Visualizations for Text Re-use

Stefan Jänicke[1], Annette Geßner[2], Marco Büchler[2] and Gerik Scheuermann[1]
[1]*Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Leipzig, Germany*
[2]*Göttingen Centre for Digital Humanities, University of Göttingen, Göttingen, Germany*

Keywords: Text Re-use, Text Visualization, Bible Visualization, Intertextuality, Digital Humanities.

Abstract: In this paper, we present various visualizations for the Text Re-use found between texts of a collection to support humanists in answering a broad palette of research questions. When juxtaposing all texts of a corpus in the 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. In contrast, the *Text Re-use Browser* allows for close reading of the Text Re-use between the two texts of a tuple. Additionally, we present *Sentence Alignment Flow*s to improve the readability for Text Variant Graphs on sentence level that are used to compare various text editions to each other. Finally, we portray findings of the humanists of our project using the proposed visualizations.

1 INTRODUCTION

Text Re-use is defined as the oral or the written reproduction of textual content (Büchler, 2013) and is roughly divided into two types. On the one hand, a text passage is re-used deliberately, like direct quotes and allusions[1], and phrases like winged words and wisdom sayings. Translations of a text into other languages also count to this group and is called interlingual Text Re-use. A very popular form of deliberate Text Re-use is plagiarism. It has gained major attention in the recent years, mainly driven by plagiarism allegations in politics. On the other hand, a Text Reuse may be unintended, like boilerplates, e-mail headers or the repetition of news agency texts when writing daily newspapers (Clough et al., 2002). Further examples are idioms, battle cries and so called multi word units.

The interdisciplinary Digital Humanities project *eTRACES*[2] that wants to discover all these kinds of intertextual similarities between historical texts of a given corpus meets two major challenges. Since manual detection of re-used text passages is virtually impossible for the collaborating humanists, the first challenge is the automation of this process for unsupervized Text Re-use detection. For this purpose, the *TRACER* tool was developed, so that traces of re-used text are annotated automatically when operating on vast collections of texts. To finally create new digital editions of texts, the humanists of the project want to analyze, evaluate and revise the found results. This leads to the second challenge: the transformation of these results into intuitive visual interfaces that support the humanists in achieving their goals.

One focus of the project is the detection and visualization of Text Re-uses within the Bible; the given corpus consists of seven different English translations (see Section 3.3). The humanists are particularly interested how specific phrases *spread* in a text – so called *Repetitive Text Re-use* – and which texts share patterns of consecutive similar sentences – so called *Systematic Text Re-use* (see Section 3.2). Furthermore, the analysis of these similar sentences regarding structure, context and used expressions is of special interest.

This paper shows, how visualizations help to answer the humanist's questions. Although particularly designed for the Bible data of the project, the following visualizations for Text Re-use can be adapted for an arbitrary collection of texts:

- **Text Re-use Grid**: a chart that juxtaposes all texts of a collection (e.g., Bible books) in relation to the number and the type (systematic and/or repetitive) of the detected Text Re-uses,
- **Text Re-use Browser**: a user interface that allows for the inspection and browsing through all Text Re-uses between two or more texts,

[1]An allusion is "an expression designed to call something to mind without mentioning it explicitly" (Oxford English Dictionary)

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

Jänicke S., Geßner A., Büchler M. and Scheuermann G..
Visualizations for Text Re-use.
DOI: 10.5220/0004692500590070
In *Proceedings of the 5th International Conference on Information Visualization Theory and Applications* (IVAPP-2014), pages 59-70
ISBN: 978-989-758-005-5

- **Sentence Alignment Flow**: a transformation of aligned sentences (e.g., Bible verses) into an interactive visualization that improves the readability of so called *Text Variant Graphs*.

2 RELATED WORK

With the steadily growing amount of digital texts that can be retrieved through web portals like Google Books[3] or HathiTrust[4], visualizations for the contents and its metadata gain more and more popularity. A lot of work has been done in the field of text visualization. We want to emphasize these ones that are topically close to our work and which were inspiring for designing Text Re-use visualizations for the Bible.

2.1 Text Re-use Visualizations

For displaying the results of an automatic detection of Text Re-uses in ancient Greek texts, Büchler provides a frontend called *CitationGraph* that visualizes the Text Re-use found for a certain author in an ancient Greek text corpus by number, citing authors, years of citing authors and passages of the book (Büchler et al., 2010). Additionally, the user can inspect individual text snippets with highlighted re-used passages. This approach has proven useful to humanists in the *eAQUA* project (Geßner, 2010).

For displaying the variants for a quote, Leskovec introduced the phrase graph (Leskovec et al., 2009), with vertices in the form of phrase clusters and directed edges for relations (inclusions) among the corresponding phrases.

Lee uses a static *Dot Plot View* for plotting Text Re-use found between Bible books (Lee, 2007). As introduced by Gibbs and McIntyre (Gibbs and McIntyre, 1970), the Dot Plot View is used in bioinformatics to compare two genome sequences to each other. 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.

The GuttenPlag Wiki (GuttenPlag, 2013) provides several visualizations for plagiarized passages of Guttenberg's dissertation[5]. 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 certain authors of the original source. The usage of colored blocks is also a common practice to visualize re-used program code passages (Freire, 2008; Ribler and Abrams, 2000).

All these visualizations focus on displaying the extent of Text Re-use of a given source text, or a text that contains lots of re-used passages. A comparative overview between all texts of a text collection is not provided. For this purpose, we present the *Text Re-use Grid* in Section 4.1.

2.2 Alignment Visualizations

Data alignments are common tasks in various research fields. The visualization of ontology alignments in the form of graphs for a better understanding of specific semantic relations is an example from the semantic web community (Lanzenberger and Sampson, 2006). In bioinformatics, there are numerous tools for the visualization of sequence alignments (Procter et al., 2010). For example, genome alignment visualizations are used to help researchers to quickly detect important genomic variations (Herbig et al., 2012). Thereby, distinct colored paths indicate distinct genomes.

