Visual Text Analysis in Digital Humanities

S. Jänicke, G. Franzini, M. F. Cheema and G. Scheuermann

1Image and Signal Processing Group, Department of Computer Science, Leipzig University, Germany
  {stjaenicke,faisal,scheuermann}@informatik.uni-leipzig.de
2Göttingen Centre for Digital Humanities, University of Göttingen, Germany
  gfranzini@gcdh.de

Abstract

In 2005, Franco Moretti introduced Distant Reading to analyze entire literary text collections. This was a rather revolutionary idea compared to the traditional Close Reading, which focuses on the thorough interpretation of an individual work. Both reading techniques are the prior means of Visual Text Analysis. We present an overview of the research conducted since 2005 on supporting text analysis tasks with close and distant reading visualizations in the digital humanities. Therefore, we classify the observed papers according to a taxonomy of text analysis tasks, categorize applied close and distant reading techniques to support the investigation of these tasks, and illustrate approaches that combine both reading techniques in order to provide a multifaceted view of the textual data. In addition, we take a look at the used text sources and at the typical data transformation steps required for the proposed visualizations. Finally, we summarize collaboration experiences when developing visualizations for close and distant reading, and we give an outlook on future challenges in that research area.

Keywords: digital humanities, survey, visual text analysis, close reading, distant reading

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology

1. Introduction

Traditionally, humanities scholars carrying out research on a specific or on multiple literary work(s) are interested in the analysis of related texts or text passages. But the digital age has opened possibilities for scholars to enhance their traditional workflows. Enabled by digitization projects, humanities scholars can nowadays reach a large number of digitized texts through web portals such as Google Books [Goo15] and Internet Archive [Arc15]. Digital editions exist also for ancient texts; notable examples are PHI [PHI15] and the Perseus Digital Library [Per15].

This shift from reading a single book “on paper” to the possibility of browsing many digital texts is one of the origins and principal pillars of the digital humanities domain, which helps to develop solutions to handle vast amounts of cultural heritage data – text being the main data type. In contrast to the traditional methods, the digital humanities allow to pose new research questions on cultural heritage datasets. Some of these questions can be answered with existent algorithms and tools provided by the computer science domain, but for other humanities questions scholars need to formulate new methods in collaboration with computer scientists.

Developed in the late 1980s [Hoc04], the digital humanities primarily focused on designing standards to represent cultural heritage data such as the Text Encoding Initiative (TEI) for texts [TEI15], and to aggregate, digitize and deliver data. In the last years, visualization techniques have gained more and more importance when it comes to analyzing data. For example, Saito [SOI10] introduced her 2010 digital humanities conference paper with: “In recent years, people have tended to be overwhelmed by a vast amount of information in various contexts. Therefore, arguments about ‘Information Visualization’ as a method to make information easy to comprehend are more than understandable.”

A major impulse for this trend was given by Franco Moretti. In 2005, he published the book “Graphs, Maps, Trees” [Mor05], in which he proposes the so-called distant reading approaches for textual data, which steered the traditional way of approaching literature towards a completely new direction. Instead of reading texts in the traditional way...
This survey observes text analysis tasks of humanities scholars – e.g., literary scholars, historians and philologists –, and the visualization techniques that have been developed in order to support these tasks. By providing a text analysis task taxonomy, categorizing applied close and distant reading techniques and outlining strategies that combine close and distant reading visualizations, we present an overview suitable for visualization scholars facing related digital humanities text analysis tasks. We further investigate the following questions:

- What are the used text sources and which data transformations are applied in order to investigate text analysis research questions with close and distant reading visualizations?
- Which experiences are reported regarding collaborations between visualization experts and humanities scholars?
- What are future challenges for visualization scholars concerning visual text analysis to further improve the support for humanities scholars?

1.1. Relation to the Previous Article

The focus of the previous version of this survey [JFCS15] was to illustrate the diversity of applied visualization techniques that support the close and distant reading of texts in digital humanities applications – enriched with used visualization tools, collaboration experiences of visualization researchers working together with humanities scholars and future challenges.

This survey extension aims to give visualization scholars new to the field of digital humanities an adequate overview of related works in order to support carrying out successful digital humanities projects. As an application domain for information visualization, these projects gain their motivation from humanities research questions on texts. Therefore, we introduce a taxonomy that groups the papers into classes of text analysis tasks in order to guide visualization researchers with similar tasks to related works that provide close and distant reading solutions. In addition, we list text sources, and we take a closer look at data transformations, which are substantial steps in order to afford designing valuable visualizations. Furthermore, we extended the collaboration experiences and future challenges sections. Finally, this survey considers 22 more related works.

2. Means of Visual Text Analysis

Close reading and distant reading are the prior means for the visual analysis of textual sources in digital humanities scenarios. While the close reading of a text is a traditional method that has its roots in antiquity when Aristotle close read the works of Plato [McC15], distant reading is a rather novel idea that was introduced by Franco Moretti at the beginning of the 21st century. In contrast to Moretti, Jockers uses the terms micro- and macroanalysis instead of close and distant reading [Joc13]. Inspired by micro- and macroeconomics, he focuses on quantitative literary text analysis using statistical analysis methods. As the methods we analyzed are more related to visualization, we decided to use the traditional, more common terms close and distant reading, but we also considered related works using different terminologies. This section introduces close and distant reading techniques and draws a line from the digital humanities to information visualization by combining both techniques.

2.1. Close Reading

The close reading of a text became a fundamental method in literary criticism in the 20th century [Haw00]. Nancy Boyles [Boy13] defines it as follows: “Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension.” In other words, close reading is the thorough interpretation of a text passage by the determination of central themes and the analysis of their development. Moreover, close reading includes the analysis of (1) individuals, events, and ideas, their development and interaction, (2) used words and phrases, (3) text structure and style, and (4) argument patterns [Jas01]. The result of a traditional close reading approach is shown in Figure 1. In this example, the scholar used various methods to annotate various features of the source text, e.g., the usage of different colors (blue, red, green) and underlining techniques that support the close and distant reading visualizations, we present an overview of different colors (blue, red, green) and underlining styles (straight or wavy lines, circles). Furthermore, numerous thoughts are written next to the corresponding sentences. Although most humanities scholars are trained in this traditional approach of close reading, today’s large availability of digitized texts and of digital editions through web portals

Figure 1: Traditional close reading of the second chapter of Charles Dickens’ David Copperfield (Figure reproduced with permission from Kehoe et al. [KG13]).
Figure 2: Digital close reading of the second chapter of Charles Dickens’ *David Copperfield* with *eMargin* (Figure reproduced with permission from Kehoe et al. [KG13]).

like Google Books [Goo15] or Project Gutenberg [Pro15], opens up new possibilities for close reading, and especially for sustainable and collaborative annotation.

Figure 2 shows a straightforward approach of visualizing various scholars’ annotations of a digital edition [KG13] within the web-based environment *eMargin* [eMa15]. There, colors are used to highlight different text features, and a pop-up window lists the comments of collaborating scholars. In Section 8.1 we outline different approaches to support text analysis with close reading by visualizing supplementary human- or computer-generated information.

2.2. Distant Reading

While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s). Moretti [Mor13] describes distant reading as “a little pact with the devil: we know how to read texts, now let’s learn how not to read them.” In 2005, he introduces his idea of distant reading [Mor05] with three examples using:

- **graphs** to analyze genre change of historical novels,
- **maps** to illustrate geographical aspects of novels, and
- **trees** to classify different types of detective stories.

