Designing Close and Distant Reading Visualizations for Text Re-use

Stefan Jänicke[1], Thomas Efer[2], Marco Büchler[3] and Gerik Scheuermann[1]

[1]Image and Signal Processing Group, Leipzig University, Germany
[2]Natural Language Processing Group, Leipzig University, Germany
[3]Göttingen Centre for Digital Humanities, University of Göttingen, Germany
{stjaenicke,efer,scheuermann}@informatik.uni-leipzig.de, mbuechler@gcdh.de

Abstract. We present various visualizations for the Text Re-use found among texts of a collection to support answering a broad palette of research questions in the humanities. When juxtaposing all texts of a corpus in form of tuples, we propose the *Text Re-use Grid* as a distant reading method that emphasizes text tuples with systematic or repetitive Text Re-use. The *Text Re-use Browser* provides a closer look on the Text Re-use between the two texts of a tuple. Additionally, we present *Text Re-use Alignment Visualization*s to improve the readability of Text Variant Graphs that are used to compare various text editions to each other. Finally, we illustrate the benefit of the proposed visualizations with four usage scenarios for various topics in literary criticism.

Keywords: Text Re-use, Text Visualization, Text Variant Graph, Literary Criticism, Digital Humanities.

1 Introduction

The conscientious analyzation and interpretation of small text passages, so called *Close Reading*, is a major technique for researches in literary criticism. But the digital age with algorithms that automatically retrieve vast amounts of data expedite *Distant Reading* methods [17] that give the observer an impression about the data distribution. The Information Seeking Mantra "Overview first, zoom and filter, details-on-demand" [22] is accomplished, when distant reading views are interactively used to switch to close reading views. The general task is to provide a visualization that shows an overview of the data, so that patterns potentially interesting for the observer are salient. A drill down on these patterns for further exploration is the bridge between distant and close reading.

In this paper, we want to outline the process of designing close and distant reading visualizations for the Text Re-use determined between all texts of a given collection. *Text Re-use* is defined as the oral or the written reproduction of textual content and is roughly divided into two categories [3]. On the one hand, a text passage is re-used deliberately, like direct quotes and phrases like winged words and wisdom sayings. Translations of a text into other languages also count to this category and are called interlingual Text Re-use. A very popular form of

deliberate Text Re-use is plagiarism, which has gained major attention in the recent years, mainly driven by plagiarism allegations in politics. On the other hand, a Text Re-use may be unintended, like boilerplates, e-mail headers or the repetition of news agency texts when writing daily newspapers [6]. Further examples are idioms, battle cries and so called multi word units.

The analysis of Text Re-use among historic texts with the goal to explore known and discover unforeseen relationships in cultural heritage has become an important task within various Digital Humanities projects. Thereby, the humanities scholars are interested which texts share patterns of consecutive similar units (systematic Text Re-use) and how specific phrases are used (repetitive Text Re-use). Furthermore, the analysis of these similar text units regarding structure, context and used expressions is of special interest. This is also a substantial task in literary criticism, when various editions of a text are cautiously compared to each other.

This paper shows, how intuitive, interactive visualizations help humanities scholars in understanding and interpreting the Text Re-use occurrences. In particular, we present the following visualizations that support close and distant reading for Text Re-use:

- **Text Re-use Grid**: a chart that juxtaposes all texts of a collection in relation to amount and type of Text Re-use,
- **Text Re-use Browser**: a user interface consistent of an interactive Dot Plot View and a Text Re-use Reader that allows for the inspection and browsing through all Text Re-uses between two texts,
- **Text Re-use Alignment Visualization**: a visualization for aligned text units that improves the readability for *Text Variant Graphs*.

2 Related Work

The discovery of relationships between different texts and the alignment and visualization of the findings has been a challenging task in various works. Xanadu, founded in 1960, can be seen as one of the pioneer projects that attend to this matter [19]. The current prototype shows the dedicated text in the center of the screen, related texts are positioned on both sides and shared patterns are aligned and highlighted using various colors. John et. al [15] propose a focus and context approach for the visualization of texts sharing similar patterns. A vertical ribbon for each text shows the distribution of these patterns, and interactively, the user can drill down to regions of interest. Cheesman offers a visualization for the alignment of multilingual text passages in Shakespeare's Othello, where the user can interactively browse through the texts of two editions [5]. Likewise, related text entities are visually linked to each other. To illustrate the computationally determined Text Re-uses in ancient Greek texts, Büchler suggests a graph to show the results for a certain author by number, citing authors, years of citing authors and passages of the book [4]. Additionally, the user can inspect individual text snippets with highlighted re-used passages. For plotting Text Re-use

between Bible books [16], Lee uses a static Dot Plot View, which was originally designed for bioinformatics to compare two genome sequences to each other [11]. A single dot marks a correlation between the genomes and multiple dots form patterns that indicate similar genomic segments. Lee utilizes this approach to highlight patterns of systematic Text Re-use. Various visualization methods also exist to highlight plagiarized passages of a given source text [12, 20]. A complete overview of the whole text is given and each page, chapter or plagiarized text passage receives its own block. Coloring is used to show the amount of re-used text or to indicate potential sources.

All the above visualization techniques focus on displaying relationships between a limited number of texts. Mostly, a certain source text is given and its correlations to other texts can be analyzed. A comparative overview between all texts of an arbitrary text collection is not provided. For this purpose, we propose the *Text Re-use Grid* as a distant reading solution for Text Re-use (Section 4). Moreover, we present the *Text Re-use Browser* that supports close reading of the Text Re-use between two selected texts (Section 5).

Text Variant Graphs are data structures that represent various editions of a text [21]. CollateX is one of the standard tools in the Digital Humanities that computes a static directed acyclic graph with vertices showing the various text fragments and edges labeled with edition identifiers connecting subsequent text fragments [8]. The plain design makes it hard for the user to follow how an edition disseminates in the graph. Furthermore, the vertices do not reflect the amount of occurrences and synonyms are not properly aligned to each other. Although extensions for user-driven annotation and modification exist [1], there are only few works that attend to the matter of designing Text Variant Graphs. A visualization, which allows for weighted nodes is the *Word Tree* [25]. It cannot be directly applied to Text Variant Graphs, since it only aligns shared beginnings of sentences in form of a tree. Each variation splits a node of the tree into several leaves. The font size of a node label reflects the number of occurrences. A plain solution to align Text Re-use is given in [4]. The original text snippet is drawn as a main branch and variations of Text Re-use candidates are sub-branches with a certain color. This approach works fine for small examples with minor variations, but it fails for major differences, especially, when multiple Text Re-uses share the same sub-branches. A similar visualization for the uncertainty in lattice graphs also supports various sub-branches [7]. But merging of multiple nodes of the same kind is not provided, although the metaphor for uncertainty could be used for this purpose.