A lot of use cases also exist for the alignment of texts. Cheesman offers a visualization for the alignment of multilingual text passages in Shakespeare's Othello in the form of a web interface, where the user can interactively browse through the texts of two editions (Cheesman et al., 2012). In contrast, Büchler provides a horizontal alignment of Text Re-uses between text passages of the same language (Büchler et al., 2010). 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 solution 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 supports also various sub-branches (Collins et al., 2007). But merging of multiple nodes of the same kind is not provided, although the metaphor for uncertainty could be used for this purpose.

A visualization, which allows for weighted nodes is the *Word Tree* (Wattenberg and Viégas, 2008). It arranges a set of sentences that start with the same set of words in the form of a tree. Each variation results in a split into several leaves. Thereby, the font size of a node label reflects the number of occurrences.

A graph visualization for Text Variant Graphs – a data structure representing various editions of a text – is proposed by Andrews (Andrews and Van Zundert, 2013). It can be seen as an extension of Büchler's

[3]http://books.google.de/

[4]http://www.hathitrust.org/

[5]In 2011, the current German Federal Minister of Defence Karl-Theodor zu Guttenberg was convicted of plagiarizing his doctoral dissertation.

approach. The graph provides a lot of interaction means for humanists to work on the automatic text alignment results like merging and splitting of vertices. However, it is hard for the user to follow how one edition disseminates in the graph. Furthermore, the vertices do not reflect the amount of occurrences and synonyms are not properly aligned to each other. In section 4.3, we propose the *Sentence Alignment Flow* that combines Andrews' concept and some of the other presented design ideas with the goal to improve the readability for Text Variant Graphs.

2.3 Visualizations of the Bible

Several visualizations exist for cross references between Bible verses, which are co-occurrences of similar events (e.g., "Jesus walks on the water"), themes (e.g., "The Tower of Babel"), or persons (e.g., "Cain and Abel"). Therefore, a cross reference can be seen as a basic form of Text Re-use. The *Bible Cross References Visualization* (OpenBible.info, 2012) is a grid with each cell containing a cross reference graph for each pair of Bible books. So, the observer gets an imagination about the amount of shared cross references between two Bible books. A single graph consists of two vertical axes; in ascending order, each verse of a book gets a position on the corresponding axis. When drawing cross references in the form of connections, the reader gets an overview about co-occurring entities, but an indication for the type of Text Re-use is not given, and an exploration of individual co-occurrences is also not possible. Harrison's visualization (Harrison and Römhild, 2008) orders all verses on a horizontal axis and represents each cross reference with an arc between the corresponding verses. Although edge coloring is used, the presence of around 64,000 arcs makes the visualization hard to read and patterns hard to discover.

In this paper, we present visualizations for Text Re-uses found in the Bible that broaden the capabilities to explore known facts and allow for the discovery of new insights.

3 TEXT CORPUS

Within the eTRACES project, computer scientists and humanists collaborate to explore and measure to what extent Text Re-use passages can be detected automatically. Intuitive, interactive visualizations are the bridge for the humanists to help understanding and interpreting the computed results. In general, the goal is to find traces of re-used text, more precisely, when, where and to what extent specific text passages were re-used. Independent of the research field, the algorithms and visualizations presented in this paper can be utilized for an arbitrary text collection.

3.1 Text Re-use Data

Let $A_1, \ldots, A_n$ denote a corpus of n texts. After splitting each text into a list of sentences, the automatic Text Re-use detection algorithm searches for Text Re-uses between each pair of sentences from distinct texts and within one and the same text. Each found Text Re-use $\{a_i, b_j\}$ consists of the two corresponding Text Re-use units a_i (i-th sentence of text A) and b_j (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 sentence lengths of a_i and b_j and their Re-use overlap, which is the proportion of matching and non-matching tokens. t is ranged in the interval $[0, 1]$; 0 means no similarity between two verses, 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 for t is outside the scope of this paper and can be found in Büchler's dissertation (Büchler, 2013).

3.2 The Type of Text Re-use

The humanists working in our project – or with Text Re-use in general – have various research questions. Therefore, we define two types of Text Re-use:

Systematic Text Re-use. The consecutive occurence of the same pattern of Text Re-uses 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 (mostly) in the same 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 $\{a, b_1\}, \{a, b_2\}, \{a, b_3\}, \ldots$

3.3 Bible Data

Since the Bible is known as one of the most often read and studied books, and therefore, easily evaluable, it was chosen as a proof of concept for the project. The sample text corpus that is used as the leading example

for this paper contains seven different English translations of the Bible:

- King James Version (KJV): Early modern English translation of the Bible with the intention to reflect the vision of the Church of England at that time (released in 1611).
- Webster's Revision (Webster): Revised KJV with grammatical changes and the replacement of archaic words into modern English (1833).
- Young's Literal Translation (YLT): Strictly literal translation of the original Hebrew and Greek texts; verses conform to Hebrew syntax (1862).
- Darby Version (Darby): Translated as exactly as possible from Hebrew and Greek texts to create a modern version for the unlearned (1890).
- American Standard Version (ASV): Version of the KJV with a strong focus on the USA (1901).
- Bible in Basic English (BasicEnglish): Bible translation with a limited English vocabulary (around 1000 words), so that more people worldwide can read and understand the text (1965).
- World English Bible (WEB): Revision of ASV with the goal of global validity (2000).

The Bible versions were reduced to a total of 28,632 verses that are included in all seven translations. The automatic Text Re-use detection algorithm iterates over all verse tuples for all pairwise permutations of Bible editions. In this context, $\{a_i, b_j\}$ indicates a Text Re-use between the i-th verse of edition A and the j-th verse of edition B.