But, the proposed methods and the intention of distant reading are controversial in the humanities [GH11a, CRS*14], as they quantify and abstract texts at an expense of reflecting the actual ambiguity and complexity of literary forms [Dru11, Mar12]. However, many works in the digital humanities domain are based upon Moretti’s idea. Figure 3 shows Posavec’s Literary Organism [Pos07] – a distant reading of Jack Kerouac’s *On the Road* in the form of a tree. While a non-interactive infographic, Posavec’s approach perfectly illustrates the idea behind Moretti’s distant reading, as it turns away from the traditional close reading by providing an abstract view of a literary text. The branching structure represents the ordered hierarchy of content objects from chapters down to words, and themes are drawn with different colors. In Section 8.2 we present a list of different distant reading techniques developed for a wide range of text analysis research questions in the digital humanities.

2.3. Combining Close and Distant Reading

In his digital humanities collaborations, Correll worked together with literary scholars who were unfamiliar with distant reading views [CAA*14]. Here, providing links to close reading was an important method to support distant reading hypotheses. During our literature research, we also discovered a multitude of works involving close reading and interfaces, which provide distant reading visualizations that allow to interactively drill down to specific portions of the data. This suggests that direct access to the source texts is indeed very important for humanities scholars when working with visualizations. For example, Bradley [Bra12] asks whether it is “possible to develop a visualization technique that does not destroy the original text in the process.” Similarly, Beals [Bea14] asks: “In an age where distant reading is possible, is close reading dead?” Coles et al. argue that distant reading visualizations cannot replace close reading, but they can direct the reader to sections that may deserve further investigation [CL13].

When distant reading views are interactively used to switch to close reading views, the Information Seeking Mantra “Overview first, zoom and filter, details-on-demand” [Shn96] is accomplished. It follows that an important task for the development of visualizations is to provide...
an overview of the data that highlights potentially interesting patterns. A drill down on these patterns for further exploration is the bridge between distant and close reading.

3. Visual Text Analysis

Figure 4 reflects a typical visual text analysis process in digital humanities workflows. The text analysis task at hand affects all steps of this process. First, the text sources are selected in accordance to the research task. Next, it is important to apply appropriate data transformations in order to design close and/or distant reading visualizations that are beneficial to investigate the given text analysis task. But, a humanities scholar often refers to both the visualizations and the underlying texts in order to gain insights concerning the research question. While the next section defines the scope of this survey, the following sections explain the various components of the visual text analysis process in detail. Section 5 lists used text sources, and Section 6 provides a brief overview of data transformation techniques. In Section 7, we provide a taxonomy of text analysis tasks. Applied close and distant reading techniques are outlined in Section 8.

4. Scope

In order to generate the research paper pool of this survey, we used the publication year of Franco Moretti’s book on distant reading techniques “Graphs, Maps, Trees” [Mor05], 2005, as a starting point. We manually scanned through the major related visualization and digital humanities journals and conference proceedings in order to generate an appropriate snapshot of existent research on visualizations for close and distant reading. To receive a processible list of related works, we required to narrow the scope of our survey.

First, we only considered information visualization publications. For example, we did not include papers from human-computer interaction issues (e.g., ACM Conference on Human Factors in Computing Systems (CHI)). Table 1 lists the number of related information visualization papers examined. We also considered approaches developed for other data types, where at least one related use case was provided for a cultural heritage dataset. The TVCG journal table entry includes all found papers presented at the IEEE Symposium on Information Visualization (InfoVis) as well as seven papers presented at the IEEE Symposium on Visual Analytics Science and Technology (VAST). Likewise, the Computer Graphics Forum entry includes all related papers also contained in the proceedings of the Joint Eurographics–IEEE VGTC Symposium on Visualization (EuroVis). No related papers were found in the proceedings of the IEEE Pacific Visualization Symposium (PacificVis) and of the International Conference on Information Visualisation (IV), nor in the IEEE Computer Graphics and Applications journal.

Second, we included related works from the major digital humanities realms into our survey. Thereby, we decided to consider the proceedings of the Annual Conference of the Alliance of Digital Humanities Organizations, which yields a suitable snapshot of the research conducted in that field, although only short papers are contained. The 60 related works presented at this major digital humanities conference (see Table 2) indicate the importance of close and distant reading visualizations for text analysis tasks. In 2014, a total of 345 individual papers were presented at the conference, of which 33 thematize information visualization techniques for cultural heritage data (9.6%), including 14 papers (4.1%) about close and/or distant reading relevant for our survey.

<table>
<thead>
<tr>
<th>Journal/Proceedings</th>
<th>#Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE Transactions on Visualization and Computer Graphics (TVCG)</td>
<td>16</td>
</tr>
<tr>
<td>IEEE Symposium on Visual Analytics Science and Technology (VAST)</td>
<td>6</td>
</tr>
<tr>
<td>Computer Graphics Forum</td>
<td>5</td>
</tr>
<tr>
<td>Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP)</td>
<td>2</td>
</tr>
<tr>
<td>Information Visualisation Journal</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Visualization papers examined.

<table>
<thead>
<tr>
<th>Journal/Proceedings</th>
<th>#Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proceedings of the Annual Conference of the Alliance of Digital Humanities Organizations</td>
<td>60</td>
</tr>
<tr>
<td>Literary and Linguistic Computing / Digital Scholarship in the Humanities</td>
<td>18</td>
</tr>
<tr>
<td>Digital Humanities Quarterly</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2: Digital humanities papers examined.
The collection is completed by related papers published in two major digital humanities journals, Literary and Linguistic Computing (Digital Scholarship in the Humanities as of 2015) and Digital Humanities Quarterly.

Third, we did not consider works published in humanities realms such as Shakespeare Quarterly or American Literary History.

4.1. Considered Research Papers

In order to be considered for our survey, a paper needed to fulfill the following requirements:

- **Textual data:** The visualization is a solution for research questions on an arbitrary text corpus, either a small text unit such as a poem, a large text unit such as a book, or a whole text collection. For example, we did not include a timeline visualization of Picasso’s works [MFM08] as it is based upon artworks.

- **Cultural heritage:** The underlying textual data has a historical value. While not considering approaches dealing with texts extracted from social media or wiki systems (e.g., a social network visualization of philosophers presented in [AL09], which is based upon relationships modeled in Wikipedia), we took into account visualizations for newspaper collections. This decision includes some visualization papers that do not directly relate to the humanities. But the proposed techniques are based upon or tested with contemporary newspaper collections, which are indeed part of cultural heritage.

- **No straightforward metadata visualization:** We only considered papers that provide a visualization that is based upon the inherent textual content. We omitted methods that only use the texts’ associated metadata. An example is given by two graphs displaying relationships between texts. Whereas the related network graph presented in [Ede14] is determined by analyzing stylistic features among the textual contents of novels, the unrelated graph visualization in [Pin10] uses Amazon recommendations to determine relationships between books.

- **No traditional charts:** In the digital humanities, the word visualization is frequently used. Traditional charts displaying statistical information such as line or bar charts are also labeled as visualizations. Based on the information visualization definitions given by Card [CMS99] and the UIUC DLI Glossary [UIU98], we only considered papers that provided computer-supported, non-traditional visual representations of abstract data. In contrast, we do not require interactive methods as humanities scholars often gain valuable insights using non-interactive visualizations. However, most proposed techniques implement certain means of interaction.

Altogether, in order to be considered, a paper required a research question on textual data of historical interest emphasizing content rather than metadata and supporting the close and/or distant reading of texts.

5. Text Sources

The focus of interest of the papers in our collection includes four major text types.

- **Single literary texts** often motivate the development of close and distant reading techniques, e.g., Adam Smith’s *The Wealth of Nations* [BJ14], Gertrude Stein’s *The Making of Americans* [CDP07], Herodotus’ *Histories* [BPB10], or *The Castle of Perseverance* [Pet14].

- **POI collections** are text corpora containing several works by a particular author (e.g., William Faulkner novels [DNCM14]), politician (e.g., the papers of Thomas Jefferson [Kle12] or the Kissinger Collection [Kau15]), or any other “person of interest” (POI), e.g., alchemical manuscripts written by Isaac Newton [WH11] or statements about Emily Carr [HSC08].