In Section 6, we propose the *Text Re-use Alignment Visualization* that utilizes some of the presented ideas with the goal to design a well readable layout for Text Variant Graphs.

3 Theoretical Basis of Text Re-use

Let $A_1, \ldots, A_n$ denote a corpus of n texts. After splitting each text into units (e.g. sentences), the Text Re-uses between all text unit tuples are determined.

Each detected Text Re-use $\{a_i, b_j\}$ consists of two corresponding Text Re-use units a_i (e.g. i-th sentence of text A) and b_j (e.g. j-th sentence of text B). The *Scoring value* $t(a_i, b_j)$ defines a weight for $\{a_i, b_j\}$ dependent on the text unit lengths of a_i and b_j and their *Re-use Overlap*, which is the proportion of matching to non-matching tokens. t is ranged in the interval $[0, 1]$; 0 means no similarity between the two units, 1 means that a_i and b_j are equal. The complete Text Re-use result list contains only relevant Text Re-uses above a certain threshold for t. A more detailed description of the underlying algorithms for Text Re-use detection and the computation of t can be found in Büchler's dissertation [3].

Researchers working with Text Re-use have various research questions that require a definition of the following Text Re-use types:

Systematic Text Re-use. The consecutive occurence of the same pattern of Text Re-use is of particular interest for researchers when comparing different texts to each other. Such type of Text Re-use could be an indication for plagiarism. For instance, the pattern $\{a_i, b_j\}, \{a_{i+1}, b_{j+1}\}, \{a_{i+2}, b_{j+2}\}$ is a *Systematic Text Re-use* of three consecutive phrases.

Repetitive Text Re-use. This type of Text Re-use appears, when the researcher is interested in analyzing a phrase that is frequently used in a certain text. The goals in this use case are to explore the contexts, in which a phrase appears as well as to what extent a specific phrase is spread in the text. *Repetitive Text Re-use* for a phrase a exists for a set of Text Re-use pairs in the form $\{a, b_1\}, \{a, b_2\}, \{a, b_3\}, \ldots$

Isolated Text Re-use. We classify a Text Re-use $\{a_i, b_j\}$ as isolated if it does not occur within a certain pattern, more precisely, if it is neither systematic nor repetitive. As systematic Text Re-use doesn't necessarily need to be consecutive in both texts due to potential insertions, deletions or changes in the ordering of the textual entities, we need to discriminate isolated from systematic Text Re-use. We define $\{a_i, b_j\}$ as isolated if there is no Text Re-use $\{a_u, b_v\}$ within a certain neighborhood ε, so that:

$$\varepsilon = \sqrt{\frac{|i-u| + |j-v|}{2}} < 10$$

Empirically, we determined 10 as the best value to separate systematic ($\varepsilon < 10$) from isolated Text Re-use ($\varepsilon \geq 10$).

4 Text Re-use Grid

The intention of this visualization is to give the researcher an overview of the Text Re-use distribution among all texts of a corpus. We transform the result of the Text Re-use detection algorithm into an intuitive, readable visual interface that immediately (1) reflects the amount of Text Re-uses between each pair of texts, and (2) provides evidence for the type of Text Re-use. For this purpose, we define three parameters:

1. Text Re-use Amount σ. σ is the number of Text Re-uses detected between two texts A and B.

2. Systematic Text Re-use Index λ. λ is an assessment for structures of systematic Text Re-use between two texts A and B with an ordered list of text units, so that $A = \{a_{first}, \ldots, a_i, \ldots, a_{last}\}$ and $B = \{b_{first}, \ldots, b_j, \ldots, b_{last}\}$. To detect these structures, we preliminary filter the Text Re-use results by removing all repetitive and isolated Text Re-uses. This filter process results in a decomposition of the remaining n Text Re-uses into m clusters $C = \{c_1, \ldots, c_h, \ldots, c_m\}$ containing more than one Text Re-use $\{a_i, b_j\}$ each. For each of these clusters c_h with $|c_h|$ Text Re-uses in total, we compute a correlation coefficient $\rho(c_h)$ as

$$\rho(c_h) = \frac{\sum\limits_{\{a_i,b_j\}\in c_h} (i - \bar{i_h})(j - \bar{j_h})}{\sqrt{\sum\limits_{\{a_i,b_j\}\in c_h} (i - \bar{i_h})^2 \sum\limits_{\{a_i,b_j\}\in c_h} (j - \bar{j_h})^2}}$$

$$\text{with} \qquad \bar{i_h} = \sum_{\{a_i,b_j\}\in c_h} \frac{i}{|c_h|} \qquad \text{and} \qquad \bar{j_h} = \sum_{\{a_i,b_j\}\in c_h} \frac{j}{|c_h|}$$

to estimate the strength of the linear relationship between the Text Re-uses in c_h. Finally, the Systematic Text Re-use Index is defined as:

$$\lambda = \sum_{h=0}^{m} \frac{|c_h|}{n} \rho(c_h)$$

λ ranges in the interval $[0, 1]$, whereas high values indicate that patterns of systematic Text Re-uses are contained.

3. Repetitive Text Re-use Index ω. ω is a measure for the amount of repetitive Text Re-use. Let N denote the number of Text Re-uses found between two texts A and B. To define ω, we remove each Text Re-use $\{a_i, b_j\}$, if both text units a_i and b_j occur only once within all Text Re-uses. Finally, we define ω in the interval $[0, 1]$ with the remaining n Text Re-uses as

$$\omega = \frac{n}{N}$$

Grid Visualization. For the visual mapping, we construct a grid with each cell representing the Text Re-uses found between two texts of a corpus. For each cell, we compute σ, λ and ω for the corresponding two texts. The cells are displayed in form of rectangles with bounds proportional to the lengths of the corresponding texts. Interactively, the user can change the display to equal-sized squares, so that even cells representing short texts are properly visible.