4 VISUALIZATION DESIGN

The conscientious analyzation and interpretation of small text passages – 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 (Moretti, 2005) that give the observer an impression about the data distribution. The Information Seeking Mantra "Overview first, zoom and filter, details-on-demand" (Shneiderman, 1996) is accomplished, when distant reading views are interactively used to switch to close reading views. The 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.

The visualizations we present in this chapter realize this process for Text Re-use data. The *Text Re-use Grid* is a distant reading visualization that highlights frequent, systematic and repetitive Text Re-uses between each text pair of a given corpus. It can be used to drill down and explore a preferred pair in the *Text Re-use Browser*. Like the *Sentence Alignment Flow* that allows to analyze various occurrences of re-used phrases, this visualization also supports close reading for Text Re-uses.

4.1 Text Re-use Grid

The intention of this visualization is to give the researcher an overview of the distribution of Text Re-uses between all texts of a corpus. We transform the result of the automatic 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.

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

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 sentences, 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 a list of Text Re-uses $\{a_i, b_j\}$ found between A and B. A Text Re-use $\{a_i, b_j\}$ is removed if:

- it contains a repeatedly re-used sentence a_i or/and b_j (repetitive Text Re-use), or
- it has no adjacent Text Re-use $\{a_u, b_v\}$ within a certain neighborhood ε (isolated Text Re-use), so that:

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

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

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

with

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

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. Low values occur, when the Text Re-uses are mostly repetitive or independent from each other.

Repetitive Text Re-use Index ω. In contrast to λ, 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 sentences 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 the 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. Although zooming is possible, the available screen space limits the number of texts that can be visually compared to each other.

Because of the importance for the humanists 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 chose an optimal classification method (Slocum et al., 2009) to group the cells into two sets of classes in dependency of σ, λ and ω. With the Jenks-Caspall-Algorithm (Jenks and Caspall, 1971) using reiterative cycling, we compute a configurable number of classes. We receive n classes $\alpha_1, \ldots, \alpha_n$ for the amount of Text Re-use with α_1 containing the cells with smallest σ and α_n containing the cells with the largest σ. We use the class α_i $(1 \leq i \leq n)$ to define the saturation of a cell color:

$$saturation = \frac{i}{n} \cdot 100$$

Thus, high amounts of Text Re-use receive highly saturated, and few amounts lightly saturated colors. Furthermore, we compute m classes $\beta_1, \ldots, \beta_m$ for the type of Text Re-use (systematic or repetitive), so that β_1 contains the cells with the smallest λ (or ω) and β_n contains the cells with the largest λ (or ω). For the mapping of these classes to colors, we facilitate color temperature. Therefore, we utilized the "Cold-Hot" color scale Diehl proposes (Diehl, 2007) for the *EpoSee* tool from blue (cold) to red (hot). We determine the hue of a cell color for a cell with class β_j $(1 \leq j \leq m)$ as

$$hue = 240 + \frac{j-1}{m-1} \cdot 120$$

So, 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 visual attraction of hot colors also fits to the importance for the humanists to discover texts with extensive systematic or repetitive Text Re-use. Finally, using $value = 100$ a cell color is defined in the HSV color space. The resultant colors for $n = m = 3$ and for $n = m = 4$ are shown in Figure 1.

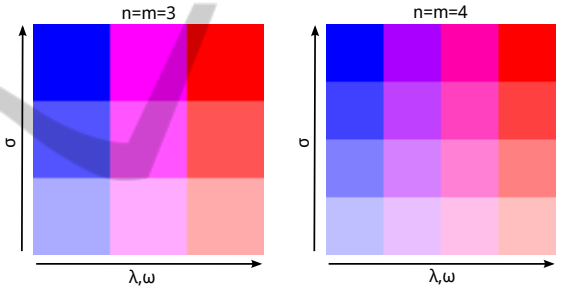
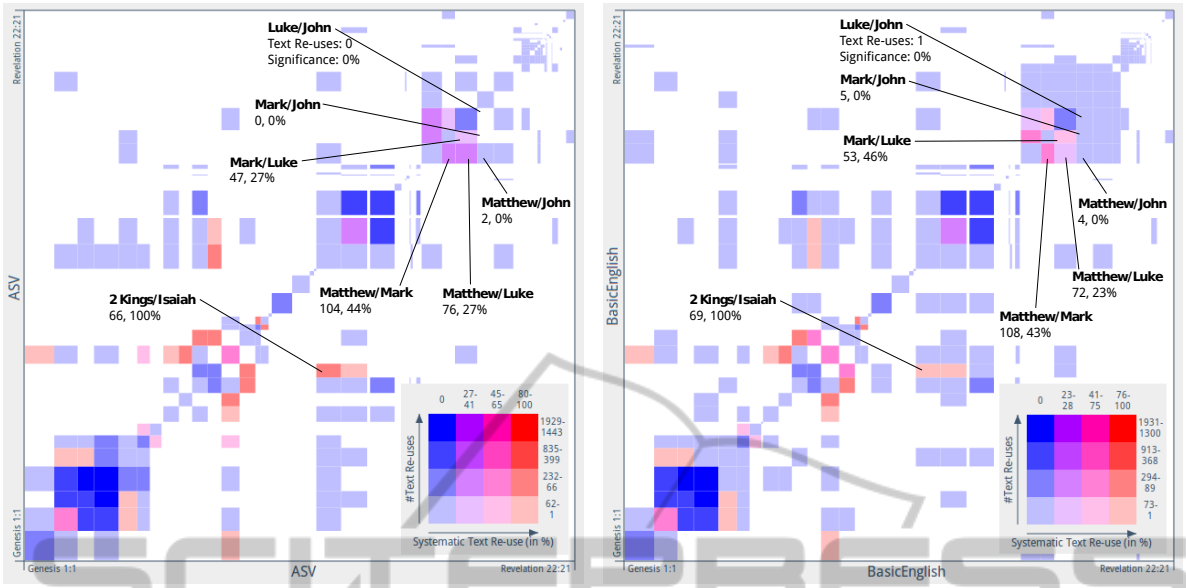


Figure 1: Resultant colors for the Text Re-use Grid.

