we order and color words within each row allows users to tune the display to address different questions.

Several designs have demonstrated the utility of variable mappings to create scalable overviews [KO*07, OJS*11, ADG11]. Due to the number and diversity of words that may occur in each collection and need to explore patterns across multiple data attributes simultaneously, we adapt the aggregate colorfield approach first presented in Sequence Surveyor [ADG11]. This approach has the added benefit that many of the components of this design have already been empirically evaluated in quantitative studies [ACG14, CAFG12]. We provide users six color and position mappings to support different pattern finding tasks (§4.1). We also identify a set of presets mappings that support common word usage analyses to help guide users in making informed mapping choices.

Colorfields provide effective overview at scale. However, humanist researchers must be able to ground insights drawn from the overview display in specific exemplars. To support users in anchoring findings from the overview in individual words, we augment colorfields with new interaction methods (§4.2).

4.1. Configurable Colorfield Design

TextDNA uses the data models outlined in Section 3 to compute a number of relevant properties for each word in each collection. Users can map the color and position of words within a sequence to any pair of these properties (§4.1.1) and compress the resulting colorfields onto the display using techniques that support visual aggregation (§4.1.2).

4.1.1. Task-Driven Data Mappings

Our collaborators identified a broad variety of possible interesting questions that could be addressed by understanding word usage across a corpus. To support these tasks, we compute six different properties associated with each instance of a word in a corpus.

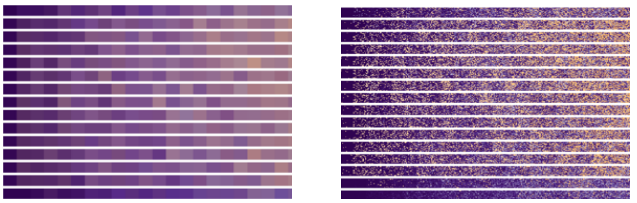
- **Word Rank** represents ordinal position of each word within each collection. In raw text corpora, this is its reading order position within the collection. In ranked corpora, this is the relative commonality of the word within the collection.
- **Word Frequency** represents how often each word appears within the corpus. In raw text corpora, this is the number of instances of the word within the corpus. In ranked corpora, this is the number of collections the word occurs in.
- **Word Count** represents the frequency of each word within the raw text of the underlying corpus.
- **Sequence Frequency** represents the number of sequences each word occurs in.
- **Sequence Co-Occurrence** represents the specific *set* of sequences that a word occurs in. We enumerate all possible subsets of collections (e.g., binary inclusions), give each set a ranking based on the binary pattern of inclusion with smaller sets ranked lower, and order words by the set they are contained in (see Fig. 1).
- **Rank in Reference** represents the Word Rank of a word within a selected reference sequence. Words not in the reference are subsequently ordered according to their Sequence Co-Occurrence.

Each of these properties can be independently mapped to either the color or position of a word within a sequence. Mapping properties to position helps to cluster words with similar properties. Color supports visual aggregation tasks and helps interesting information pop out. Pairing position and color in this way not only supports users in exploring multiple properties of the data simultaneously, but creates high-level gradients and other visual structures to highlight interesting, large-scale trends across collections (see §6 for examples).

Reconfigurable color and position mappings in configurable colorfields allow users to address a broad variety of questions. Two of these encodings, Sequence Co-Occurrence and Rank in Reference, were drawn directly from applications in biology. Our collaborators found these two mappings frequently revealed compelling patterns across broad collections of words. The applicability of these mappings across domains points to the utility of configurable colorfields for a variety of applications. TextDNA can also read in other computed properties of words, such as TF-IDF values, which users can explore through color mapping.

We additionally allow users to select the color ramps used in the visualization from a set of eight sequential, one categorical, and six diverging ColorBrewer ramps [HB03] and a plain grey ramp to reduce visual noise from unimportant words. For properties that could be divided into two meaningful sets (words appearing or not appearing in the reference for Rank in Reference and words in all collections or a subset of collections in Sequence Co-Occurrence), we allow users to map each set to a different ramp to control perceptual properties of the visualization. Dynamically specifying color ramps enables users to tailor the visualization to their aesthetic, perceptual, and data-specific needs.

Users can configure this overview visualization by specifying mappings through dropdowns and interactive selection. The system architecture enables rapid reconfiguration to support fluid analysis (§5). This configurability both supports user preferences and allows expert knowledge to drive the exploration. However, the flexibility afforded by colorfields represents a trade-off: how can we help users choose the right mappings to support their analysis goals? To address this issue, we observed mapping combinations our literature collaborators found particularly useful. We provide preset combinations of these mappings to serve as starting points for different types of exploration.