Because of the importance for the researchers to detect and analyze texts with extensive systematic or repetitive Text Re-use, we use a specific coloring for the grid cells, so that the type of Text Re-use (represented by λ and ω) and the amount of Text Re-use (σ) can be easily recognized. As the human's ability

to discriminate colors is limited, we chose a class based approach to compute a limited number of cell colors. As proposed by Slocum et. al, we use an optimal classification method [23] to group the cells into two sets of classes in dependency of σ, λ and ω. With the Jenks-Caspall-Algorithm [14] using reiterative cycling, we compute a configurable number of classes. We receive n classes $\alpha_1, \ldots, \alpha_n$ for the type of Text Re-use (systematic or repetitive), so that α_1 contains the cells with the smallest λ (or ω) and α_n contains the cells with the largest λ (or ω). Furthermore, we compute m classes $\beta_1, \ldots, \beta_m$ for the amount of Text Re-use with β_1 containing the cells with smallest σ and β_m containing the cells with the largest σ. We determine the color for a grid cell in the HSV color space dependent on these classes as follows:

$$H = 240 + \tfrac{i-1}{n-1} \cdot 120 \qquad S = \tfrac{j}{m} \cdot 100 \qquad V = 100$$

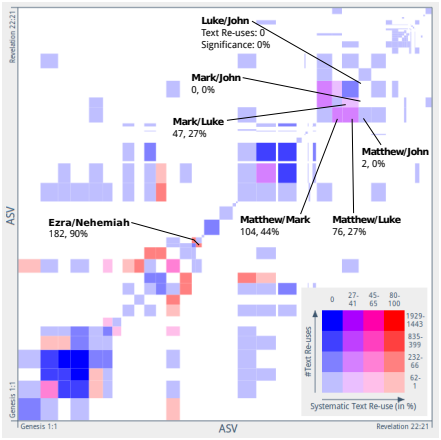
The type of Text Re-use defines the hue on the "Cold-Hot" color scale Diehl proposes [9] for the *EpoSee* tool from blue (cold) to red (hot). Thus, we receive cold hues for cell colors with less, and hot hues for cell colors with a lot of systematic (or repetitive) Text Re-use between the corresponding texts. The amount of Text Re-use defines the saturation, so that cells with a high number of Text Re-use receive highly saturated, and cells with less Text Re-uses lightly saturated colors. For the examples in this paper we used $n = m = 4$.

In Figure 1, the resultant Text Re-use Grids for the Bible books of the American Standard Version compared to each other highlighting systematic (Figure 1(a)) and repetitive Text Re-use (Figure 1(b)) can be seen. With the help of a legend, the user is able to immediately categorize type and amount of Text Re-use between two Bible books. Interactively, the user can change from highlighting systematic to highlighting repetitive Text Re-use. By mouse clicking onto a cell, the user has the ability to switch from the distant reading grid view to the closer Text Re-use Browser view that is explained in the next section.

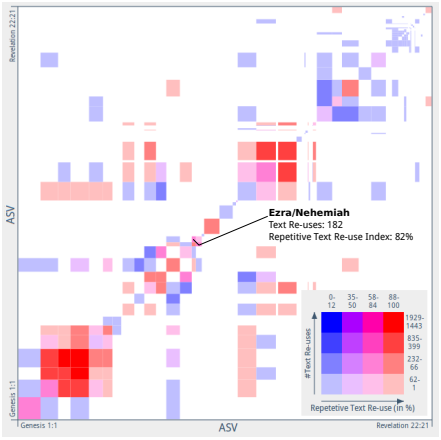
5 Text Re-use Browser

Whereas, the Text Re-use Grid allows for distant reading of all Text Re-uses occurring within a text collection, the Text Re-use Browser provides a closer look at the Text Re-uses found between two texts $A = \{a_{first}, \ldots, a_i, \ldots, a_{last}\}$ and $B = \{b_{first}, \ldots, b_j, \ldots, b_{last}\}$. This still complies to the characteristics of distant reading, but the Text Re-use Browser can be utilized to drill down to a limited set of Text Re-uses, which supports close reading. In particular, the Text Re-use Browser provides two panels for this purpose:

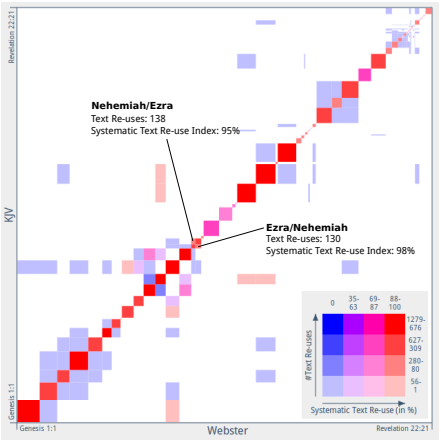
1. Dot Plot View. We also utilize the approach of a Dot Plot View to emphasize the types of Text Re-use between the given texts. In contrast to Lee [16], we provide an interactive chart, where the number $|A|$ of text units of A defines the range for the x-axis, and the number $|B|$ of text units of B defines the range for the y-axis. Each Text Re-use for a text unit pair is drawn as a single dot. As in bioinformatics, specific dot patterns indicate specific Text Re-use



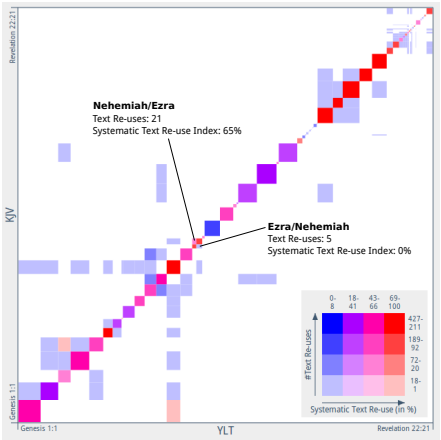
(a) ASV/ASV: Systematic Text Re-use



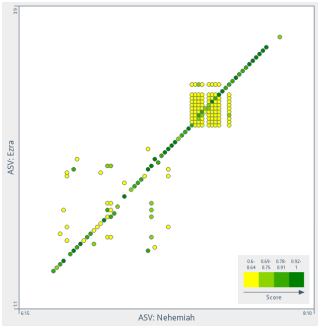
(b) ASV/ASV: Repetitive Text Re-use



(c) KJV/WBS: Systematic Text Re-use



(d) KJV/YLT: Systematic Text Re-use



(e) Dot Plot View

(f) Text Re-use Reader

Fig. 1. Text Re-use Grids for the juxtaposition of various Bible editions (Figures 1(a)-1(d)) and panels of the Text Re-use Browser showing systematic and repetitive Text Re-use patterns detected between the books *Ezra* and *Nehemiah* (Figures 1(e)-1(f)).

types. Diagonal patterns highlight sections that contain systematic Text Re-use, whereas vertical and horizontal patterns appear for phrase repetitions. In Figure 1(e) we detect patterns for both types of Text Re-use. By selecting a dot via mouse click, a popup with the corresponding text units and a Text Re-use Alignment Visualization (see Section 6) is shown. Interactively, the user is also able to zoom into a rectangular region of interest.