- **Text editions** are collections of variants of the same source text, e.g., research questions concerning German translations of Shakespeare’s *Othello* [GCL13], English translations of the *Bible* [JGBS14], or reprints of Nathaniel Hawthorne’s short story *The Celestial Railroad* [Cor13].

- **Text collections** contain a multitude of texts, which usually belong to a particular typology. Among others, this includes the collection of poems [Wil15a], biographies [Boo13], tales [W13a], novels [Ede14], medieval texts [JW13], or news articles [KLB14].

This variety of source texts reflects the diversity of text analysis research questions raised by humanities scholars. On the other hand, it suggests the requirement of multifarious close and distant reading techniques.

While more and more digital editions and archives are made available online [GTW13, FMT15], distant reading studies are few and far between due to the limited number of available text collections [Und15]. It follows that the more and the larger the accessible text collections, the bigger the scope for close and distant reading analyses and the higher the probability to generate new research challenges for visualization scholars. The issue with this seemingly logical conclusion, however, is the unavoidable and enormous amount of manual work required in order to assemble [Und15], encode and upload [Wei15] such text collections to the web for others to generate derivative outputs. Some of the papers provide information about used publicly available data sources, which we list in Table 3.

6. Data Transformation Aspects

Many papers in our collection do not provide sufficient information about applied preprocessing steps to transform the given textual data into the visualization’s input format. Occasionally, a visualization directly processes annotated text in XML format (e.g., [Pie10, Boo13, BGHH14]). In particular, many techniques are based upon documents in the XML-based TEI format (e.g., [BHW11, CGM12]).
## Table 3: Publicly available data sources.

<table>
<thead>
<tr>
<th>Data Source URL Used by</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perseus Digital Library</td>
<td>[BPBI10]</td>
</tr>
<tr>
<td>Project Gutenberg</td>
<td>[CTA<em>13, GCL</em>13, KO07, GTAHS15]</td>
</tr>
<tr>
<td>TextGrid Repository</td>
<td>[TFK15, JKH*15]</td>
</tr>
<tr>
<td>HathiTrust</td>
<td>[AGZH15]</td>
</tr>
<tr>
<td>The Swinburne Project</td>
<td>[WH11]</td>
</tr>
<tr>
<td>The Chymistry of Isaac Newton Project</td>
<td>[WH11]</td>
</tr>
<tr>
<td>The Papers of Thomas Jefferson</td>
<td>[Kle12]</td>
</tr>
<tr>
<td>Internet Shakespeare Editions</td>
<td>[RSDCD*13]</td>
</tr>
<tr>
<td>Bob Gibson Collection of Speculative Fiction</td>
<td>[HFM16]</td>
</tr>
<tr>
<td>The Poetess Archive</td>
<td>[FS11, CGM*12]</td>
</tr>
<tr>
<td>Folklore and Mythology Electronic Texts</td>
<td>[RFH14]</td>
</tr>
<tr>
<td>Eighteenth Century Collection Online</td>
<td>[JOL*15]</td>
</tr>
<tr>
<td>MySword (Bible Collection)</td>
<td>[JG15]</td>
</tr>
<tr>
<td>1641 Depositions</td>
<td>[ÓML14]</td>
</tr>
<tr>
<td>Digital National Security Archive</td>
<td>[Kau15]</td>
</tr>
<tr>
<td>PoliInformatics</td>
<td>[Poi15]</td>
</tr>
<tr>
<td>Projekt Deutscher Wortschatz</td>
<td>[KLB14]</td>
</tr>
<tr>
<td>VroniPlag Wiki</td>
<td>[RPSF15]</td>
</tr>
<tr>
<td>Europe Media Monitor</td>
<td>[KBK11]</td>
</tr>
<tr>
<td>Digital National Security Archive</td>
<td>[Kau15]</td>
</tr>
<tr>
<td>Collected Project Corpora</td>
<td>[KZ14]</td>
</tr>
<tr>
<td>Trading Consequences</td>
<td>[HAC*15]</td>
</tr>
<tr>
<td>British National Corpus</td>
<td>[Bea08, ARLC*13]</td>
</tr>
<tr>
<td>Documenting the American South</td>
<td>[CTA*13]</td>
</tr>
<tr>
<td>Metadata Offer New Knowledge</td>
<td>[VCPK09]</td>
</tr>
<tr>
<td>The Internet Movie Script Database</td>
<td>[HPR14]</td>
</tr>
</tbody>
</table>

The internet movie script database

The internet movie script database

The internet movie script database

The internet movie script database
which has become the humanities leading technology to map the structure of a digital text edition [Sin13]. XSLT stylesheets are basic ways to transform the TEI encoded information into a meaningful visualization of an individual text (e.g., [Cor13, Pie13, HKT14]). But most distant reading techniques require more sophisticated preprocessing steps – an brief summary of common data transformation approaches is shown below.

**Tokenization and normalization** are rather rudimentary natural language processing methods first applied to segment raw text sources. The then determinable frequency distribution of words is a valuable basis for various tasks (e.g., stylometric analyses [CEI*14, Ede14]), and can be clearly visualized in the form of tag clouds to support the exploration of word statistics (e.g., [Bea08, GTAHS15, JBR*15]). Vector space models can be used to list term frequencies per text and support a variety of text analysis tasks (e.g., [DFM*08, GLC*13, KOTM13]). On the other hand, counting n-grams allows to draw more specific statements about a text corpus (e.g., [CDP*07, Bea12, MH13]). But tokenization and normalization are also necessary steps for the data transformation methods described below.

**Sequence alignments** are computed when investigating research questions concerning textual re-use (e.g., [BGHE10, JGBS14]) and for the analysis of similarities and differences among various text editions (e.g., [WJ13b, JGF*14]). In such scenarios one typically applies the Gothenburg model, which includes tokenization, normalization, alignment, analysis and visualization [Got15]. One of the web-based tools implementing this model is CollateX [HDvHM*15].

**Part-of-speech (POS) tagging** is a frequently applied preprocessing technique to automatically annotate the words of a corpus according to their part of speech category. The use of tools like the Stanford POS tagger [TKMS03] is a mandatory basis for investigating diverse research questions. Typically, words and their relationships are explored (e.g., [KKL*11, MH13]) or linguistic patterns are extracted from a corpus (e.g., [Murl11, RFH14]). Furthermore, POS tagging is used to analyze phonetic features [CTA*13] or for research questions concerning stylometry [K007].

**Named entity recognition (NER)** is the practice of extracting named entities such as places or persons from texts. Preprocessing steps like part-of-speech tagging can be applied to automatically list named entity candidates [GH11b]. With the help of lexicons, named entities are subsequently classified. For example, the Pleiades gazetteer [Ple15] is used for the extraction of ancient place names [EJ14], and DBPedia [LIJ*14] supports the discovery of commodities in [HAC*15]. The Stanford Named Entity Recognizer [FGM05] is a popular tool for automatic named entity extraction. The manual collection of named entities is not uncommon and guarantees the highest precision (e.g., [JW13, Wil15b]). Occasionally, named entities are already present when using annotated TEI files as data sources (e.g., [BB15a, TFK15]).

**Topic modeling** algorithms are fundamental for topic-related analyses of text collections. The Latent Dirichlet allocation is the most often applied topic model [Bnj03]. It requires a predefined number of topics, which are determined automatically based on the words contained in the text corpus (e.g., [AKV*14, BJ14]). The topic model can be used to cluster texts thematically [Wol13] – as happens with text classification methods (e.g., [PSA*06, DFM*08]) – or to define the similarity among the texts of a corpus [Joc12]. When temporal metadata is provided, the change of topics can be analyzed (e.g., [ARR*12, CLWW14]).