(a) Averaging represents dominant patterns in the data.

(b) Color weaving reveals blocks of words with interesting variation.

Figure 3: *Scholars can alternate between two aggregate representations of sequence data to explore patterns at different levels of detail.*

4.1.2. Supporting Visual Aggregation

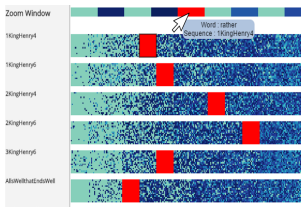
Color provides a compact encoding that can reveal interesting high-level structures in data. While the compactness of colorfields allows for the visualization of over 100 collections simultaneously, an individual sequence often contains more words than there are pixels available to display them.

TextDNA overcomes this limitation by first mapping the data in a sequence to pixel space, creating a fixed mapping of pixels to data values. Then contiguous datapoints are grouped across fixed pixel windows, creating a series of screen-space data "blocks." This screenspace blocking preserves local sequence variations that may be difficult to characterize using statistical techniques. Users can choose to compress data within each block using either averaging or color weaving (Fig. 3). **Averaging** colors a block according to the average color value of its component words. This allows users to explore high-level trends in the dataset with minimal visual clutter. **Color weaving** permutes all color values within a block at the pixel level, facilitating overview while preserving low-level variations from individual words. Color weaving takes advantage of the visual system's ability to summarize color to support multiscale analysis: viewers can explore corpus-level patterns at a glance and also identify important local variance by focusing on a block's distribution of precise, pixel-level information.

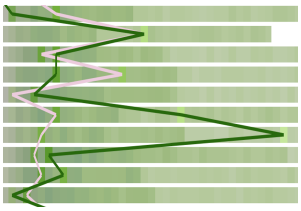
Users can interactively switch between averaging and color weaving to explore different patterns at the overview level. They can then drill down into blocks to recover patterns specific to individual words (§4.2).

4.2. Connecting to Individual Words

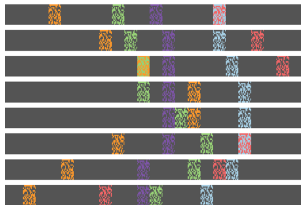
Both prior work [AKV*14, CWG11] and our discussions with collaborators underscored the need to ground corpus-level discoveries in specific exemplars. This introduced an additional user requirement: how can an overview visualization help viewers understand a given word usage pattern at both a corpus and detail level?



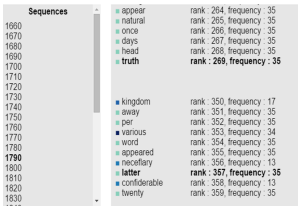
(a) Brushing in the Zoom Window highlights individual words in the overview.



(b) Filtering and linking facilitates comparison between specific words.



(c) Qualitative filtering highlights patterns in larger sets of specific words.



(d) A Words List contains detailed information for all words in a given block and can help identify co-occurring words in other collections.

Figure 4: *TextDNA provides a number of different interaction techniques for exploring patterns among specific words in the context of the larger collection.*

We provide several interaction techniques for identifying and exploring specific sets of words in the context of large datasets (see Fig. 4 and the supplemental video for an overview). To identify locations of interesting words across the corpus, brushing over an aggregate block provides a tooltip that highlights all blocks in the visualization sharing at least one word with the brushed block. It also visualizes the contents of the block in the Zoom Window as an unaggregated colorfield (Fig. 4(a)). If a block illustrates a pattern of particular interest, users can click on a block to lock its contents to the Zoom Window.

Brushing over individual words in the Zoom Window highlights their locations in the dataset, allowing users to contextualize usage of each word in the overview. To compare word-scale patterns across collections, users can click on a word in the Zoom Window to draw a line between all blocks containing that word in the overview visualization (Fig. 4(b)). This overlays a representation similar to a vertical line graph on top of the overview visualization. Users can also directly filter for specific words. Filtering reduces the opacity of all words not in the list to make patterns in the specified words more salient.

While filtering and linking helps identify patterns in specific words, it is generally difficult to differentiate more than three or four words using this technique. To explore larger collections of specific words, users can also apply a *qualitative filter* to specific words (Fig. 4(c)). Qualitative filtering maps the color of a specified set of words to a qualitative color ramp and maps all other words to grey. This makes the specified subset of words salient in comparison to the rest of the corpus. Users can then compare the behavior of these words over their mapped positions across collections.

To identify words that share similar properties across collections, users can scan a Words List associated with each block (Fig. 4(d)). This list allows users to explore sets of words that co-occur with the selected block across all collections through a textual list of all words that co-occur with those in the selected block and any associated metadata. Words are presented in the order of the active position encoding and annotated using the color encoding. With this list, users can identify specific interesting co-occurrences within the dataset and, for raw text data, gather information about the context of interesting words within the original document.