2. Text Re-use Reader. This panel allows for browsing A and B in two opposite windows. Whenever a re-used text unit appears in the viewport of one window, a connection to the opposite text unit is drawn. A click on a connection scrolls both texts, so that the text units of the corresponding Text Re-use are placed on the same horizontal level and a step-by-step exploration of consecutive Text Re-use is possible. A mouseover highlights matching tokens in both units. An additional overview for the texts gives an impression about all occurring Text Re-uses, and can be utilized to directly jump to a dedicated position. In both views, an accumulation of parallel lines is an indication for systematic Text Re-use, and *hub*s (a single unit of one text that is connected to multiple units of the opposite text) occur for repetitive Text Re-use. These features can be seen in Figure 1(f).

Both panels are linked to each other. A dot selection in the Dot Plot View triggers a scrolling of the texts to the corresponding positions, whereas a connection selection in the Text Re-use Reader opens the popup for the corresponding dot. For coloring the Text Re-use glyphs (dots, connections), we use again a class based approach. We group the Text Re-uses in dependency of their scoring value t into p classes $\gamma_1, \ldots, \gamma_p$, so that γ_1 contains Text Re-uses with the smallest t, and γ_p these ones with the largest t. In order to avoid misinterpretations, we chose a different color scheme in comparison to the Text Re-use Grid. The glyph colors are defined in the HSV color space as:

$$H = 60 + \tfrac{k-1}{p-1} \cdot 60 \qquad S = 100 \qquad V = 100 - \tfrac{k-1}{p-1} \cdot 50$$

Thus, the hue of a glyph color for a Text Re-use with class γ_k ($1 \leq k \leq p$) ranges from yellow to green. To gain visually distinctive colors, the color value ranges between 100 and 50. For the examples in this paper we used $p = 4$.

Some text juxtapositions contain lots of Text Re-uses that form various patterns. To visually filter for specific Text Re-uses, we allow to hide glyphs of repetitive or systematic Text Re-use. Additionally, Text Re-uses with low scoring values can be hidden and a slider can be used to hide isolated Text Re-uses without adjacent Text Re-uses in a certain neighborhood.

6 Text Re-use Alignment Visualization

One of the substantial tasks in the field of textual criticism is called collation, which is the cautious comparison of various editions of a text. Since it is an extremely laborious approach for humanities scholars to do this manually, few projects investigate methods that compute alignments for the digitized texts

and visualize the resulting directed acyclic Text Variant Graph in a simple manner (see Section 2). In this section, we present an intuitive, for the humanities scholars easily readable and comprehensible layout for Text Variant Graphs.

A text edition is a specific kind of Text Re-use as it was derived from a specific source text, but also occurrences of repetitive Text Re-use can be aligned and visualized with this approach. Let $\{e^1, \ldots, e^n\}$ denote a set of editions for a re-used text unit. Various alignment algortihms exist (e.g. [2, 18]), we use a brute force approach that performs well for small text units. After tokenization and normalization, we insert each edition $e^i = e^i_1 \ldots e^i_{|e^i|}$ in form of a directed path $v(e^i_1) \ldots v(e^i_{|e^i|})$ with vertices representing tokens in the initial Text Variant Graph. Then, we iteratively merge vertices of different paths with equal tokens and choose this alignment that reaches a maximum number of merge iterations while keeping the Text Variant Graph acyclic. Each vertex $v = \{e^i_s, e^j_t, e^k_u, \ldots\}$ of the graph is an alignment of the tokens $\{e^i_s, e^j_t, e^k_u, \ldots\}$. The *token degree* $|v|$ is the number of tokens assigned to v and $v(e^i_s)$ is the corresponding vertex for the s-th token of edition i. Figure 2(b) shows such a graph for five editions of the first Bible verse (Figure 2(a)).

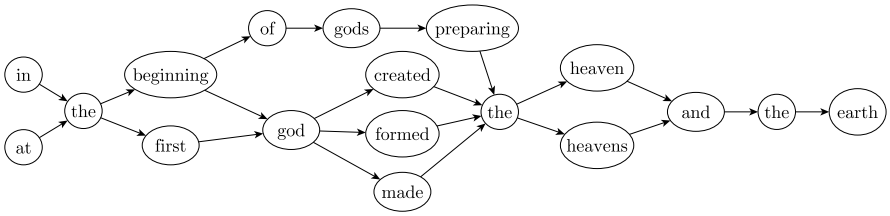
Graph Visualization. The token of a vertex is used for labeling. As Wattenberg proposes [25], we use font size as a metaphor to reflect the number of occurrences of a token. We layout the vertices by placing the corresponding labels onto horizontal layers. The height of a layer depends on the maximum height of the labels placed on it. We start by placing the labels for the vertices $v(e^i_1), \ldots, v(e^i_{|e^i|})$ of an arbitrary edition e^i in left-to-right order on layer 0 (main branch). By default, we choose the edition e^i with the maximum value for

$$\sum_{s=1}^{|e^i|} |v(e^i_s)|$$

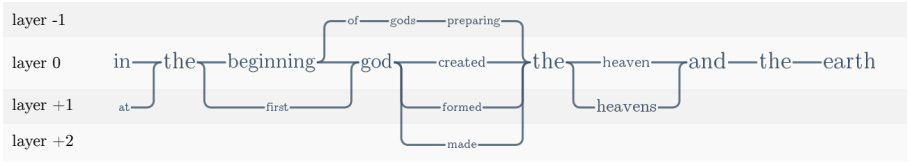
which means e^i has lots of tokens assigned to vertices with large token degrees. Afterwards, we iteratively determine the subpaths of the edition with most vertices already assigned to layers. Each subpath $\{v_1, \ldots, v_n\}$ has assigned layers for v_1 and v_n and the layer for the vertices of $p = \{v_2, \ldots, v_{n-1}\}$ needs to be determined. Let i denote the layer of v_1 and j the layer of v_n. We aim to place p as close as possible to its adjacent vertices v_1 and v_n. Starting with layer $k = \lfloor (i + j)/2 \rfloor$, we iteratively search for a layer with enough free space for the labels of the vertices of p in the order $k, k + 1, k - 1, k + 2, k - 2$, and so on. If the total width of the labels of p is larger than the space between v_1 and v_n, we stretch the distance between v_1 and v_n. After the proper layer is found, we move all vertices of the graph horizontally, so that (1) the labels do not overlap each other, (2) a minimum space of configurable width between all adjacent vertices is given, and (3) each vertex is placed in the barycenter of its neighbors. We perform this process for all subpaths of all editions to complete the layout for the Text Variant Graph. We draw undirected edges (for the user the direction is obvious) between adjacent vertices of the same layer in form of horizontal lines. To ensure a good readability of the graph, we use horizontal and vertical