In Figures 2(a) and 3, the resultant Text Re-use Grids for the Bible books of the American Standard Version compared to each other highlighting systematic and repetitive Text Re-use 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 texts. 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 a close reading browser view that is explained in the next section.

4.2 Text Re-use Browser

In order to allow the inspection of Text Re-uses found between two texts $A = \{a_{first}, \ldots, a_i, \ldots, a_{last}\}$ and $B = \{b_{first}, \ldots, b_j, \ldots, b_{last}\}$, the Text Re-use Browser provides two panels for this purpose: a *Dot Plot View* and a *Text Re-use Reader*.



(a) Juxtaposition for ASV/ASV. (b) Juxtaposition for BasicEnglish/BasicEnglish.

Figure 2: Text Re-use Grid showing juxtapositions of Bible books highlighting systematic Text Re-use.

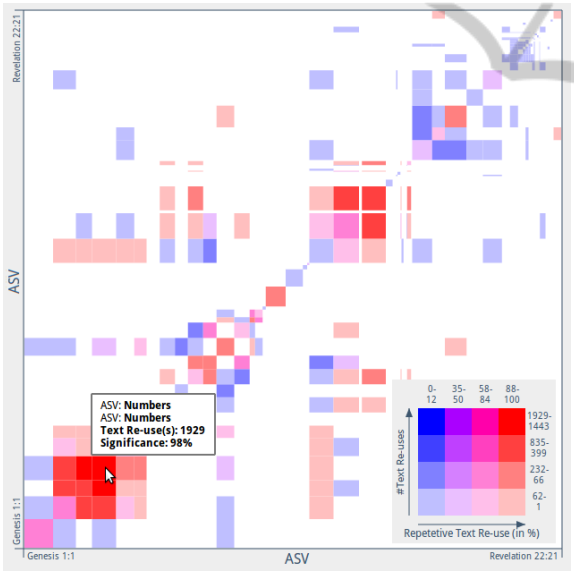
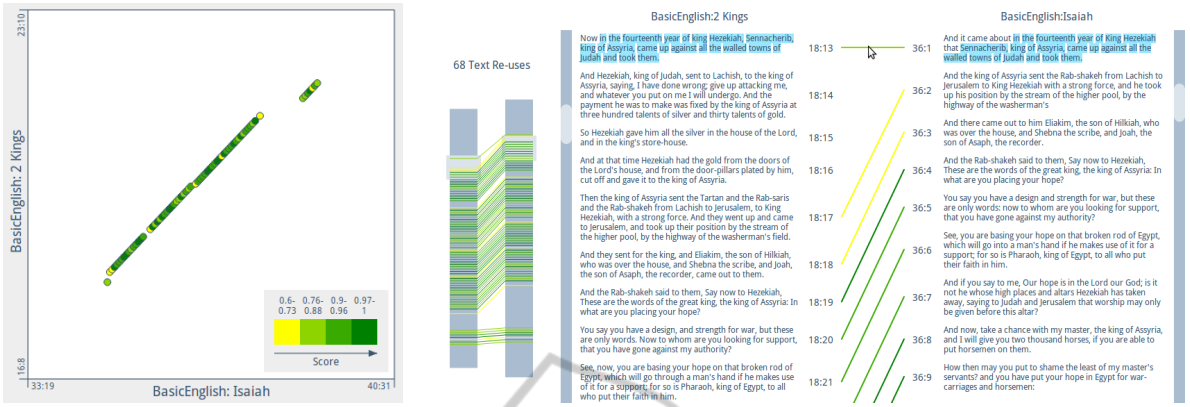


Figure 3: Text Re-use Grid showing juxtaposition for ASV/ASV highlighting repetitive Text Re-use.

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 (Lee, 2007), we provide an interactive chart, where the number $|A|$ of sentences of A defines the range for the x-axis, and the number $|B|$ of sentences of B defines the range for the y-axis. Each Text Re-use for a sentence pair is drawn as a single dot. As in bioinformatics, specific patterns indicate specific Text Re-use types. Diagonal patterns highlight sections that contain systematic Text Re-use (Figure 4(a)), whereas vertical and horizontal dot arrangements appear for phrase repetitions (Figure 5(a)). By selecting a dot via mouse click, a popup with the corresponding sentences and a Sentence Alignment Flow (see Section 4.3) is shown. Interactively, the user is also able to zoom into a rectangular region of interest (ROI).

Text Re-use Reader. This panel allows for browsing A and B in two opposite windows. Whenever a re-used sentence appears in the viewport of one window, a connection to the opposite sentence is drawn in the central area between the windows. A click on a connection scrolls both texts, so that the sentences 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 these words in both sentences, for which matches were detected with the sentence alignment algorithm (see Section 4.3). 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 (Figure 4(b)), and *hub*s (a single sentence of one text that is connected to a plenty of sentences of the opposite text) occur for repetitive Text Re-use (Figure 5(b)).