5. System Architecture

Our collaborators wanted a tool that was accessible online for ready distribution, accessibility, and to ease replication of their findings for remote collaborators. This required architecting a system capable of supporting flexible mappings and real-time interaction over a large number of words (upwards of 900,000 in some datasets) in the browser. To address this challenge, we implemented a two-layered front-end rendering scheme: a raster visualization layer leveraging GPU acceleration and an SVG layer to manage interaction. The prototype TextDNA system is open source and available at `http://vep.cs.wisc.edu/TextDNA`.

5.1. The Raster Layer

Because users interactively manipulate color and position mappings in TextDNA, the visualization must render quickly. Minimizing latency when configuring colorfields requires that the system efficiently render aggregate representations of large collections of words. We accomplish this by accelerating overview aggregation and rendering using the GPU. We developed two fragment shaders that compress blocked data into aggregate glyphs (see §4.1.2 for details on the glyph designs). Each shader uses a data texture, representing color mapping values for each word within a block, and a color texture, employing a user-selected color scheme, to generate an aggregate glyph representing the words in a block.

TextDNA system reads data from a JSON object containing collections and data about their component words. Once a color and order mapping are specified, the system creates a list of color mapped values for each word ordered according to the specified order property for the collection. The resulting sequence is then aggregated using the GPU. For averaged blocks, the aggregation shader samples the highest level of the mipmap, using this value to linearly interpolate the color scheme texture. For color weaving, the system generates a one-dimensional index into the largest possible block ($blockheight \times blockwidth$) and randomly permutes those values into a permutation texture. The shader then uses the permutation texture as an index into the data texture at each pixel and linearly interpolates the corresponding color value. Using a permutation texture reduces the overhead of generating a unique permutation for each block and ensures visual continuity between blocks of similar composition. By separating the data and color textures, TextDNA allows users to rapidly apply different color ramps.

5.2. The Vector Layer

To support fluid interaction with large collections of words, we layered a transparent SVG image using D3.js overtop of the raster visualization. With this approach, we can bind an SVG rectangle to the data in each aggregate block and class the rectangle according to the set of words represented by the block. The system then aligns the rectangle with its corresponding location in the rendered image.

This SVG architecture removes the need to rerender each block when the interaction changes. Additionally, by binding the data and position of each block to the DOM, we can use existing selection mechanisms to support efficient searching, brushing, and linking. The resulting system supports interactive performance with upwards of 960,000 words in a web browser.

6. Case Studies

We measure the success of our approach by observing its utility for our collaborators in the humanities. In this section, we present five case studies that illustrate the analyses enabled by TextDNA. We collected these case studies from findings generated by the three literature collaborators who we worked with to refine the system in the course of their own research as well as from researchers interacting with the system at various Digital Humanities workshops and conference demonstrations and from other literature researchers.

These case studies survey word usage patterns identified by scholars across different levels of detail and three datasets. The first dataset is collected from over three million documents in Google Books [MSA*11], from which we extracted the top 5,000 most popular words per decade between 1660 and 2009 (175,000 total words over 35 decades). The second dataset explores the collection of plays written by William Shakespeare (961,304 words in raw text over 36 plays). The third compares raw text patterns in each chapter of *She: A History of Adventure* [Hag86], a serialized novel that exemplifies an important literary style (57,335 words over 15 chapters).

6.1. Google Books

N-grams count how often strings of n words occur in a text. The Google N-Grams dataset [MSA*11] provides these numbers (plus metadata such as publication date) for Google Books. Most methods for exploring this kind of data visualize a handful of words over time, focusing on patterns for specific words rather than specific texts or time periods. Our collaborators found exploring the data in such a large corpus powerful for characterizing the evolution of written language. It also provided an engaging and relatable dataset for scholars from various disciplines to explore during interactive public demonstrations.

To generate this dataset, we grouped the Google N-Grams data by decade from 1660 to 2009, ordered the words within each decade by their frequency within the corpus, and discarded all but the top 5,000 words per decade. In the overview visualizations discussed in this section, decades are ordered chronologically from top to bottom, with each decade represented as a single row. We present three major findings that our collaborators generated using TextDNA with this data.