**Semi-automatic approaches** reflect the importance of integrating the humanities scholar into the data transformation process. For instance, the scholar’s knowledge is required when manually generating or validating a training set to produce an appropriate data mining classifier (e.g., [PSA*06, KKL*11, KJW*14]). Other methods include semi-automatic alignments [GZ12] and the annotation of TEI documents (e.g., [To13, OGH15]). Sometimes, even the visualization entirely depends on manually collected data through crowdsourcing (e.g., [WMN*14, RPF15, HFM16, JFS16]).

### 7. Taxonomy

There has been extensive research done in developing taxonomies for information visualization in the last decades. Unfortunately, these taxonomies were either too general (e.g., [BM13, RAW*15]) or too specific (e.g., [LPP*06, KKC15]) to be used for our paper collection. Therefore, we defined a taxonomy focusing the underlying text analysis tasks in the digital humanities domain (see Figure 5). A detailed classification of papers is given in Table 4. Papers focusing a single text analysis task are grouped to a single – the best fitting – category. The few works providing visualization methods for two text analysis tasks each appear in two categories [RRRG05, WH11, Wol13, Kau15]. The taxonomy consists of five major categories:

![Figure 5: Taxonomy of text analysis tasks.](image-url)
Table 4: Classification of papers according to the taxonomy of text analysis tasks.

The analysis of named entities is a common text analysis task of humanities scholars that is supported with close and distant reading visualizations. When extracting places, fictional or reported geographies of a single text or a whole collection can be explored. The extraction of persons is required to analyze social networks of individuals or of characters in a story. Miscellaneous tasks focus on other (e.g., encyclopedia entries [HAHB15]) or on multiple named entities (e.g., commodities and locations [HAC*15]).

The analysis of topics inherent in a text corpus supports text analysis tasks that require both close and distant reading techniques. Popular tasks are topic extractions, so that major topics in the source texts can be tracked and topic-related text passages can be discovered. The presence of temporal data allows for the analysis of topical evolution, and on the basis of the found topics, a topical clustering of a corpus is possible.

An analysis of similar patterns that includes the discovery, the alignment and the visualization of similar text segments among the texts of a given collection is a typical text analysis task in the digital humanities. Dependent on the length of patterns, we divide the tasks belonging to that category into three sets. While the analysis of linguistic patterns concerns short phrases, text re-use analysis focuses on determining deliberately re-used text segments (e.g., quotes or plagiarized passages). In papers grouped to the category text edition comparison, the humanities scholar is more interested in analyzing the variants between the text editions.

Some tasks focus an individual literary work, which we call text of interest. Sophisticated close reading techniques are often applied but are sometimes coupled with distant reading representations of the textual content. The underlying research tasks vary from visualizing text interpretations to the analysis of sound of literary works (mostly poems),

S. Jänicke, G. Franzini, M. F. Cheema & G. Scheuermann / Visual Text Analysis in Digital Humanities

© 2016 The Author(s)
Computer Graphics Forum © 2016 The Eurographics Association and John Wiley & Sons Ltd.
<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Font size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyphs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connections</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat maps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag clouds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timeline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom-up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-down &amp; Bottom-up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-down</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**named entities**

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>places</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>persons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miscellaneous</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**topics**

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>extraction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evolution</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>clustering</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**similar patterns**

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>linguistic patterns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text re-use</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text edition comparison</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**text of interest**

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>interpretation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>story flow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miscellaneous</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**corpus analysis**

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td>word statistics &amp; relationships</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exploration</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Close Reading</th>
<th>Distant Reading</th>
<th>Combining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 38 4 5 9</td>
<td>26 20 18 18 41 9</td>
<td>5 8 25</td>
</tr>
</tbody>
</table>

Table 5: Applied close and distant reading techniques according to text analysis tasks.
and to the story flow analysis of a given source text. Miscellaneous tasks, for example, support the thorough analysis [KJW14] or enhance the close reading [GWF14] of a literary work.

The final category are corpus analysis tasks. Usually, distant reading visualizations are used to explore text corpora containing a high number of texts. A major research task is the analysis of word statistics & relationships among them. Further interests concern text similarity between the texts of a corpus, others require platforms for corpus exploration to facilitate knowledge discovery.

8. Applied Visualization Techniques

This section provides an overview of close and distant reading visualizations examined in the papers in our collection. In addition, we outline strategies for combining close and distant reading visualizations that facilitate a multifaceted analysis of the underlying textual data. Table 5 shows a distribution of the applied techniques according to the text analysis task taxonomy (see Table 4). For some research tasks, favored visualization methods stand out. A detailed overview about what techniques are suited for which text analysis tasks is given below.

8.1. Close Reading Techniques

A visualization that allows to close read a text requires that the structure of the text be retained in order to facilitate a smooth analysis. With additional information in the form of manual annotations or of automatically processed features of textual entities or relationships among them, a plain text can be transformed into a comprehensive knowledge source.

As can be seen in Table 5, the application of close reading techniques is particularly important when analyzing similar patterns. For text edition comparison, close reading is necessary to discover occurring similarities and differences, and a close reading of similar linguistic patterns or text re-use patterns helps to analyze the contexts in which these patterns were used. When focusing a text of interest, close reading techniques are applied to illustrate various text features. Other research tasks concerning named entities, topics and corpus analysis rather investigate generic features of a corpus, and apply mostly distant reading techniques. Still, close reading is sometimes helpful to connect a computationally gained distant view with the underlying source texts.

While ten visualizations provide only plain close reading views without additional information, 56 visualizations attend to the matter of enhancing the close reading capabilities of the humanities scholars. To visualize such additional information for a great variety of purposes, the researchers made use of the techniques listed below.

Color is the visual attribute most often used to display the features of textual entities and it is applied in different ways. In most cases, a colored background is used to express various types of information about a single word or an entire phrase (Figure 2). The tool Serendip [AKV14] varies the transparency of background colors to encode the importance of individual words (Figure 6a). Font color is also frequently used for this purpose (Figure 6b left). Colored circumcircles (Figure 6c) around words are used only once [MLCM16]. When displaying digital editions of literary texts, insertions are underlined. This might be the reason that this metaphor of underlining words is also rarely used to enhance close reading [CWG11]. Overall, coloring is a suitable method to express a great variety of textual features. Among other purposes, coloring is used for the analysis of similar patterns, e.g., to mark common words (e.g., [JRS09, Mur11]) and aligned text segments (e.g., [ZNMS15, RPSF15]) in parallel texts, or when exploring a text of interest, e.g., to highlight

(a) Colored backgrounds and backgrounds with varying transparency (Figure provided by Alexander et al. and based on [AKV14]).

(b) PRISM uses color to highlight the classification of words and font size to encode the number of annotations (Figures under CC BY 3.0 license based on [WMN14]).

(c) Circumcircles in the Poem View (left) and connections in the Path View (right) highlight rhyme sets in poems (Figure produced with Poemage [Poe16] based on McCurdy et al. [MLCM16]).
the automated or manual classification of words or phrases (e.g., [KJW*14, WMN*14]), or to visualize various sound patterns in poems (e.g., [CTA*13, Ben14]).

Font size is another method of visualizing features of textual entities. Adopted from tag cloud design [VWF09], this metaphor serves best to highlight the significance or weight of a textual entity in relation to the given text or corpus. In the design of a variant graph [JGF*15, JG15], which is a directed acyclic graph that is used for text edition comparison as it visualizes differences and similarities among text variants, font size encodes the number of occurrences of a word in all editions (Figure 9a). Within the web-based tool PRISM [WMN*14], users collaboratively group the words of literary texts into different categories. The collected statistics are used to display the number of annotations of each word by variable font size (Figure 6b right). In [CWG11], varying font size is used to visualize the importance of text passages according to the user’s preferences.