King James Version (KJV) In the beginning God created the heaven and the earth.
American Standard Version (ASV) In the beginning God created the heavens and the earth.
Bible in Basic English (BBE) At the first God made the heaven and the earth.
Young's Literal Translation (YLT) In the beginning of God's preparing the heavens and the earth –
Smith's Literal Translation (SLT) In the beginning God formed the heavens and the earth.

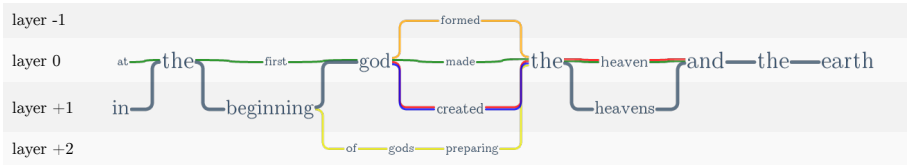
(a) Five English translations of *Genesis 1:1*



(b) Resultant Text Variant Graph



(c) Visualization using the KJV as main branch (layer 0)



(d) Visualization using the BBE as main branch (layer 0)

Fig. 2. Text Re-use Alignment Visualizations for five editions of *Genesis 1:1*.

links connected with bends for edges between vertices of different layers. More details about the edge routing algorithm can be found in [13]. The edges of the Text Variant Graph are drawn in gray, and to identify different edition flows, we use colors of the 12-color palette for categorial usage suggested by Ware [24] to facilitate maximal visual differentiation by the user.

We provide several means of interaction for the analysis of the resultant alignment visualization. The user is able to decide between various methods to display the edges of the Text Variant Graph. This includes an edge overview like shown for the resultant Text Re-use Alignment Visualization for the five editions of the first Bible verse in Figure 2(c), and a display of all edges drawn in the corresponding edition colors (e.g. Figure 6). Furthermore, thick grey majority edges that are passed by a minimum number of editions can be drawn (Figure 2(d)), so that only varying paths are highlighted. Hovering a token removes all edges from editions not passing it and selecting a token via mouse click shows the information about its corresponding editions in a popup window (Figure 6(b)).

7 Usage Scenarios

The Text Re-use Visualizations presented in this paper are utilized in various Digital Humanities projects. In this section, we take a look at four usage scenarios from these projects.

7.1 Various English Translations of the Bible

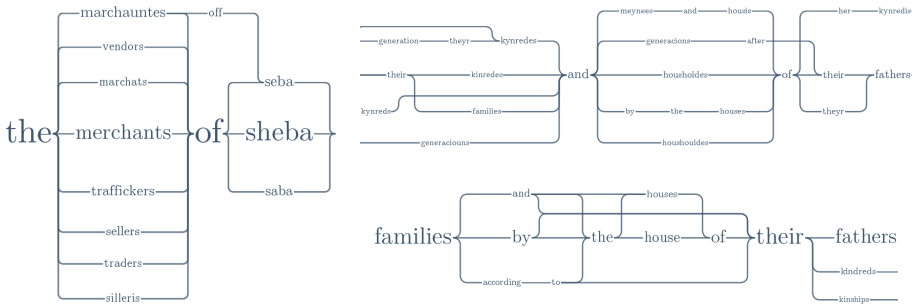
The interdisciplinary Digital Humanities project *eTRACES*[1] wants to discover, analyze and evaluate intertextual similarities in form of Text Re-use among historical texts of a given corpus. Since the Bible is known as one of the most often read and studied books, and therefore, potential findings are easily evaluable, it was chosen as a proof of concept for the project. The text corpus contains 23 different English translations of the Bible covering a time period from the 14th (Wycliffe Bible) to the 21th century (World English Bible). Since each translation was driven by a specific motivation, the involved humanities scholars had a great variety of research questions to be answered. In this section, we present some of their findings.

Of particular interest for the humanities scholars was the comparison of Bible books of the same edition regarding systematic Text Re-use. The Text Re-use Grid shows for the three evangelists *Matthew*, *Mark* and *Luke* strong interdependencies, whilst *John* has few or no Text Re-use at all with those three – confirming a well known fact by visualizing it. These interdependencies were detected for the ASV (Figure 1(a)) and for various other editions the visualization showed a similar pattern. The visualization reveals further insights by highlighting other cells of the grid. For example, there is an indication for vast systematic Text Re-use between the books *Ezra* and *Nehemiah*. Also, there is evidence for repetitive Text Re-use given (Figure 1(b)). Picking the corresponding cell in the Text Re-use Browser allows for close reading and reveals a rectangular cluster of repeatedly used phrases to be compared using the Text Re-use Alignment Visualization, and a large systematic Text Re-use pattern between the beginning of *Ezra* and the middle section of *Nehemiah* (*Ezra 2:1/Nehemiah 7:6 - Ezra 2:70/Nehemiah 7:73*). When juxtaposing the KJV and its revision by Webster (Figure 1(c)) the systematic Text Re-use pattern for *Ezra* and *Nehemiah* is still highlighted. For the juxtaposition of the KJV and YLT, which uses Hebrew syntax, the overall number of Text Re-uses strongly decreases and a systematic Text Re-use pattern for *KJV:Ezra* and *YLT:Nehemiah* is not detected (Figure 1(d)). Interestingly, for the juxtaposition of *KJV:Nehemiah* and *YLT:Ezra* a small part of the systematic Text Re-use pattern is still preserved. Those are results causing the user to analyze Text Re-use further and to gain knowledge that wasn't expected or even looked for.

The Text Re-use Alignment Visualization turned out to be very useful for philological matters since syntactic similarities and differences between repetitive Text Re-use occurrences and verses of different editions can be analyzed easily.