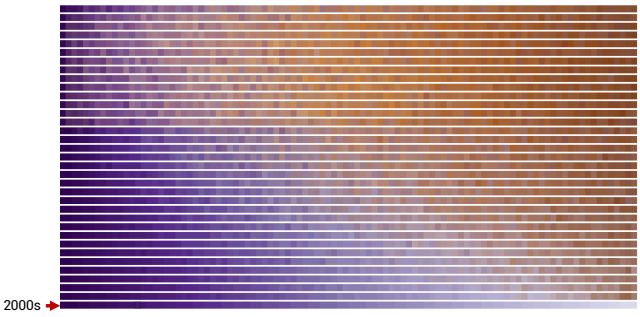


Figure 5: *The top 5,000 words per decade from Google Books ordered by their commonality in each decade (most common on the left) and colored by their commonality in the 2000s (purple are common in the 2000s, orange are not). The orange words form a nearly linear boundary in the upper right, suggesting that language evolves steadily over time.*

6.1.1. CS1: How Writing Evolves

Common words change over time as a function of culture, historical events, and a number of other factors. By comparing the commonality of words in past decades to that of more modern decades, people can assess, for example, how quickly written language is changing and what historical or cultural events might drive this change.

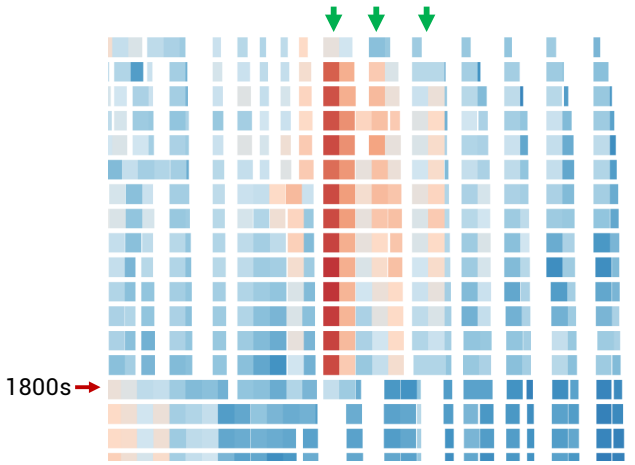
To explore language evolution in the Google corpus, our collaborators ordered words within each decade by their relative commonality (Word Rank, most common on the left) and colored the dataset by the commonality of each word within the most modern decade in the dataset (Rank in Reference, 2000-2009). Figure 5 shows the resulting visualization, where words in purple are among the 5,000 most popular in that decade and words in orange are not.

The encoding shows a roughly linear boundary between the orange and purple words, providing an approximation of how quickly words come into and fall out of use. The regularity of this phenomena within the dataset was both useful and surprising to our collaborators as it suggested that written language evolves steadily over time. Earlier decades have increasingly fewer popular words in common with the 2000s. Our collaborators referred to the decline of a word from modern language as *word death*. An additional angled light purple band follows this boundary in more recent decades. Our collaborators found these words intriguing as they characterize a set of modern terms that have been steadily fading out of use since the mid-19th century.

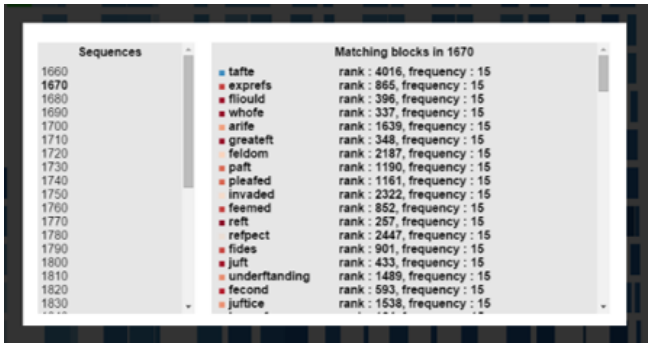
While prior tools enabled scholars to investigate word death for individual words, TextDNA allowed researchers to characterize this phenomena across a vocabulary several orders of magnitude larger than previous explorations. This scalability enabled researchers to generate more holistic conclusions about the evolution of written language.

6.1.2. CS2: Shifting Typographic Conventions

Applying color weaving to the previous example revealed variation (i.e., oranges in fields of purple and purple in fields of orange)



(a) Three red columns in Figure 1 contain words that are very common within the decades that they occur but only in the 5,000 most common words in a subset of decades.



(b) The Words List for these blocks shows that these words are largely instances of the Long S typography convention.

Figure 6: *TextDNA allowed scholars to quickly isolate instances of the Long S typography convention in the Google N-Grams dataset. They used findings from this visualization to develop heuristics for accounting for these errors in their analyses.*

in this pattern, suggesting areas for further exploration. For example, Figure 3(b) shows the most common words in first 15 decades. Some of these orange words cluster to the left of the visualization, indicating that they were once extremely common. To explore these terms, we ordered words within each row according to the sets of decades they appear in (Sequence Co-Occurrence, words in all decades appear to the left and words in a single decade appear to the right) and colored these words according to their relative popularity within each decade (Word Rank, red words are most common and blue are less common).