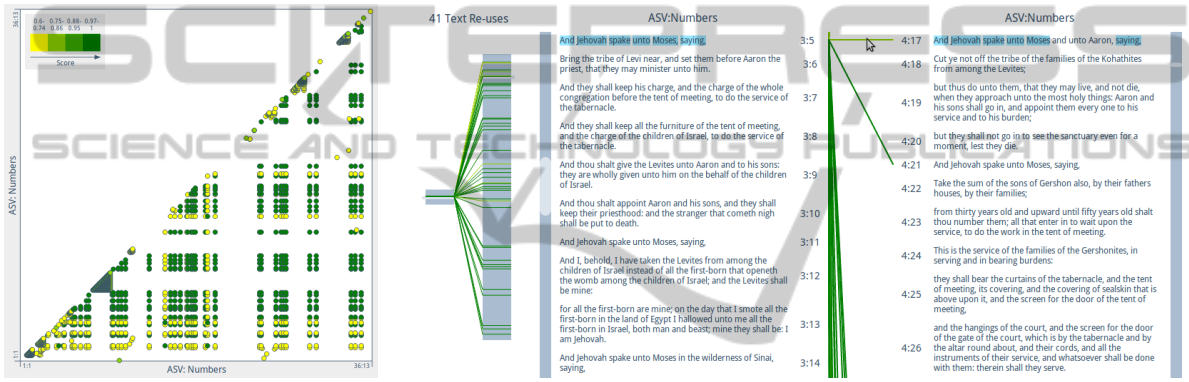
Both panels are linked to each other. A dot selection in the Dot Plot View triggers a scrolling of the texts to the corresponding positions, whereas a connection selection in the Text Re-use Reader opens the



(a) Diagonal pattern in the Dot Plot View.

(b) Parallel lines in the Text Re-use Reader.

Figure 4: Text Re-use Browser components showing systematic Text Re-use found for the Bible books *2 Kings* and *Isaiah*.



(a) Vertical and horizontal patterns in the Dot Plot View.

(b) Spread of the re-used verse *3:5* is visible in the Text Re-use Reader.

Figure 5: Text Re-use Browser components showing repetitive Text Re-use found in the Bible book *Numbers*.

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 compared to the Text Re-use Grid (Section 4.1). 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:

$$hue = 60 + \frac{k-1}{p-1} \cdot 60$$

To gain visually distinctive colors, the color value ranges between 100 and 50

$$value = 100 - \frac{k-1}{p-1} \cdot 50$$

and with $saturation = 100$ all glyph colors are defined in the HSV color space. Figure 6 shows the resultant colors for $p = 3$ and $p = 4$.

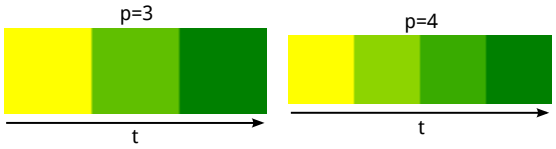


Figure 6: Resultant colors for the Text Re-use Browser.

Some text juxtapositions contain a lot of Text Re-uses that also create various patterns for Text Re-use. Therefore, we enable the user to visually filter for specific Text Re-uses. Firstly, we allow to hide glyphs of repetitive Text Re-use, and secondly, a slider can be used to hide isolated glyphs without adjacent glyphs in a certain neighborhood. Thirdly, the user is able to only display the Text Re-uses for a specific phrase. Thus, a drill-down to highlight only significant systematic or repetitive Text Re-use patterns is possible.

4.3 Sentence Alignment Flow

Humanists are interested in phrasing variants of a re-used entity and the contexts in which a specific phrase appears. The *Sentence Alignment Flow* is an interactive user interface that supports this task by visualizing a sentence alignment for a set of Text Re-uses. Furthermore, we provide several means of interaction to explore and modify the visualization and the underlying data structure.

Let $S = \{s_1, \ldots, s_n\}$ denote a set of sentences that share the same re-used entity. Preliminarily, we convert each sentence to lower case and remove punctuation characters. Afterwards, the sentences are split into tokens.

Sentence Alignment. We construct a directed acyclic graph $G = (V, E)$ as data structure for the sentence alignment. Each vertex $v = \{s_i, t_j, u_k, \ldots\} \in V$ is an aggregation of aligned tokens $\{s_i, t_j, u_k, \ldots\}$ that are equal to each other. The *token degree* $|v|$ is the number of tokens assigned to v, and $v(s_i)$ is the corresponding vertex in G for a sentence token s_i. We use a brute force algorithm to align the sentence tokens to each other. Thereby, we merge tokens of different sentences and use this alignment solution that reaches a maximum number of merge iterations while keeping G acyclic. Finally, each token of each sentence is a component of exactly one vertex of G. We insert a directed edge between two vertices, if they contain consecutive tokens for at least one sentence. Figure 7 shows such a graph for seven editions of the first Bible verse.

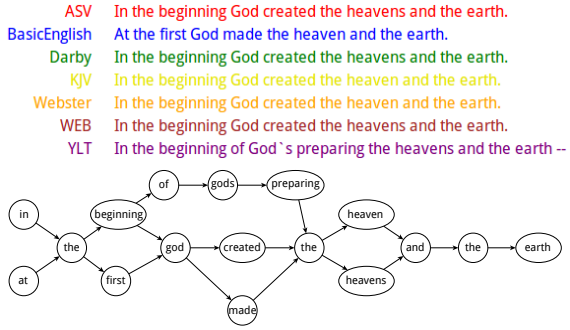


Figure 7: Sentence alignment DAG for seven editions.

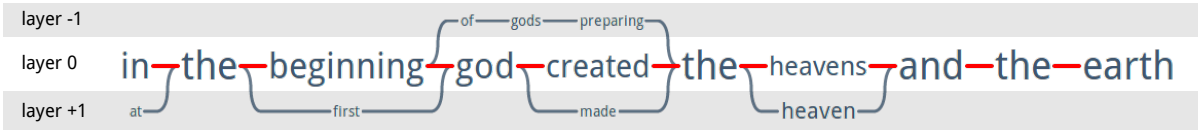
Graph Visualization. The corresponding token of a vertex is used for labeling. As Wattenberg proposes (Wattenberg and Viégas, 2008), we also use font size as a metaphor to reflect the number of occurrences of individual tokens. We layout the vertices of G by placing the corresponding labels onto horizontal 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 $v(s_1), \ldots, v(s_{|s|})$ for an arbitrary sentence $s \in S$ in left-to-right order on layer 0 (main branch). By default, we choose the sentence s with the maximum value for