[1] http://etraces.e-humanities.net/

Variations are easy to detect, for example many synonyms for "merchants" as seen in Figure 3(a) for the opening of *Ezekiel 27:22*. Most of the early Bible versions translated into Middle English used the token "marchauntes" except the Wycliffe Bible, which used "silleris". Most often, the early Middle English translations vary strongly from modern English translations which approves a separate analysis. For a repeatedly used phrase in *Numbers 1*, we detect a stronger variance for the 6 early translations compared to the 17 modern translations. The number of uses in different translations of the Bible implies that the long, possibly more precise and most often used phrase "families by the house of their fathers" (8 times) in the modern translations could be the most literal translation of the original text, an impression that can now be researched and verified or falsified (Figure 3(b)).



(a) The opening of *Ezekiel 27:22*. (b) Frequent phrase in *Numbers 1*.

Fig. 3. Text Re-use Visualizations show the variability of English Bible translations

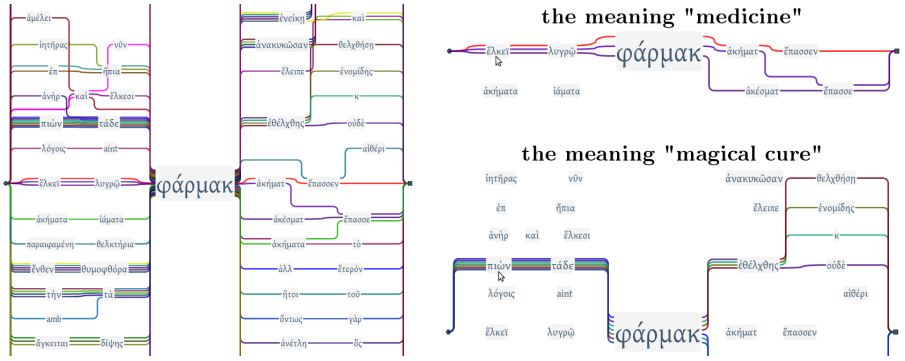
The humanities scholars also stated that the Text Re-use visualizations can help to determine, whether English versions of the Bible that claim to translate the Hebrew and ancient Greek original very literally, do this in a similar way or not and which one could be considered the most literal one.

7.2 The Meaning of Terms in Ancient Greek Texts

The purpose of the Digital Humanities project *eXChange*[2] is to explore the various meanings of a term, how these meanings changed over the years and how they were transferred in the past. Based upon ancient Greek texts, the meanings of a term are described by a set of various words.

One of the examples interesting for the collaborating classical philologists is the various meanings of *φάρμακον*. Utilizing the Text Re-use Alignment Visualization, text snippets containing the truncated form *φάρμακ* can be analyzed (Figure 4(a)). Immediately, the known two meanings of *φάρμακ* to be a

[2] http://exchange-projekt.de/



(a) Aligned text snippets with *φάρμακ* (b) Detecting meanings in the graph

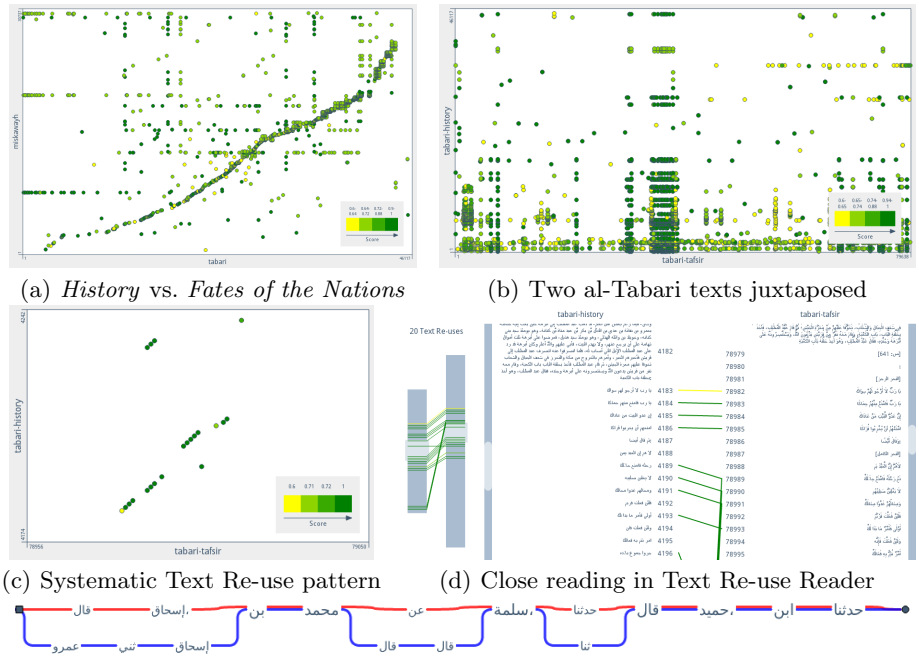
Fig. 4. Utilizing Text Re-use Alignments to discover concepts in ancient Greek texts

"medicine" or a "magical cure" can be perceived. Interactively, one is able to highlight the corresponding branches in the visualization (Figure 4(b)). The first meaning can be identified by the token *ἔλκει* (wound) and various variants of *ἀκημα τα* (cure). The token *ἐθέλχθης* (cast a spell) clusters text snippets for the second meaning. An important role for the given meanings plays the administration of a drug. If a drug is applied (*ἔπασσε*), it is a medicine, whereas if it is drunk (*πίων*), it is a magical cure. In contrast to the traditional methods of analyzing the contexts of the given terms' occurrences, the visualization facilitates a rapid comprehension of its opposing meanings by clustering related tokens that define the meaning of a term.

7.3 Text Re-use in Historic Arabic Texts

To explore the Text Re-use among historic arabic texts for the first time digitally, historians from the Aga Khan University utilized Büchler's algorithm for detecting Text Re-use and the Text Re-use Browser for visualizing and exploring the results. Predominantly, the focus was on the analysis of systematic Text Re-use. On the one hand, known facts were confirmed to assess the reliability of the visualization, on the other hand, unexpected and unknown patterns were analyzed further.