The resulting visualization is shown in Figure 1 and in detail in Figure 6(a). There are three significant columns of largely red words, indicating a set of words that are quite common in the early decades of the dataset yet abruptly fall out of popularity after roughly 1800. This indicates a significant change in writing around this time. While less popular words only appearing in the top 5,000 words in 40% of the decades is not unusual—words "die" (fall out of popular usage) reasonably frequently—highly common words tend to be those central to written English, such as 'so', 'the', and 'and.' Drilling into these clusters, our collaborators found a large number of words such as 'fo' and 'alfo' that are representative of the "Long S" phenomena—a typography convention that fell out of use around 1800. This convention often causes 's' characters to be misinterpreted as 'f' by OCR, creating words like 'fo' and 'faid' from 'so' and 'said.'

The long S phenomena has been demonstrated in previous work [MSA*11]; however, TextDNA allowed our collaborators to validate the system by finding a known phenomena and also to develop an unprecedented broad picture of the phenomena. They have used this broad picture to reason about heuristics for characterizing the long S in digitized texts by searching the visualization for complementary columns in later decades.

6.1.3. CS3: Cultural Influences on Word Usage

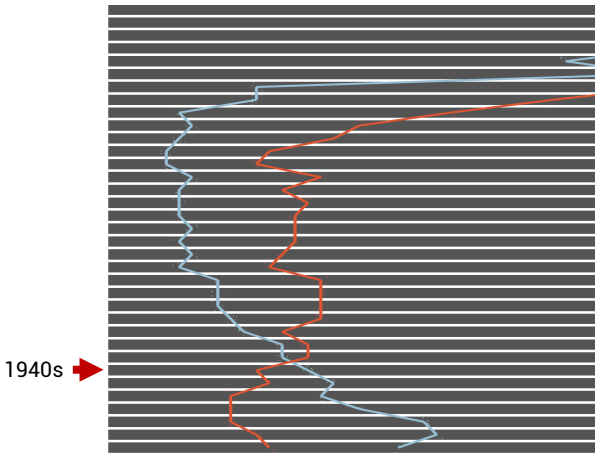
Several collaborators hypothesized that cultural events help drive which words fall in to or out of usage. They explored this hypothesis using a combination of position encodings and interactions. For example, when ordering according to Sequence Co-Occurrence (Fig. 1), they explored collections of words that were unique to interesting sets of decades. For instance, the word "germ" is common only in the 1890s and 1910s, corresponding to the outbreak of the Russian and Spanish flus. In previous systems, such correlations could only be identified by researchers correctly guessing correlated terms. TextDNA's co-occurrence metrics instead divide corpora into sets of correlated words that researchers can use to generatively identify cultural correlations.

Linking and qualitative filtering paired with Word Rank ordering helped users directly compare word usage over time to identify cultural correlations. For example, "army" first became more common than "military" in the 1940s, correlating with World War II (Fig. 7(a)). "Women" first became more popular than "wife" in the 1910s corresponding to global women's suffrage movements: women gained voting rights in 25 countries in this decade [VPH11]. While such correspondences could be traced with previous tools, researchers used TextDNA to explore these detailed patterns and generate corpus-level findings using a single system.

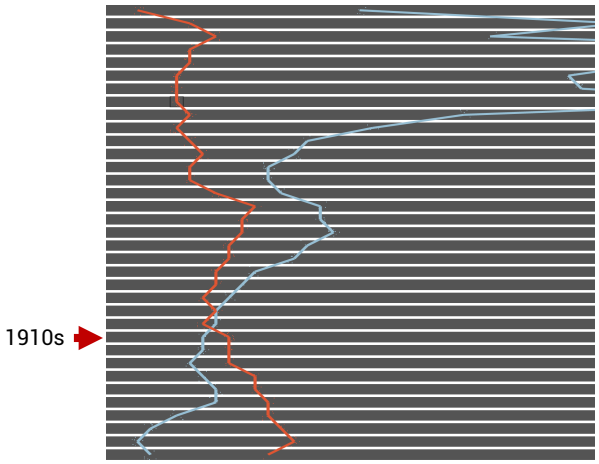
6.2. CS4: Shakespeare

TextDNA enables simultaneous exploration of a corpus and its individual documents. Using color weaving, a collaborator discovered surprising word frequency trends in a raw text dataset of 36 Shakespeare plays. Our collaborator wanted to know not only what words were most frequent in the plays, but also *where* these words occurred. To complete this task, our collaborator used color to represent the position (or Word Rank) of words within the plays. Then the collaborator arranged words in the plays by Sequence Co-Occurrence.