Glyphs attached to individual textual entities are convenient techniques to visualize abstract annotations that are hardly expressible with plain coloring or varying font size. All examples we found enhance the close reading of a text of interest, mostly poems. In [ARLC*13], phonetic units are drawn atop each word using color to classify phonetic types (Figure 7). Additionally, pictograms illustrate phonetic features. The Myopia Poetry Visualization tool uses rectangular blocks to visualize poetic feet and the spoken length of syllables [CGM*12]. For the visualization of a poem’s hermeneutic structure, Piez deploys glyphs in the form of rectangular and circular maps [Pie10]. An example is given in Figure 8. Goffin explores the placement and design of so-called word-scale visualizations, which are small glyphs enriching the base text with additional information [GWFI14]. For example, the background color of words contained in digital copies speaks for OCR certainty. Furthermore, small interactive bar charts illustrate variants of observed words.

Connections aid to illustrate the structure among textual entities most often applied to support text analysis tasks concerning similar patterns. One usage of connections in close reading is to highlight subsequent words in a variant graph to track variation among text editions [BGHE10]. As shown in Figure 9a, colored links can help to identify certain editions [JG15, JGF*15]. Other approaches juxtapose the texts of various editions and visually link related text passages [WJ13b, HKTK14, JGBS14], as instantiated in Figure 9b. Furthermore, connections can also be used to vi-

![Figure 7: Poem Viewer uses glyphs to encode phonetic units, and connections show phonetic and semantic relationships (Figure reproduced with permission from Abdul-Rahman et al. [ARLC*13]).](image)

![Figure 8: Close reading example with glyphs illustrating the hermeneutic structure of a poem (Figure provided by Piez and based on [Pie10]).](image)

![Figure 9: Connections used for text alignment.](image)
ualize sentence structure [KZ14]. Two close reading visualizations use connections for the analysis of sound in poems. While Abdul-Rahman [ARLC*13] illustrates phonetic and semantic relations within poems (Figure 7), McCurdy [MLCM16] draws paths between words of a poem sharing the same tones (assonances) to highlight sonic patterns (Figure 6c).

8.2. Distant Reading Techniques

A visualization that displays summarized information of the given text corpus facilitates distant reading. The process of transforming such information into complex representations can be based upon a large variety of data dimensions, e.g., various types of metadata of textual entities, automatically processed or manually retrieved relationships between textual elements, or quantitative and qualitative statistics about unstructured textual contents.

An overview of applied distant reading visualizations according to the text analysis task taxonomy is given in Table 5. The overall usage of such techniques suggests their importance for nearly all text analysis tasks in digital humanities, even when the close reading of a text is more important, e.g., when focusing a text of interest or analyzing similar patterns.

Within our research papers collection we found 132 visualizations providing a distant reading view of a given text corpus. We extracted and grouped various approaches found to visualize summarized information into the six following categories.

Heat maps or block matrices are often used to highlight text snippets, especially, when analyzing similar patterns and in corpus analysis tasks. Thereby, a heat map may reflect structural elements of a text [JRS*09, VCPK09] or the structure of an entire corpus [CDP*07, Mur11, BGHJ*14]. In such scenarios, the coloring of rectangular blocks helps to analyze the distribution of specific textual patterns [CWG11, MH13, JKH*15]. Another example is the usage of heat maps to show relationships among various texts in a corpus. The similarity for each tuple of texts within the corpus can be determined by counting similar text passages, and the result can be visualized as a heat map [GCL*13, FKT14], e.g., to highlight the similarity between Shakespearean plays [RRRG05]. Heat maps are also applied to visualize similarities or differences among text editions [JG15, PMMR15], or to highlight re-used passages between the texts of a corpus [JGBS14, RARC*15, ZNMS15]. For the analysis of potentially plagiarized texts, so called Diffines reveal structural differences between several suspicious text fragments and their alleged originals in a Focus+Context view [RPSF15], an example is shown in Figure 10. A further heat map variant are fingerprinting techniques as introduced in [KO07] in order to visualize characteristic textual features of literary works. In text analysis tasks concerning named entities, heat maps can be used to analyze places mentioned in texts [AGZH15], or to reveal interpersonal relationships between characters in prose literature [OKK13]. For the analysis of topics, Alexander et al. [AKV*14] propose two matrix representations. The RankViewer illustrates the ranking of words belonging to topics and the CorpusViewer shows relations to certain topics for each document of a corpus. Heat maps are also used in [MSR*15] to display “high-level summaries” of topic modeling results. Finally, heat maps are used to analyze a text of interest [KJW*14], e.g., to visualize the similarity [CTA*13] or the flow [FS11, Ben14] of sound in poems.

Tag clouds are intuitive visualizations to encode the frequency of words within a selected section, a whole document or an entire text corpus by using variable font size.

Figure 10: Heat map highlighting potential plagiarized text passages in a PhD thesis (Figure reproduced with permission from Riemann et al. [RPSF15]).

Figure 11: Comparing the tags in five Shakespeare plays (Figure based on [JBR*15]).
Tag clouds are therefore a suitable method for corpus analysis tasks [VCPK09, Bea12, FKT14, GTAHS15]. TagPies – a tag cloud arranged in a pie chart manner – support the comparative analysis of the co-occurrences of search terms [JBR*15], an example of which is shown in Figure 11. Other approaches visualize the temporal evolution of tags in tag clouds, either listing tags per time period [CVW09], or by attaching a time graph to each tag [LRKC10]. Beaven uses tag clouds to illustrate collocational relationships of a single word [Bea08] and to compare the collocates between two words [Bea11]. Tag clouds are also applied when analyzing topics, e.g., by displaying a topic’s characteristic tags [BJ14, ESK14, JOL*15, MSR*15] (an example can be seen Figure 16b), or by summarizing the major tags for certain time periods [CLT*11, CLWW14]. The usage of tag clouds to explore the classification of speculative fiction anthologies [HFM16] is shown in Figure 13. Tag clouds are rarely applied for the analysis of named entities [HAC*15] (see Figure 12a) or when focusing a text of interest [KJW*14]. In some of the above mentioned works, tag coloring is used to express additional information such as the temporal evolution of a word’s significance or the classification of tags.

Maps are widely used to display the geospatial information contained in a text. Most often, maps support the analysis of named entities extracted from a text or an entire corpus. Two works illustrate the geographical areas which are associated to persons [Wil15b], e.g., by mapping the places of activity of musicians manually extracted from musicological literature in order to support the geospatial comparison of musicians’ activity regions [IFS16]. But usually, places mentioned in texts are analyzed. With the help of contemporary (e.g., GeoNames [Geo15]) and historical gazetteers (e.g., Pleiades [Ple15]), the extracted placenames can be enriched with geographical coordinates, and their visualization on a map supports the analysis of the (fictional) geographic space described in the source text(s). Some approaches use thematic [ÖML14] or density maps [GH11b, BB15b] for this purpose, but the usage of glyphs in the form of circles is more frequent [Tra09, DWS*12, HAC*15, Wil15a] as it simplifies the interaction with individual plotted places (e.g., see Figure 12a). In [HAHT*15], circles are used to map discrete London places occurring in fictional literature, and polygons represent wider spaces such as neighborhoods or districts of London. The coloring of shapes indicates collected emotions to these places (see Figure 12b). In [JW13], various glyphs encode various types of places occurring in medieval texts. Two works that focus on mapping the geographical knowledge of ancient Greek authors draw connections between glyphs to illustrate travel routes [EJ14] or to highlight the strength of the relationship between placenames, which is reflected by the number of co-occurrences [BPBI10]. In contrast to the previous works, the geospatial metadata associated with individual corpus texts (text creation timestamp) can be used for mapping [MBL*06]. The visualization of Faulkner’s fictional Yoknapatawpha County includes various means of geographic mapping [DNCM14]: on the one hand, the imagined geography and, on the other, the placenames displayed on the geographic levels region, nation and world (Figure 12c). In addition to named entity analysis, maps are used for the analysis of topics [DFM*08, GDMF*14] and corpus exploration [JHSS12].
Timelines are appropriate techniques to visualize historical text corpora carrying various types of temporal information. Often, timelines support the analysis of named entities. In [JW13], uncertain datings of medieval texts are visualized to support the analysis of cited places. Sometimes, the temporal information about events reported in a text need to be extracted in order to visualize (fictional) calendars [DNCM14, GDMF14, ÖML14, HAC15] (e.g., see Figure 12a). For the exploration of placenames in Herodotus’ Histories, a timeline is used to show where certain placenames occur in the text [BPBI10]. Timelines also support the temporal analysis of topics [HFM16], e.g., the exploration of events in news articles [ESK14] (see Figure 16b). Furthermore, timelines are used for corpus analysis tasks [JHSS12]. A somewhat abstract timeline view is shown in [HSC08]. Here, a so-called tree cut section, whereby each ring represents a decade, visualizes statements from and about Emily Carr’s life and work (Figure 14). Streamgraphs are popular techniques that produce aesthetic visualizations and allow to track the evolution of topics over time [BW08], thus generating enhanced versions of the timeline metaphor. Such visualizations are often based on newspaper sources [CLT11, KBK11, DWS12, CLWW14] or political text archives [Kau15, Poi15] to support the analysis of contemporary topic changes. Using a research paper pool [ARR12], the changing importance of research topics can be explored. Streamgraphs may also be used to support the analysis of storylines in a text of interest to illustrate plot evolution and changing locations [LWW13]. Based upon Hollywood screenplays, the tool ScripThreads visualizes action lines of movie characters [HPR14].