Figure 5(a) shows evidence for systematic Text Re-use in form of a diagonal pattern between two chronologies: *History*, called *Ta'rīkh al-rusul wa-l-mulūk* by al-Tabari (839–923), and *Fates of the Nations*, called *Tajārib al-umam* by Miskawayh (932–1030). Modern historians have often argued that Miskawayh relied heavily on al-Tabari's text. The analysis with the visualization suggests a more complex picture, namely, that Miskawayh more selectively copied al-Tabari's text. With the interaction capabilities of the Text Re-use Browser, the historians discovered that Miskawayh copied al-Tabari's text directly for the Umayyad (661–750) and Abbasid (750–1517) periods but copied very little of it for the period up to 651, which includes Iran's pre-Islamic history and the history



(a) *History* vs. *Fates of the Nations* (b) Two al-Tabari texts juxtaposed

(c) Systematic Text Re-use pattern (d) Close reading in Text Re-use Reader

(e) Alignment visualization for a Text Re-use occurrence

Fig. 5. Text Re-use Browser to discover systematic Text Re-use in Arabic texts

of the early Muslim community. It seems possible that Miskawayh wanted a fresh reading. To examine this judgment, the historians plan a further Text Re-use analysis, including a comparison of Miskawayh's text against a larger pool of digitzed Arabic texts.

Another research question tries to discover what conclusions can be drawn from common passages in a single author's works. In Figure 5(b), the Text Re-use between al-Tabari's *History* and his *Commentary on the Qur'an*, called *Jāmi' al-bayān 'an ta'wīl āy al-Qur'ān*, is shown. After removing vast occurrences of repetitive Text Re-use, especially stock phrases, the remaining systematic Text Re-use patterns can be analyzed. An example is given in Figure 5(c). The pattern begins with the statement *"According to what someone with knowledge claimed ..."* (Figure 5(d)). Read on its own, one might think this was al-Tabari's introduction to a topic or report. This might be the case, with al-Tabari repeating himself. It seems at least as likely, however, that this small bit of introduction derives from an original source, which al-Tabari (or perhaps a member of his editorial workshop) copied into both of his texts. Detecting chunks like this across his text, and comparing them to textual units in other classical Arabic texts, might give a sense of the size of units that passed through the tradition.

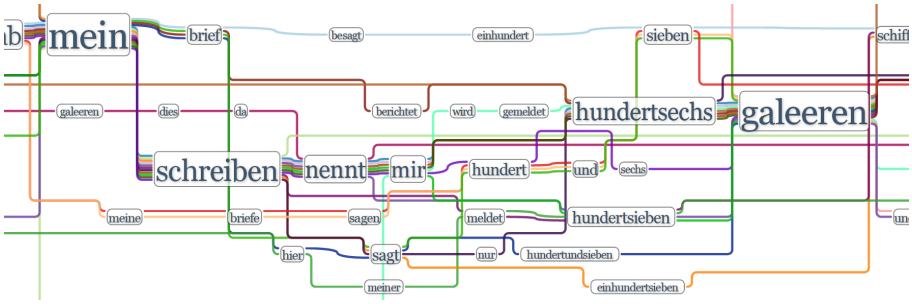
For both the examples, the Text Re-use Visualization Alignment was an effective method to investigate syntactic differences of a detected Text Re-use. An example, taken from the latter use case, is given in Figure 5(e).

7.4 The Comparison of German Shakespeare Translations

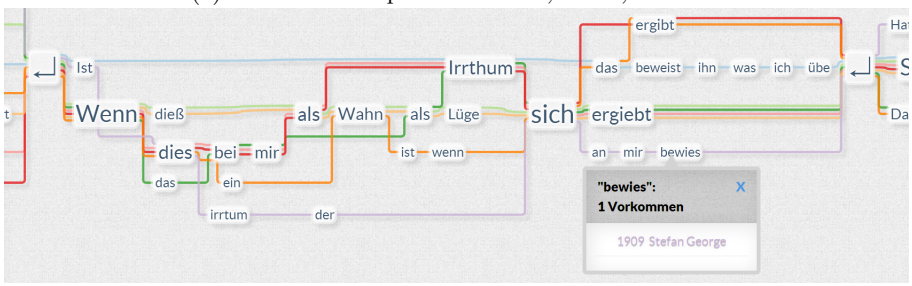
In the occasion of the 450th anniversary of William Shakespeare's birth and the 300th anniversary of the German Shakespeare Society[3], researchers from the Max Planck Institute for Mathematics in the Sciences, Leipzig and the Natural Language Processing Group of the Leipzig University developed novel language analysis methods and applications on Shakespearean works [10]. The contribution described rather formal, block entropy based measures for language complexity, but also had a focus on making structural aspects of the dramatic progression in Shakespeare's plays and the used language visible to the user in an interactive environment. As part of the latter, the Text Re-use Aligment Visualization was used to point out the differences and similarities of several German translations. Figure 6(a) shows the parallel text of over 20 different German versions of an Othello scene. Even this small segment shows the variety of visible aspects of the language. Regarding word position and sequence, the so called syntagmatic level bears fixed and grammar-related expressions, such as singular "mein brief" (my letter) versus plural "meine briefe" (my letters) or active constructions "nennt mir" (tell me) versus passive "mir wird gemeldet" (is reported to me), which can be tracked visually, mainly along the horizontal axis. The vertical axis yields word substitutions and chosen alternatives, the so called paradigmatic level, for example the nouns "brief" (letter) and "schreiben" (writing). The visualization lets the user estimate the relative popularity of sequences and alternatives and the longer-ranging effects of the choice of words. As a very interesting find, the original "hundred and seven galleys" are quite often changed to 106 in translations. Most probably, this is done for its shorter syllable count and the resulting softer pronunciation with "ga-lee-ren". However, other (as valid) variations (like 108) were never chosen, hinting at a high level of influence within the concerned translations.

The reception of this intuitive form of text variant representation was very positive. This lead to Text Re-use Aligment Visualizations being included as a central part in a follow-up cooperation with the Department of British Culture at the University of Bamberg focusing on Shakespeare's Sonnets. To compare works of poetry, it has proven useful to add a special line break marker that helps segmenting the verses visually and that also provides "incentives" for the alignment algorithm, to adhere to the verse structure, as can be seen in Figure 6(b). The interactive features of the Text Re-use Aligment Visualization were reported as very useful for quick inquiries into the data. The hover-based edge filtering clears the visual representation to show only the set of relevant parallel versions while retaining the broader context on all other used tokens.

[3] Deutsche Shakespeare-Gesellschaft, http://shakespeare-gesellschaft.de/



(a) Part of Shakespeare's *Othello, Act 1, Scene 3*



(b) Part of Shakespeare's *Sonnett 116*

Fig. 6. Text Re-use Alignments for various German translations of Shakespeare texts

The click-induced popup box provides the details for all crossing versions unambiguously. Both interactions are of great use in scenarios showing ten, twenty and more editions, which are common in the translation analysis of very popular authors, like Shakespeare.