Sequence Co-Occurrence highlighted interesting variation in the Shakespeare raw text dataset. The majority of the first blocks within sequences were yellow, indicating that for most of the plays the



(a) Relative usage of "army" (blue) and "military" (orange).



(b) Relative usage of "women" (blue) and "wife" (orange).

Figure 7: *Ordering words by Word Rank (most common on the left, 1,800th most common on the right) and using qualitative filtering allowed users to identify correlations between word usage and historical events.*

most frequent co-occurring words first appeared near the beginning of the plays. What stood out most, however, were first blocks that were markedly speckled with colors from all parts of the ramp. The color distribution of those blocks suggests that the most frequent co-occurring words were distributed more evenly, perhaps more towards the middle and end, than the other plays. The variation was strongest for the history and tragedy plays *1 Henry VI*, *Julius Caesar*, *King John*, and *Titus Andronicus*. These plays were written early in Shakespeare's career, some suspected to be collaborations. Further work this pattern inspires can be to analyze word co-occurrence with plays of suspected collaborators in Shakespeare attribution studies.

6.3. CS5: She: A History of Adventure

At smaller scales, TextDNA can be used to understand thematic and structural aspects of literary works. Our collaborators hypothesized

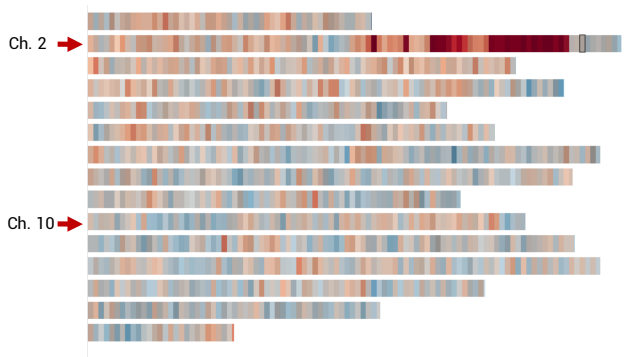


Figure 8: *The raw text of* She: A History of Adventure *with one chapter per sequence. Red areas contain predominantly words in less than half of the chapters, whereas blue are common terms. As the novel progresses, the language shifts to contain larger numbers of blue terms, characteristic of a growing familiarity of the protagonist with the "lost world."*

that the novel *She: A History of Adventure* had substantial linguistic shifts over the course of the story tied to the overall thematic structure of an archetypal "Lost World" story [Hin72]. However, the only methods available to them for validating this hypothesis were intuitions built from close reading.

Dividing the text into its constituent chapters supported analysis of linguistic patterns across the story arc. Ordering the words in the text according to their relative position (Word Rank, first word in a chapter on the left and last on the right) and coloring according to Sequence Co-Occurrence results in the visualization shown in Figure 8. Reddish blocks represent regions of the text where the words are found in less than half of the chapters on average.

Our collaborators found significant support for their hypothesis using TextDNA. The dark red band in the second chapter corresponded with a flashback scene introducing the protagonist to the "lost world" and starting the primary story. The third and fourth chapter contained large clusters of red, which described the flora and fauna of the new world. Later chapters are predominantly composed of more common blue terms with smaller clusters of red indicating references to the supernatural or divine. This shift to common words correlates with the protagonist's growing familiarity with the world. However, the tenth chapter is roughly divided between blue terms in the beginning and red terms in the end suggestive of evil. This band of red terms contains a significant turning point in the plot where the protagonist engages in dark magic ritual.

7. Discussion

In this paper, we introduce TextDNA, a system for exploring usage patterns across large collections of words. This system demonstrates how configurable colorfields can be applied to text analysis in order to support scalable exploration of complex patterns at different levels of detail.

The breadth of findings generated with this system illustrate the effectiveness and generalizability of this approach. Our case studies

demonstrate how configurable colorfields allow users to see how patterns form across collections (§6.1.1), identify co-occurrence across large sets of words (§6.1.2), characterize stylistic patterns between collections (§6.2), and ground larger scale findings in precise examples (§6.1.3). Applying colorfields to different data models facilitate different kinds of analyses and at different scales, ranging from identifying aggregate patterns over millions of books (§6.1) to exploring narrative structure within a single novel (§6.3). Our collaborators noted how the ability to explore supplemental data and to explore patterns in both raw texts and ranked word counts will enable researchers to answer a broad variety of questions about different text corpora.