Graphs are valuable visualizations for all text analysis tasks. They are most often applied to visualize certain structural features of a text corpus. A common usage is the visualization of relationships between the texts (represented as nodes) of a corpus in the form of a tree [HFM16] (see Figure 13) or a network [WH11]. Proximity can be used to express the similarity of these texts [EX10, MRMK15], e.g., based upon stylistic closeness [Joc12, CEJ14]. For instance, Eder visualizes a network of poetic texts based on stylistic features [Ede14], thereby highlighting connected editions and sequels (Figure 15a). Graphs can also be used to illustrate topics (as nodes) and proximities of topic models applied to text corpora [Kau15]. In other works, nodes represent the words of a corpus and links are drawn to reflect semantic relationships [AGL07, RFH14, OGH15] or co-occurrences [VCPK09, KKL11, WJ13a, BB15b].
In [HAHB15], a graph visualizes cross references in historic encyclopedia by linking related entries. Further applications are the visualization of scene changes and character movements in Shakespearean plays [RRED05], as well as the display of conceptual [Arm14], contextual [HSC08] or multilingual [GZ12] information. Phrase nets connect textual entities that appear in the form of a user-specified relation (syntactic or lexical) [vHWV09]. All aforementioned works apply force-directed algorithms for the placement of nodes. Radial graphs can be used to unveil the relationships of words within poems [MFM13], or, again, to highlight the similarity among texts and, in this case, as nodes radially grouped by authors [Wol13]. The Word Tree [WV08], also used in [MH13], visualizes sentences sharing the same beginning in the form of a tree. In contrast to the variant graph, a technique that supports close reading of textual editions, the Word Tree is a distant reading technique as it dissolves the order of sentences. Finally, we found a method that visualizes plain event trigrams extracted through phrase mining algorithms and thus providing metaphors to display uncertain information [MLSU13]. When analyzing named entities, graphs are the means of choice to visualize the relationships between people in the form of social networks. Such representations are widely applied in the digital humanities to illustrate the relationships between characters in literary texts [CSV08, Tót13, BF05, TFK15]. In these graphs, the size of a node can be used to encode the frequency of a character name in the text [BHH11, Pie14], the thickness of an edge [Cob05] (Figure 15b) or the proximity of the nodes [Pie15, JFS16] can serve to reflect the strength of a relationship, and edge style can be used to classify the type of relationship [KOTM13]. As per the aforementioned works, Kochtchi uses a force-based graph layout to visualize social networks automatically extracted from newspaper articles [KLB14]. In contrast, radial layouts and parallel coordinates are used in [Boo13]. For the visualization of Thomas Jefferson’s social relationships (Figure 15c), the nodes placed on a vertical axis are connected with arcs [Kle12]. Riche proposed a layout for Euler diagrams, which can also be utilized to visualize relationships between characters extracted from Shakespearean texts [RD10]. Finally, GeneaQuilts smartly visualizes large genealogies extracted from literary texts such as the Bible [BDF10].

Miscellaneous methods also produce beneficial results for certain text analysis tasks, most often for the exploration of similar patterns. In [JGBS14], an interactive dot plot interface is used to visualize and explore patterns of text re-use between two texts (Figure 16a). In [GCL07], parallel coordinates and a dot plot view, which is used for filtering purposes, visualizes the similarity of parallel text sections. Sankey diagrams are used to compare the categories of words contained in two books [HCC14], and to highlight plagiarized text passages when juxtaposing a PhD thesis to potential sources [RPSF15]. For the analysis of repetitions in Gertrude Stein’s The Making of the Americans [CDP07], parallel coordinates visualize the frequency of phrases across sections, and TextArc [Pal02] is used to explore the repetition of individual words. Two miscellaneous methods are applied to analyze topics. For the exploratory thematic analysis of historical newspaper archives [ESK14], an application of the dust-and-magnet metaphor [YMSJ05] yielded useful results (Figure 16b). Another topical analysis technique uses a landscape metaphor to visualize the topology-based clustering of articles taken from the New York Times Corpus [OST10] (Figure 16c). Various methods were also developed to support the analysis of a text of interest. The tool PlotVis allows users to model and interact with XML-encoded literary narratives in 3D [PB14]. A further complex tool named “Simulated Environment for Theatre (SET)” supports the story flow simulation of theatrical plays [RSDCD13]. It consists of various 2D interfaces.
illustrating the “line of action” and a 3D interface populated by character avatars. For the analysis of word statistics & relationships, tree maps are used to illustrate the occurrences of adjectives in fairy tales in [WJ13a]. The Column Explorer introduced in [JFS16] supports the analysis of named entities, in that case by comparatively visualizing biographical profiles of musicians.

8.3. Techniques for Combining Close and Distant Reading

Most of the visualizations we find provide either a close or a distant reading of a text corpus. Still, an important feature for literary scholars when working with distant reading visualizations is direct access to source texts or, in other words, close reading. Among the papers in our collection providing close and distant reading, some visualizations combine both techniques – most often in the form of coordinated views [WBBK00]. We do not consider the methods in [WH11,CTA*13,Ben14,BJ14] as the presented visualizations for close and distant reading serve different purposes, and are not connected to one another. Table 5 orders the remaining 37 remaining techniques, which are outlined in detail below, according to the given text analysis tasks in three groups.

**Bottom-up** methods focus primarily on close reading, especially, when focusing similar patterns. In [GCL*13], the user selects a desired text passage in Shakespeare’s Othello, which is shown in various German translations. Distant reading visualizations are processed (parallel coordinates view, dot plot view, heat map) based on that selection. Another bottom-up approach supports the semi-automatic alignment of early new high German text variants [MRMK15]. A graph displaying the similarities between text editions is updated as annotations are collected in close reading sessions. In [Mur11], the literary scholar selects a certain phrase during the close reading process. Next, that phrase is searched within the text corpus and the phrase’s distribution is shown in the form of a heat map. Two approaches provide bottom-up strategies to support the analysis of named entities. When annotating literary texts [AGZH15], places related to Edinburgh are marked, and a linked heat map that displays the distribution of all annotations is accordingly updated. In [OGH15], the user explores automatically tagged named entities of scientific papers in close reading mode. After editing, a graph reflecting contained entities and relationships among them is generated.