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

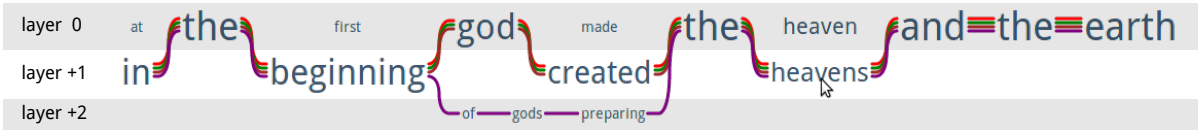
which has lots of tokens assigned to vertices with large token degrees. Then, we iteratively search for the shortest path $\{v_1, \ldots, v_n\} \in G$ with assigned layers for v_1 and v_n and the vertices of the subpath $p = \{v_2, \ldots, v_{n-1}\}$ without an assigned layer. Let i denote the layer 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 $k = \max(|i|, |j|)$, we iteratively search for a layer with free space for the labels of the vertices of p in the order $k, k+1, k-1, k+2, k-2$, etc. If the total width of the labels of p is larger than the space between v_1 and v_n, we preliminary stretch the distance between v_1 and v_n. After the proper layer is found, we move all vertices of G 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 paths containing vertices without assigned layers to complete the layout for the Sentence Alignment Flow. We draw undirected edges (for the user the direction is obvious) between the vertices in the form of horizontal lines of the same layer. To ensure a good readability of the graph, we use a horizontal line with a connection in the form of a Bézier curve for edges connecting vertices of different layers.

One application for this visualization is the alignment of seven editions of an individual Bible verse. For the color selection to identify the seven different sentence flows, we chose the following colors of the 12-color palette for categorial usage suggested by Ware (Ware, 2004) to facilitate maximal visual differentiation by the user: red, blue, green, yellow, orange, brown, and purple. Furthermore, we use a gray hue to draw the edges of the graph. The resultant Sentence Alignment Flows for seven editions of the first Bible verse using the ASV and BasicEnglish edition as main branch on layer 0 are shown in Figure 8.

To also support the work of researchers for textual criticism, who are interested in comparing different text editions to each other, we enable the user-driven modification of the underlying data structure. By dragging the labels over the surface, the user is able to merge vertices when it doesn't create cycles. We use cyan and pink colors to signalize feasible and permitted user-driven merge interactions. Furthermore, the splitting of a vertex v with a token degree $|v| > 1$ is



(a) Sentence Alignment Flow using the ASV as main branch (layer 0).



(b) Sentence Alignment Flow using the BasicEnglish edition as main branch (layer 0). Sentence flows containing the token "heavens" are highlighted.

Figure 8: Sentence Alignment Flows for the first verse of seven Bible editions.

also possible. Thus, the graph can be modified stepwise to gain the desired alignment and to correct potential errors of the alignment algorithm.

4.4 Implementation Notes

We implemented the proposed visualizations in JavaScript in the form of modules to facilitate the integration into web-based research platforms that are widely used in the Digital Humanities. We provide a JSON interface for the Text Re-use data; within our project, an Apache Solr[6] backend dynamically serves the required information. We use the Raphaël JavaScript library[7] for rendering all glyphs in the form of Scalable Vector Graphics.

Thus, the reponse time when loading a visualization depends on the used client browser and the number of glyphs to be displayed. For the Bible use case, the approximate number of rectangles in the Text Re-use Browser is 300, and the number of dots (lines) in the Dot Plot View (Text Re-use Reader) ranges from 0 to around 2,000.

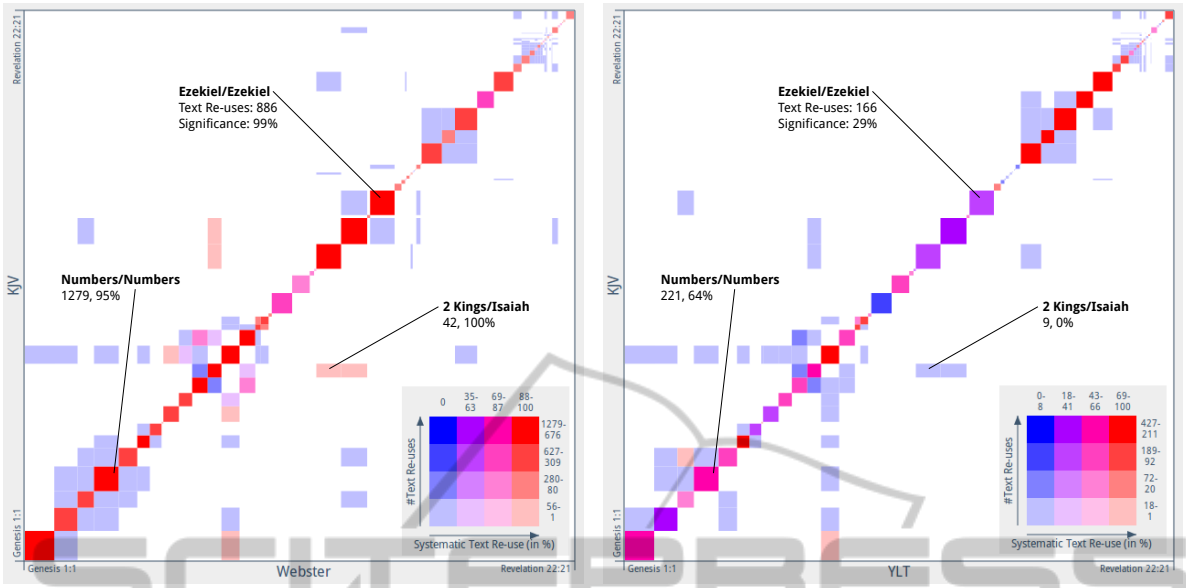
5 RESULTS

We worked together with three humanists experienced in the field of textual criticism in order to develop and improve the usability of the presented Text Re-use visualizations as well as to ensure their scientific benefit. Initially, we discussed in what way we could support the humanists in answering their research questions. In face to face sessions, we demonstrated the current status and gave some time to work with the visualizations. In subsequent interviews, we figured out problems and discussed potential enhancements of the design (e.g. color mapping). In this section, we present the humanist's final evaluation of the visualizations and their findings for the Bible data.