The overview design of TextDNA also demonstrates the generalizability of configurable colorfields across a variety of potential applications. In our demonstrations, researchers from different domains saw the system's promise for analyses in disciplines such as literary studies, history, and even discourse analysis in psychology and education. While TextDNA focuses predominantly on word localization, co-occurrence, and frequency analyses, this approach could be applied to text at different data granularities, such as n-grams, sentences, chapters, or scenes. Allowing users to configure the color, position, and aggregation of a data sequence addresses a diversity of analysis tasks that integrate different data dimensions. The perceptual affordances of this design allow users to visually aggregate across data collections to identify patterns in large datasets. We find that, in practice, users actively manipulate color, position, and aggregate representation and value the flexibility of configurable colorfields. This, coupled with the popularity of mappings inspired by applications in biology (§4.1), suggests that configurable colorfields have broad utility across a variety of domains.

Our users also appreciated the ability to explore patterns over specific collections of words. These findings support prior work in emphasizing the need to tie overview findings back to the text. We found that, in practice, users explored patterns both starting from the overview and drilling down to specific instances, but also by first searching for interesting words and using patterns in these words to guide analysis at larger scales. By contextualizing patterns related to interesting words in the overview, users were able to identify patterns that cut across multiple levels of detail.

The ideas of TextDNA suggest how the configurable colorfield approach might be applied in other domains. While the data model is specific to the word usage application, it suggests how similar, sequence-oriented models may be used. The methods for exploring details will be useful in any application where both broad patterns and specific elements of the collections are of interest. Further, using the GPU to afford efficient colorfield rendering and re-ordering can make the approach practical in domains with large datasets.

7.1. Limitations & Future Work

The work discussed here represents first steps in understanding how visualization can support users in exploring word usage in text corpora. There are a number of limitations in the current approach that could be addressed in future work.

For example, TextDNA provides a preliminary set of techniques for understanding the specific context in which interesting words are used. Future work should establish techniques that allow users to smoothly drill down from interesting patterns into the text of the original documents. This work could employ word-scale visualizations [GWFI15] to help ground such small-scale explorations in the overview.

Additionally, text copora are currently modeled in a preprocessing step. Future work could explore how visualization systems might guide users in fluidly exploring patterns across different data models or levels of detail, such as moving from aggregate ranked sequences over time to the raw text of a subset of documents.

The number of possible text sequences that can be visualized at a given time is currently bounded by the available screenspace (we have successfully visualized roughly 100 sequences simultaneously on a standard desktop). Future work should explore automated methods for further increasing the scalability of word usage analysis to support comparing larger numbers of collections.

While configurable colorfields enable a broad variety of explorations, the system can also be overwhelming to inexperienced users. We anticipate presets will provide an accessible entry point for new users and are actively working with literature scholars to develop tutorials showing how the system can be used for various tasks. Our experiences with interactive demonstrations and the feedback we have received suggest that users are excited about the system and can readily engage with the tool to generate interesting findings. However, further understanding the balance between the exploratory power offered by configurable colorfields and ease of use of the resulting system is important future work.

8. Conclusion

This paper presents TextDNA, a visualization system for comparing large collections of words. In designing TextDNA, we make three primary contributions. First, we introduce a data model that supports aggregating word frequency data across large collections of documents. Second, we show how configurable colorfields can support comparisons between large collections and address usability needs for humanist scholarship. Finally, we demonstrate how such an interactive system can be architected for the browser by leveraging the GPU to accelerate aggregate rendering. The resulting system supports scholars in exploring thematic, cultural, and stylistic patterns across a broad variety of texts at dramatically larger scales than previously possible.

Acknowledgments:

We would like to thank Catherine DeRose and Robin Valenza for their engagement with and support of this work and Ben Bederson for inspiration for part of Case Study 3. This work was funded by NSF award IIS-1162037 and a grant from the Andrew W. Mellon Foundation.

References

[ACG14] ALBERS D., CORRELL M., GLEICHER M.: Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 551–560. 3

[ADG11] ALBERS D., DEWEY C., GLEICHER M.: Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2392–2401. 2, 3

[AKV*14] ALEXANDER E., KOHLMANN J., VALENZA R., WITMORE M., GLEICHER M.: Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2014), IEEE, pp. 173–182. 2, 4

[AMA*14] ALSALLAKH B., MICALLEF L., AIGNER W., HAUSER H., MIKSCH S., RODGERS P.: Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. In *EuroVis - STARs* (2014), Borgo R., Maciejewski R., Viola I., (Eds.), The Eurographics Association. 2

[CAFG12] CORRELL M., ALBERS D., FRANCONERI S., GLEICHER M.: Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 1095–1104. 3

[CCP09] COLLINS C., CARPENDALE S., PENN G.: Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 1039–1046. 2

[CMH12] CHUANG J., MANNING C. D., HEER J.: Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), ACM, pp. 74–77. 2