8 Conclusion And Future Work

To support humanities scholars in exploring and analyzing the Text Re-use among historic text collections, we designed various close and distant reading visualizations. The *Text Re-use Grid* is a novel distant reading approach to discover type and amount of Text Re-use between each pair of texts of a given text corpus. At the researcher's convenience, one is able to highlight either grid cells with systematic or repetitive Text Re-use. The *Text Re-use Browser* facilitates a further exploration of such Text Re-use patterns between two texts and allows for close reading of individual text passages. This bridge between both perspectives, which fulfills Shneiderman's Information Seeking Mantra, turned out to be an important aspect for the collaborating humanities scholars. The *Text Re-use Alignment Visualization*, a further close reading visualization, allows for exploring similarities and differences between various editions of a given text. Thereby, we focused on improving the design and the readability for so called Text Variant Graphs. Instead of vertices, we place the vertices' labels with variable font size that reflect the number of occurences on horizontal layers. We attached great

importance to the vertical alignment of variations to allow an easy detection of synonyms.

During the development phase, the collaborating humanities scholars steadily evaluated the design of the Text Re-use visualizations. We wanted to ensure creating intuitive and flexible interfaces to be able to help answering a broad palette of research questions. Four usage scenarios for various English Bible editions, the meaning of terms in ancient Greek texts, the Text Re-use among historic Arabic texts and the comparison of German Shakespeare translations confirm the benefit of this iterative process and the adaptability of the visualizations independent on the language of the texts in the given corpus.

In the future, we will direct our attention on the development of distant reading visualizations for Text Re-use Alignments. This would allow for comparing the various editions of a text on a different level and for detecting global patterns among those editions.

Acknowledgments. The authors like to thank Sarah Bowen (Aga Khan University), who utilized the presented Text Re-use Visualizations for historic Arabic texts, Eva Wöckener-Gade (Leipzig University), who worked with the Text Reuse Alignment Visualization to analyze the various meanings of ancient Greek terms and Annette Geßner (Göttingen Centre for Digital Humanities) for the collaboration when designing the Text Re-use Visualizations for English Bible translations. This research was funded by the German Federal Ministry of Education and Research.

References

1. Andrews, T.L., Mac, C.: Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. Literary and Linguistic Computing (2013)
2. Bourdaillet, J., Ganascia, J.G.: Practical block sequence alignment with moves. In: Loos, R., Fazekas, S.Z., Martn-Vide, C. (eds.) LATA. vol. Report 35/07, pp. 199–210. Research Group on Mathematical Linguistics, Universitat Rovira i Virgili, Tarragona (2007)
3. Büchler, M.: Informationstechnische Aspekte des Historical Text Re-use (2013)
4. Büchler, M., Geßner, A., Eckart, T., Heyer, G.: Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1(2) (2010)
5. Cheesman, T., Flanagan, K., Rybicki, J., Thiel, S.: Six Maps of Translations of Shakespeare. In: Wiggin, B., Macleod, C., DiMassa, D., Theis, N. (eds.) Un/Translatables: New Maps for Germanic Literatures. Northwestern University Press (2014)
6. Clough, P., Gaizauskas, R., Piao, S.S.L., Wilks, Y.: METER: MEasuring TExt Reuse. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 152–159. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)

7. Collins, C., Carpendale, S., Penn, G.: Visualization of Uncertainty in Lattices to Support Decision-Making. In: Proceedings of the 9th Joint Eurographics / IEEE VGTC conference on Visualization. pp. 51–58. EUROVIS'07, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland (2007)
8. Dekker, R.H., Middell, G.: Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements. Supporting Digital Humanities (2011)
9. Diehl, S.: Software Visualization: Visualizing the Structure, Behaviour, and Evolution of Software. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2007)
10. Efer, T., Heyer, G., Jost, J.: Text Mining am Beispiel der Dramen Shakespeares. In: Christa Jansohn (ed.) Proceedings of the Symposium "Shakespeare unter den Deutschen" (2014)
11. Gibbs, A.J., McIntyre, G.A.: The Diagram, a Method for Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences. Eur J Biochem 16(1), 1–11 (1970)
12. GuttenPlag: GuttenPlag Wiki Visualizations (2013), `http://de.guttenplag.wikia.com/wiki/Visualisierungen` (Retrieved 2013-06-10)
13. Jänicke, S., Büchler, M., Scheuermann, G.: Improving the Layout for Text Variant Graphs. In: VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources. pp. 41–48 (2014)
14. Jenks, G.F., Caspall, F.C.: Error on Choroplethic Maps: Definition, Measurement, Reduction. Annals of the Association of American Geographers 61(2) (1971)
15. John, M., Heimerl, F., Müller, A., Koch, S.: A Visual Focus+Context Approach for Text Comparison Tasks. In: VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources. pp. 29–32 (2014)
16. Lee, J.: A Computational Model of Text Reuse in Ancient Literary Texts. In: Association for Computational Linguistics (ed.) Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 472–479 (2007)
17. Moretti, F.: Distant Reading. Verso (2013)
18. Needleman, S.B., Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of molecular biology 48(3), 443–453 (1970)
19. Nelson, T.H.: Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning, and Deep Re-Use. ACM Computing Surveys (CSUR) 31(4es), 33 (1999)
20. Ribler, R.L., Abrams, M.: Using Visualization to Detect Plagiarism in Computer Science Classes. In: Proceedings of the IEEE Symposium on Information Vizualization 2000. pp. 173–178. INFOVIS '00, IEEE Computer Society, Washington, DC, USA (2000)
21. Schmidt, D., Colomb, R.: A data structure for representing multi-version texts online. International Journal of Human-Computer Studies 67(6), 497 – 514 (2009)
22. Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Visual Languages, Proceedings. pp. 336–343 (1996)
23. Slocum, T.A., McMaster, R.B., Kessler, F.C., Howard, H.H.: Thematic Cartography and Geovisualization. Prentice Hall Series in Geographic Information Science, Prentice Hall, 3rd, international edn. (2009)
24. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2004)
25. Wattenberg, M., Viégas, F.B.: The Word Tree, an Interactive Visual Concordance. IEEE Transactions on Visualization and Computer Graphics 14(6), 1221–1228 (Nov 2008)