**Top-down & bottom-up** approaches taken within one visualization entity allow for switching between close and distant reading while taking into account manipulations of the preceding view. Some of these approaches support the analysis of similar patterns. In [JRS*09] and [PMMR15], the user can switch between heat map (distant reading) and text view (close reading). A side-by-side navigation between source text (close reading) and a distant reading graph showing the relationships among textual entities is illustrated in [WV08,RFH14]. Here, textual entities can be selected in both the graph and the text, triggering mutual updates. Other text analysis tasks also benefit from the combination of top-down and bottom-up approaches. A typical use case are visual analytics methods. The Varifocal Reader [KJW*14] hierarchically visualizes a document with the help of distant views (structural overview, tag clouds) and close reading techniques (use of color, digital copy), thus supporting hierarchical navigation. In close reading mode, automatically acquired classifications of textual entities can be manually modified, which subsequently affects distant views. The same applies to social networks automatically extracted from newspaper articles [KLB14]. The user browses the graph, opens close reading views associated with individual nodes and annotates the source text, which, again, affects the distant view and is used for classifier training. WordSeer [MH13] allows for a multifaceted perusal of a text corpus. For selected textual entities, several close and distant reading views can be used to browse the corresponding source texts. Within the close reading views, the user can group words into classes, which can then be used as a starting point for text corpus analysis.

**Top-down** strategies support nearly all types of text analysis tasks. They are mostly applied to combine close and distant reading visualizations. Such methods implement the Information Seeking Mantra in its original meaning. Initially, a distant view on the textual data is shown, the user can often manipulate the visualization by means of filtering and zooming, and finally retrieve the details-on-demand by clicking on a potentially interesting data item. In some cases, the texts are simply shown at the end of the information seeking pipeline [HSC08,DWS*12,MFM13,RSDCD*13,FKT14,Wil15a,Wil15b,HFM16]. Observed words or text patterns are often highlighted in the close reading view by way of coloring [VCPK09,GZ12,Will13,AKV*14,HPR14,HAC*15,JBR*15,KH*15]. Various colors can thereby illustrate word categories [CDP*07], e.g., types of toponyms in the Herodotus Timemap [PPB10] or topological cluster information [OST*10]. In some systems, close reading is more closely related to the preceding distant reading. In [BGHJ*14], the connection between close and distant reading is achieved by zooming. The distant view, a structural overview, highlights certain patterns, and zooming allows the close reading of individual passages. In [JGBS14], a grid-based heat map visualizes similarities between the texts of a corpus, and clicking on a grid cell opens a close reading view showing the corresponding two texts juxtaposed with connections between related text passages. Similarly, the navigation between distant plagiarism overviews and the close reading of plagiarized passages is organized in [ZNMS15] and [RPSF15]. A distant reading visualization illustrating the variance of verses among multiple Bible editions provides distant views as heat maps on various text hierarchy levels (entire Bible, book, chapter) [JG15].
chapter view, the close reading of individual verses is possible. The CorpusSeparator presented in [CWG11] is a distant view used to generate a weighted tag list (dependent on corpus statistics). Based upon these weights, the close reading view of a text (illustrated with Shakespeare’s A Midsummer Night’s Dream) is manipulated by coloring and sizing lines.

9. Collaboration Experiences

Within our collection, we examined papers about the research experiences reported by visualization researchers in order to provide suggestions that might help visualization scholars new to the field of digital humanities to develop successful visualizations. Some projects reveal valuable insights into collaboration experiences. Excellent design studies are outlined by Abdul-Rahman et al. [ARLC*13], McCurdy et al. [MLCM16] and Hinrichs et al. [HFM16]. All applications were successfully presented in visualization and digital humanities issues. Other publications also share important experiences. A collective overview of the gained insights regarding various aspects of the development phase are outlined below.

Project start. The beneficial, initial decision of carrying out a user-centered design study [Mun09] is reported in various works (e.g., [ARLC*13, JFS16]). This leads to a very close collaboration between researchers of the different fields, which helps to avoid gearing the development of a visualization into false directions. Further important tasks at the beginning of a digital humanities project are discussions about the research questions and perspectives for which a visualization, be it for close or distant reading, can be beneficial [JGBS14]. These discussions include the analysis of the data features [VCPK09] as well as the setup of regular project meetings to work on and extend a collaborative idea. A typical problem of digital humanities projects is reported in [MLCM16]. The “initial conversations [between visualization and humanities scholars]” were broad and open-ended,” also, because the humanities scholars “did not have specific goals” in mind. Furthermore, the humanities scholars were sceptical that visualization can support their research, and there was also an “anxiety that the computer would inhibit the qualitative experience of the poetic encounter.” After humanities scholars presented examples of interesting features and computer scientists “established methods for computationally detecting and analyzing the devices that most interested them,” a common project basis and tasks had been generated. In such circumstances, special workshops can also help computer scientists and humanities scholars get acquainted with each others’ tasks, mindsets and workflows [ARLC*13]. Abdul-Rahman reports the importance of visualization researchers participating in poetry readings and in-depth discussions with literary scholars to discover “a variety of interesting problems that might be subject to visualization solutions.” Also, a small corpus generated for literary scholars was helpful for Abdul-Rahman to examine research questions without the aid of existent visualizations. The generation of a text corpus is often an enduring humanities scholars’ task that begins with a project launch [HFM16]. As a consequence, visualization researchers start with a small training set, and should therefore design a visualization as flexible as possible in order to enable potential changes of humanities scholars’ research interests, and to avoid limitations. In the best case, the text corpus to be analyzed is already available in digital form and a precise research question is at hand, as outlined in [JFS16].

Iterative development of prototypes. The involvement of humanities scholars in various stages of the development is necessary to ensure creating an intuitive visualization that will be used. For example, regular face-to-face sessions between computer scientists and humanities scholars can help to identify problems and potential enhancements of the prototype design [JGBS14]. Such a session should be composed of a demonstration and trials of the visualization prototype as well as intense discussions in order to gather the levels of detail and complexity that a visualization should ideally reach [ARLC*13]. Geßner [JGF*15] stated that such a process finally helps to gain an intuitive result “even for the inexperienced, maybe sceptical user.” When designing a profiling system for musicians [JFS16], a frequent interdisciplinary get-together was important for the visualization researchers to communicate their own concerns and to iteratively redesign the underlying mathematical basis (similarity measures) – thereby ensuring that aspects of data transformation retained comprehensible for the collaborating musicians. That the scholarly exchange is of particular importance if the textual data source evolves throughout the project time, is outlined by Hinrichs et al. [HFM16]. On the one hand, the archival work of humanities scholars when working on the dataset may further develop hypotheses that trigger new visualization ideas, and on the other hand, a visualization has the potential of changing the humanities scholars’ research processes and their perspective on a text collection. For the development of Poemage [MLCM16], frequent meetings helped visualization researchers in understanding the problem space and engaged literary scholars in working with the visualization, and finally, in developing “an interface that reflected their interests, aesthetics, and values.” The authors also document that the departure from well established design principles such as regarding “ambiguity as a fundamental source of insight” or “not restricting the tool to avoid clutter” was necessary in raising the value of the visualization. Another example is given by scholars involved in the development of Neatline [NMG*13], which is based upon Omeka [Ome15], a content management system for online digital collections. The stepwise development of Neatline led to advancements of Omeka itself, thus benefiting a far wider audience than originally anticipated.

Evaluating visualizations with humanities scholars. The evaluation sessions provide important insights into design, intuitiveness, the utility of visualizations and into
Adapting existing visualization techniques. For some of the text analysis research questions posed in the digital humanities, the adoption of existing techniques proposed in visualization research papers is beneficial. A positive example is the Trading Consequences project [HAC15]. Involved visualization scholars designed a system inspired by VisGets [DCCW08] and made use of Parallel Tag Clouds [CVW09]. Both visualization techniques were not primarily developed for digital humanities data, but they were beneficially adapted to support humanities scholars. Occasionally, new techniques for close and distant reading are designed while appropriate, sophisticated visualizations unrelated to digital humanities data already exist. For future research tasks, the inclusion of these visualizations into the workflows of humanities scholars could lead to faster hypotheses generation due to the limited time for development. As an example, the Sequence Surveyor, which provides a dendrogram to explore genomic structures [ADG11], could support future research. Each leaf of the dendrogram shows a heat map illustrating genome distributions. This metaphor could be used to visualize both the rhyme structure of a poem in dendrogram form and the heat maps displaying phonetic patterns. Other possible adaptations of existing visualization techniques for digital humanities research can be found in the previous version of this survey [IFCS15].