One of the major purposes was the utilization of the Text Re-use visualizations for various research questions. These can be divided into different perspectives: the user has the opportunity to either compare the same or different sections of the same text (e.g., an edition of the Bible), or texts from different editions (e.g., various editions of the Bible). Furthermore, the possibility to determine the relevance of the results either by the amount of systematic or repetitive Text Re-use, and the division into different visualizations that allow for a "more distant" or a "more close" view on the text is of great interest.

Biblical books of the same or two different editions that have a lot of systematic or repetitive Text Re-use are easy to find by using the Text Re-use Grid. Focusing on the diagonal line of books being compared with their pendant in the same or another Bible version and the squares next to them, some "clusters" of Biblical books seeming to have systematic and/or repetitive interdependencies can be figured out easily. Especially when comparing books of the same Bible version regarding systematic Text Re-use, the visualization 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. We detect these interdependencies for the ASV Bible (Figure 2(a)); the simplified language of the BasicEnglish edition yet increases this effect (Figure 2(b)). But the visualization also reveals other insights by highlighting other cells of the grid. For example, there is an indication for vast systematic Text Re-use between the book *2 Kings* and *Isaiah* in both Bible editions. Picking the corresponding cell in the Text Re-use Browser allows for close reading and reveals a large systematic Text Re-use pattern starting from *2 Kings 18:13* and *Isaiah 36:1* (Figure 4). Those are results caus-

[6] http://lucene.apache.org/solr/
[7] http://raphaeljs.com/



(a) Juxtapositions for KJV/Webster. (b) Juxtapositions for KJV/YLT.

Figure 9: Text Re-use Grid showing juxtapositions of Bible books for different editions highlighting systematic Text Re-use.

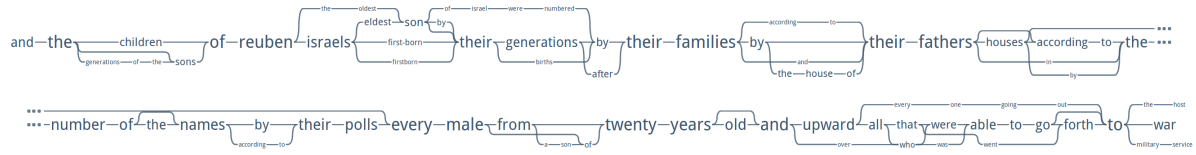
ing the user to gain knowledge that wasn't expected or even looked for[8]. When comparing various editions of the Bible to each other, the majority of Text Re-uses depends on the same Bible verses in the same Bible books. The Text Re-use Browser can be used to determine how similar or different translations of the Bible are. Especially here the Sentence Alignment Flow can be used to further explore the syntactic similarity between verses.

When juxtaposing the KJV and its revised version Webster (Figure 9(a)) the systematic Text Re-use pattern for *2 Kings* and *Isaiah* is still highlighted. For the juxtaposition of the KJV and the YLT that uses a different sentence syntax, the overall number of Text Re-uses strongly decreases and a systematic Text Re-use pattern for *2 Kings* and *Isaiah* is not detected (Figure 9(b)).

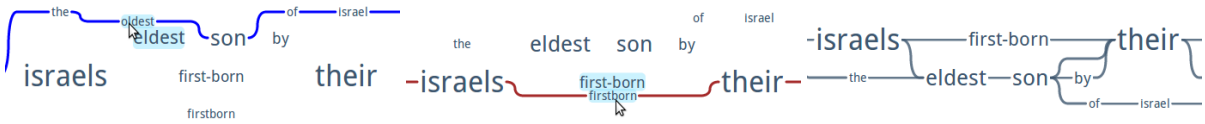
Researchers interested in biblical expressions and phrases are interested in those results with a high relevance concerning repetitive Text Re-use. Those seem to be found mainly inside one book of one edition, for instance, the book Numbers of the ASV offers 1,929 results (Figure 3) that can be compared in the Text Re-use Browser (Figure 5(a)). The Text Re-use Reader supports the process of exploring how a specific phrase is spread in a book (Figure 5(b)).

An even closer look at the specific structure of the same verses in different translations can be done with the Sentence Alignment Flow that is very useful for philological matters. Variations are easy to detect, for example many synonyms as seen in Figure 10(a) for the verse 1:20 of *Numbers* like "eldest son" and the variation "oldest son", "first born" and the variation "first-born". Now – depending on the research question – those words can be differentiated to determine how many translators used which variation to determine different translation techniques. But it is also possible to merge single variations by dragging and dropping the matching words and simple variations like "eldest" and "oldest" or "first-born" and "firstborn" to concentrate on variations more complex (Figure 10(b)). A merge of "israels" and "israel" would create a cycle in the data structure, and therefore, is not possible due to various sentence structures (Figure 10(c)). In this verse, the words between "families" and "their fathers" are of great interest because they vary a lot, using the single words "and" (once) or "by" (twice) or the phrases "according to" (once) and "by the house of" (three times). The number of uses in different translations of the Bible implies that the long, possibly more precise and most often used phrase "by the house of" could be the most literal translation of the original text, an impression that can now be researched and verified or falsified (Figure 10(d)).

[8]This so called *serendipity* effect is "the occurrence and development of events by chance in a happy or beneficial way" (Oxford English Dictionary)



(a) The complete Sentence Alignment Flow for *Numbers 1:20*.



(b) Valid merge steps for ”oldest”/”eldest” (left) and ”firstborn”/”first-born” (middle) and the final structure (right).

(c) Invalid merge attempt for ”israels”/”israel”.