[CVW09] COLLINS C., VIEGAS F. B., WATTENBERG M.: Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), IEEE, pp. 91–98. 2

[CWG11] CORRELL M., WITMORE M., GLEICHER M.: Exploring collections of tagged text for literary scholarship. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 731–740. 2, 4

[DZG*07] DON A., ZHELEVA E., GREGORY M., TARKAN S., AUVIL L., CLEMENT T., SHNEIDERMAN B., PLAISANT C.: 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* (2007), ACM, pp. 213–222. 2

[GCL*13] GENG Z., CHEESMAN T., LARAMEE R. S., FLANAGAN K., THIEL S.: Shakervis: Visual analysis of segment variation of german translations of shakespeareâĂŹs othello. *Information Visualization* (2013), 1473871613495845. 2

[GLK*13] GORG C., LIU Z., KIHM J., CHOO J., PARK H., STASKO J.: Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics 19*, 10 (2013), 1646–1663. 2

[GWFI15] GOFFIN P., WILLETT W., FEKETE J.-D., ISENBERG P.: Design considerations for enhancing word-scale visualizations with interaction. In *IEEE Transactions on Visualization and Computer Graphics* (2015). 9

[Hag86] HAGGARD H.: Rider. she: A history of adventure. *Serialized in The Graphic* (1886). 6

[HB03] HARROWER M., BREWER C. A.: Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003), 27–37. 4

[Hin72] HINZ E. J.: Rider haggard's "she": An architypal "history of adventure". *Studies in the Novel* (1972), 416–431. 8

[JFCS15] JÄNICKE S., FRANZINI G., CHEEMA M. F., SCHEUERMANN G.: On close and distant reading in digital humanities: A survey and future challenges. 2

[JKM12] JANKOWSKA M., KESELJ V., MILIOS E.: Relative n-gram signatures: Document visualization at the level of character n-grams. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 103–112. 2

[KBK*11] KRSTAJIĆ M., BERTINI E., KEIM D., ET AL.: Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2432–2439. 2

[Kei96] KEIM D. A.: Pixel-oriented visualization techniques for exploring very large data bases. *Journal of Computational and Graphical Statistics 5*, 1 (1996), 58–77. 2

[KO*07] KEIM D., OELKE D., ET AL.: Literature fingerprinting: A new method for visual literary analysis. In *2007 IEEE Symposium on Visual Analytics Science and Technology* (2007), IEEE, pp. 115–122. 2, 3

[LGS*14] LEX A., GEHLENBORG N., STROBELT H., VUILLEMOT R., PFISTER H.: Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), to appear. 2

[Li92] LI W.: Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory 38*, 6 (1992), 1842–1845. 3

[LRKC10] LEE B., RICHE N. H., KARLSON A. K., CARPENDALE S.: Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1182–1189. 2, 3

[MSA*11] MICHEL J.-B., SHEN Y. K., AIDEN A. P., VERES A., GRAY M. K., PICKETT J. P., HOIBERG D., CLANCY D., NORVIG P., ORWANT J., ET AL.: Quantitative analysis of culture using millions of digitized books. *Science 331*, 6014 (2011), 176–182. 2, 3, 6, 7

[OJS*11] OELKE D., JANETZKO H., SIMON S., NEUHAUS K., KEIM D. A.: Visual boosting in pixel-based visualizations. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 871–880. 2, 3

[Pal02] PALEY W. B.: Textarc: Showing word frequency and distribution in text. In *IEEE Symposium on Information Visualization Poster Session* (2002), vol. 2002. 2

[SMDS14] SADANA R., MAJOR T., DOVE A., STASKO J.: Onset: A visualization technique for large-scale binary set data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), to appear. 2

[VPH11] VILLANI L., PROVOST C., HILAIRE E.: A timeline of women's right to vote - interactive. *The Guardian* (July 2011). URL: http://www.theguardian.com/global-development/interactive/2011/jul/06/un-women-vote-timeline-interactive. 7

[VW08] VIÉGAS F. B., WATTENBERG M.: Timelines tag clouds and the case for vernacular visualization. *Interactions 15*, 4 (2008), 49–52. 2

[WJS*15] WANNER F., JENTNER W., SCHRECK T., STOFFEL A., SHARALIEVA L., KEIM D. A.: Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis. *Information Visualization* (2015), 1473871615576925. 2

[WLS*10] WEI F., LIU S., SONG Y., PAN S., ZHOU M. X., QIAN W., SHI L., TAN L., ZHANG Q.: Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), ACM, pp. 153–162. 2

[WV08] WATTENBERG M., VIÉGAS F. B.: The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1221–1228. 2