Novel techniques for close reading. Various publications outline that close reading benefits from visualization, e.g., by highlighting crowdsourcing statistics [KG13, WMN14] or displaying information about textual features and structure [ARLC13, JGF15] alongside the source text. Although close reading is an essential task for humanities scholars, in most cases only simple visualization techniques, such as color coding textual entities, are provided. Few works attend to the matter of enhancing close reading in a beneficial manner. For example, the work on word scale visualizations is a promising technique [GWFH14] from which many humanities scholars may profit. But despite the proposed annotations of individual words with statistics or of country names with polygons, the concept needs to be expanded to annotating other kinds of named entities. For example, providing supplementary information about (1) acting persons and their relationships, (2) artifacts mentioned in texts, or (3) occurring references could be interesting features for humanities scholars. Future work in visualization should include the development of design methods to meet such use cases, and studies that measure the benefit of glyph based approaches for close reading in comparison to using color or font size to express certain text features.

Visualizing transpositions in parallel texts. When observing similarities and differences among various editions of a text, one focus is to detect transpositions of textual entities. Such transpositions may occur on various text hierarchy levels, e.g., changed word order, modified argumentation structures, or even when exchanging whole paragraphs or sections. Although suitable methods exist for the first two
hierarchy levels (words, sentences) [WJ13b, JGF*15], there are no visualization techniques capable of coherently visualizing transpositions on all hierarchy levels by combining means of close and distant reading.

**Geospatial uncertainty.** Many visualizations deal with placenames extracted from literary texts to illustrate the geographical knowledge of a particular era. Here, various mapping issues arise [JW13]. Texts may contain placenames of varying granularity (e.g., country, region, city) or type (e.g., points for cities, polygons for areas, polylines for rivers) or even fictional placenames, which are hard to represent. Furthermore, placenames can themselves carry uncertainty of varying degrees, e.g., the exact locations of “Sparta” and “Atlantis” have yet to be discovered. Another form of uncertainty is defined by contextual information, e.g., expressions like “in London” and “close to London” cover various geospatial ranges. The development of a design space providing solutions to visualize these various types of geospatial uncertainty is one of the current primary challenges in digital humanities. Such a design space could be built upon the ideas of MacEachren for visualizing geospatial uncertainty [MRO*12].

**Temporal uncertainty.** The visualization of temporal uncertainty is an equally important future task. Such uncertainties occur, for instance, when dating cultural objects, such as historical manuscripts [JW13, BESL14]. Temporal metadata, in fact, can be provided in multifarious manners, e.g., 1450, before 1450, after 1450, around 1450, 15th century, first half of the 15th century, etc. One can try to transform such temporal formats into machine-parsable time ranges, but the visualization of such uncertainties is a crucial issue as it comprises considerable risks of misinterpretation. Applying methods capable of visualizing temporal uncertainty as proposed by Slingsby [SDW11] can be a first step, but their utility for humanities applications needs to be investigated.

**Reconstructing workflows with visualization.** In two visualization papers, authors related situations where, during their conducted case studies, humanities scholars mentioned the importance of visualization features that emulate the scholar’s workflow. In [KJW*14], users liked the display of digital copies as this builds trust in the visualization. When working with genealogy visualizations [BDF*10], historians “insisted on redundant representation of gender ... that is consistent with their current practices.” Both situations illustrate the future challenge of inventing visualization techniques for digital humanities applications that the humanities scholar can easily adapt. An important task for the computer scientist is not only to incorporate a scholar’s workflow when designing the visualization, but to also communicate all aspects of data transformation, so that a scholar is able to generate trustworthy hypotheses. The importance of this issue is documented in [GO12].

**Usability studies.** Although the utility of most visualizations considered within our survey was illustrated by usage scenarios, we found little evidence about conducted usability studies to, for example, justify taken design decisions. The number of humanities scholars participating in such studies is potentially very small due to the multifarious research interests scholars may have on a large body of texts belonging to different eras and genres. Generating a user study format that caters for the interests of many different scholars is required to gain valuable insights into guidelines for designing visualizations for the digital humanities. When it comes to tool building, in fact, the digital humanities community poses interesting and complex challenges by virtue of its interdisciplinary nature. It embraces a wider range of disciplines, so the techniques it offers should address the larger scope. It also welcomes contrasting mindsets, methods and cultures. While sharing similar logical and analytical methods, computer scientists tend towards problem solving, humanities scholars towards knowledge acquisition and dissemination [Hen14]. No one community should operate in subservience to the other but together, complementing each others’ approaches. For these reasons and in this context, specialist terminology, assumptions and technical barriers should all be avoided. It is in this sense that tool usability should be understood not only as improved functionality or aesthetics but as a transparent guide to utility [GO12].

**Qualitative studies.** The number of projects that include visualization components as valuable means of text analysis indicates the potential of visualization to support digital humanities research. Some scholars suggest the role of visualization as providers of new perspectives on the texts that facilitate text comprehension and hypothesis generation. For example, humanities scholars involved in the development of the PoemViewer [ARLC*13] mentioned that “they would not likely look for insight from the tool itself ... they would look for enhanced poetic engagement, facilitated by visualization.” Hinrichs et al. [HFM16] state that “information visualizations ... are not a means to an end but a starting point to explore, interpret, and discuss literary collections.” Similarly, Sinclair [SRR13] argues that “a visualization that produces a single output for a given body of material is of limited usefulness; a visualization that provides many ways to interact with the data, viewed from different perspectives, is better; a visualization that contributes to new and emergent ways of understanding the material is best.” Comprehensive case studies that scientifically debate the actual influence and impact of visualization could further specify its role and further strengthen its value as part of humanities research.

11. Conclusion

Computer scientists and humanities scholars seemingly do not have many things in common. Although they share some methodologies, they are geared towards different goals. But the digital age created a platform that brings people from two research areas together: the digital humanities.
During our survey, we had the opportunity to take a look at various fascinating digital humanities projects proposing visualization techniques that support a number of text analysis tasks. We classified the papers providing visualizations for historical texts according to our proposed taxonomy for text analysis tasks in digital humanities, categorized applied close and distant reading techniques and analyzed methods of combining both views to allow for multifaceted data analyses. In the process, we derived insights into a research area that requires the design of intuitive interfaces, but visualizations for textual data as part of the cultural heritage are rarely published in the visualization community. Figure 17 shows the temporal distribution of the papers in our collection. The trend of related works published within the digital humanities reflects the increasing value of close and distant reading visualizations for text analysis tasks in the recent years. Until now, the visualization community did not notice or consider these needs. The reason for this may lie in the obstacles encountered in publishing application papers with a digital humanities background because the often demanded quantitative evaluations are hard to perform due to the usually limited number of collaborating humanities scholars.

To strike a balance between our discussed shortcomings, we listed future challenges to support humanities scholars’ tasks with close and distant reading. Developing solutions could provide beneficial contributions to both research fields. Furthermore, we outlined collaboration experiences reported by visualization researchers working in the field of digital humanities as a means of singling out the important ingredients for a successful project.

References


© 2016 The Author(s)

Computer Graphics Forum © 2016 The Eurographics Association and John Wiley & Sons Ltd.
J. Nänä, G. Franzini, M. F. Cheema & G. Scheuermann / Visual Text Analysis in Digital Humanities


[Jan01] JASINSKI J.: Rhetoric and Society: Sourcebook on


© 2016 The Author(s) Computer Graphics Forum © 2016 The Eurographics Association and John Wiley & Sons Ltd