(d) Most often used phrase ”by the house of”.

Figure 10: Features of the Sentence Alignment Flow for seven editions of *Numbers 1:20*.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented three visualizations for Text Re-use. The *Text Re-use Grid* is a novel 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 frequent systematic or repetitive Text Re-use. However, the Text Re-use Grid can only be applied to a limited text corpus, since the user's available screen space constrains the size of the cells to be displayed. A dynamic grid allocating more screen space to cells that are relevant dependent on the research question could be an alternative. The *Text Re-use Browser* facilitates a further exploration of the Text Re-use between two texts. In contrast to the Text Re-use Grid, which is a distant reading visualization, the Text Re-use Browser allows for close reading of individual text passages. This switch between both perspectives turned out to be an important aspect for the collaborating humanists. With the *Sentence Alignment Flow*, we developed a further close reading visualization for so called Text Variant Graphs. In comparison to the approach of Andrews (Andrews and Van Zundert, 2013), we focused on improving the readability of the visualization. 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 of editions to allow easy detection of synonyms. To support the collation process for the researchers of textual criticism, we provide an interactive interface that allows for a user-driven modification of the alignment to potentially create new editions of the given text. For a broad deployment of the Sentence Alignment Flow in the humanities, we need to extend the visualization with more means for the annotation of editions.

During the development phase, the humanists of our project steadily evaluated the design of the Text Re-use visualizations. We wanted to ensure creating an intuitive and flexible system to be able to help answering various research questions. The findings of the humanists listed in Section 5 confirm the benefit of this iterative process that should be always performed when developing visualizations for humanistic applications. The humanists also stated that our 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. Furthermore, the visualizations could also be used trying to determine how exactly literature is cited, when looking for indirect transmission doing textual criticism (Geßner, 2010).

Designed for the Bible data of the eTRACES project, we will test and evaluate the proposed visualizations for other text collections in the future. Firstly, the humanists aim at investigating the Text Re-use of Bible passages in the works of Friedrich Schiller. Secondly, an extraction and analyzation of Text Re-uses among the historical texts within the Perseus Digital Library[9] is planned.

[9]http://www.perseus.tufts.edu/

ACKNOWLEDGEMENTS

The authors like to thank Christian Heine for fruitful discussions and Markus Ackermann and Muhammad Faisal Cheema for proofreading. This research was funded by the German Federal Ministry of Education and Research.

REFERENCES

Andrews, T. L. and Van Zundert, J. J. (2013). An Interactive Interface for Text Variant Graph Models. In *Proceedings of the Digital Humanities 2013*.

Büchler, M. (2013). *Informationstechnische Aspekte des Historical Text Re-use*.

Büchler, M., Geßner, A., Eckart, T., and Heyer, G. (2010). Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2).

Cheesman, T., Thiel, S., Flanagan, K., Zhao, G., Ehrmann, A., Laramee, R. S., Hope, J., and Berry, D. M. (2012). Translation Arrays: Exploring Cultural Heritage Texts Across Languages. In *Proceedings of the Digital Humanities 2012*.

Clough, P., Gaizauskas, R., Piao, S. S. L., and Wilks, Y. (2002). METER: MEasuring TExt Reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 152–159, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collins, C., Carpendale, S., and Penn, G. (2007). Visualization of Uncertainty in Lattices to Support Decision-Making. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC conference on Visualization*, EUROVIS'07, pages 51–58, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Diehl, S. (2007). *Software Visualization: Visualizing the Structure, Behaviour, and Evolution of Software*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Freire, M. (2008). Visualizing Program Similarity in the AC Plagiarism Detection System. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, pages 404–407, New York, NY, USA. ACM.

Geßner, A. (2010). Das automatische Auffinden der indirekten Überlieferung des Platonischen Timaios und die Bedeutung des Tools CitationGraph für die Forschung. In Schubert, C. and Heyer, G., editors, *Das Portal eAQUA*, pages 26–41.

Gibbs, A. J. and McIntyre, G. A. (1970). The Diagram, a Method for Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences. *Eur J Biochem*, 16(1):1–11.

GuttenPlag (2013). GuttenPlag Wiki Visualizations. `http://de.guttenplag.wikia.com/wiki/Visualisierungen` (Retrieved 2013-06-10).

Harrison, C. and Römhild, C. (2008). The Visualization of the Bible. `http://www.chrisharrison.net/index.php/Visualizations/BibleViz` (Retrieved 2013-06-10).

Herbig, A., Jäger, G., Battke, F., and Nieselt, K. (2012). GenomeRing. *Bioinformatics*, 28(12):i7–i15.

Jenks, G. F. and Caspall, F. C. (1971). Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61(2).

Lanzenberger, M. and Sampson, J. (2006). AlViz - A Tool for Visual Ontology Alignment. In *Proceedings of the conference on Information Visualization*, IV '06, pages 430–440, Washington, DC, USA. IEEE Computer Society.

Lee, J. (2007). A Computational Model of Text Reuse in Ancient Literary Texts. In Association for Computational Linguistics, editor, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 497–506, New York, NY, USA. ACM.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

OpenBible.info (2012). Bible Cross References Interactive Visualization. `http://www.openbible.info/labs/cross-references/visualization` (Retrieved 2013-06-10).

Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., and Barton, G. J. (2010). Visualization of multiple alignments, phylogenies and gene family evolution. *Nature methods*, 7(3 Suppl):S16–S25.

Ribler, R. L. and Abrams, M. (2000). Using Visualization to Detect Plagiarism in Computer Science Classes. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 173–178, Washington, DC, USA. IEEE Computer Society.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Visual Languages, Proceedings*, pages 336–343.

Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. (2009). *Thematic Cartography and Geovisualization*. Prentice Hall Series in Geographic Information Science. Prentice Hall, 3rd, international edition.

Ware, C. (2004). *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Wattenberg, M. and Viégas, F. B. (2008). The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228.