A Survey of Text Alignment Visualization

Yousef, Tariq; Janicke, Stefan

Published in:
IEEE Transactions on Visualization and Computer Graphics

DOI:
10.1109/TVCG.2020.3028975

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Yousef, T., & Janicke, S. (2021). A Survey of Text Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1149-1159. https://doi.org/10.1109/TVCG.2020.3028975

A Survey of Text Alignment Visualization

Tariq Yousef and Stefan Jänicke

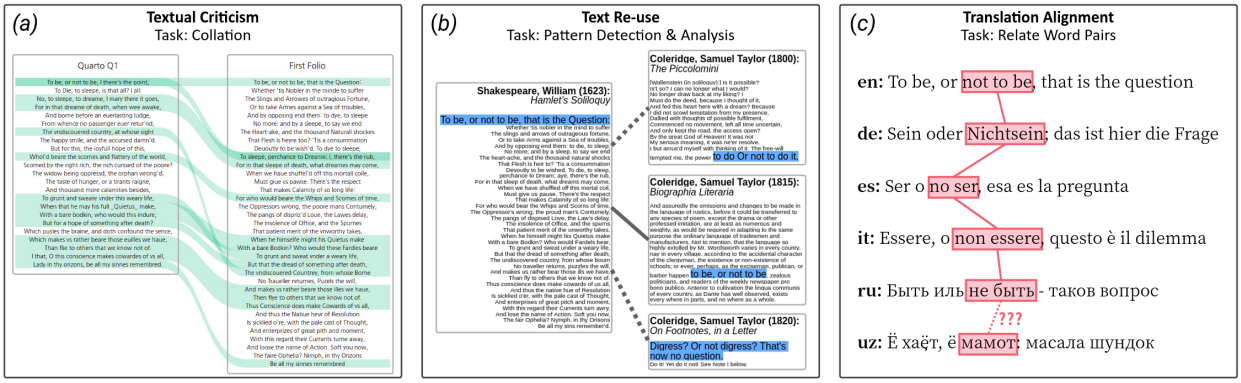


Fig. 1. Three text alignment scenarios focusing on Hamlet's Soliloquy: (a) collation scenario shows alignment of similar or equal lines among the Quarto Q1 and First Folio editions of Hamlet, (b) text re-use scenario highlights variants of the "To be, or not to be"-quote in three different texts of Samuel Taylor Coleridge (the straight line indicates an exact quote, variants are linked with dashed lines), and (c) translation scenario links different translations of "not to be" in the first line of the speech (in Uzbek, "not to be" translates to *death*).

Abstract—Text alignment is one of the fundamental techniques text-related domains like natural language processing, computational linguistics, and digital humanities. It compares two or more texts with each other aiming to find similar textual patterns, or to estimate in general how different or similar the texts are. Visualizing alignment results is an essential task, because it helps researchers getting a comprehensive overview of individual findings and the overall pattern structure. Different approaches have been developed to visualize and help making sense of these patterns depending on text size, alignment methods, and, most importantly, the underlying research tasks demanding for alignment. On the basis of those tasks, we reviewed existing text alignment visualization approaches, and discuss their advantages and drawbacks. We finally derive design implications and shed light on related future challenges.

Index Terms—Text Alignment, Text Visualization, Collation, Text Re-Use, Plagiarism Analysis, Translation Studies

1 Introduction

Alignment refers to the discovery of similar and diverging patterns among two or more data objects. It is a fundamental and widely used technique in many fields, one of which is bioinformatics, in which sequences of DNA, RNA, and proteins are aligned to detect regions of similarity that may be considered as an evidence of structural or functional relationship [69]. A number of works outline the benefit of visualizing sequence alignments [9,23,60,95]. However, text alignment scenarios are different from sequence alignment in bioinformatics; thus, the algorithms as well as the means of visualization vary.

The *first scenario* for text alignment refers to the collation task in textual criticism aiming to investigate textual variation across different versions of a text [99]. The first attempt to (manually) collate texts was the Wimbledon method of collation by using one finger to trace the lines in two texts synchronously to detect differences. In the late 1940s, Charlton Himan invented an optical collator that used strobe lights and mirrors to detect differences between two documents [98]. The process has become much easier with the advent of computers and digitized texts, and many algorithms have been developed for this purpose since 1970. The Needleman-Wunsch algorithm [71] was one of the first algorithms developed to align sequences to find the optimal matching using dynamic programming technique. However, Dekker outlines that the application of such standard sequence alignment methods to the collation task poses technical and methodological problems relating to (1) exchanged text fragments, so-called transpositions, (2) word order independence, and the necessity to (3) flexibly match tokens [50].

The *second scenario* is concerned with the discovery and analysis of re-used text passages within a collection of texts [42]. This oral or written reproduction of textual content is called *text re-use* [29]. Deliberate text re-use appears in the form of direct quotes and phrases like winged words and wisdom sayings. A prominent application of text alignment in this context is plagiarism detection [45], for which a text with unacknowledged, re-used passages is compared with a reference documents database. Text re-use can also be unintended as in the cases of boilerplates, e-mail headers, repetition of news agency texts, the use of idioms, battle cries, etc. Compared to the first scenario, it is not known if different versions of text fragments exist prior to analyzing a text collection. Algorithms are tailored to overcome the challenges of detecting paraphrases, text re-use across languages, or plagiarized ideas [29,48].

The *third scenario* of text-based alignment scenarios is translation alignment, which is a fundamental task in machine translation systems [27]. Text fragments are aligned with their translations at word, sentence, or paragraph level. Such algorithms produce lists of translation pairs that can be re-used in future machine or human translations, or to create dynamic dictionaries and translation memories. First and foremost, the alignment challenges relate to the vocabularies in different languages. This includes not only morphological or syntactic phenomena being hard to align [61], but also the overall aim that related sentences sometimes *only* convey the same meaning [64].

- *Tariq Yousef is with Leipzig University, Leipzig, Germany. E-mail: tariq.yousef@uni-leipzig.de.*
- *Stefan Jänicke is with University of Southern Denmark, Odense, Denmark. E-mail: stjaenicke@imada.sdu.dk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Though the driving forces for developing solutions for the three scenarios are different, the methodological approaches to discover and analyze text alignments partially overlap. Text alignment visualizations are indispensable in all scenarios for targeted users who typically have a background in non-technical domains like the humanities, social and political sciences to make sense of automatically derived patterns. We reviewed 40 visual interfaces that support text alignment analysis aiming (1) to discuss seven state-of-the-art visualization strategies for text alignment, (2) to highlight the uniformity of them across several disciplines, and (3) to reveal the current redundancy of solutions. Drawing from the body of related works, we extract design implications to guide future developments for text alignment visualization and open challenges relating to insufficiently supported research tasks.

2 Scope of the Survey

Alignment in its basic form is the process of finding correspondences between two data objects, and visualization supports the analysis of alignment patterns. The alignment of data fragments has been a fundamental application in several fields, and the approaches to align patterns depend on the data being processed. In bioinformatics, biological sequences are aligned to detect identical positions referring to functional, structural, or evolutionary relationships between the sequences [69]. Many tools in the field support the visual analysis of aligned sequences.[1] Though the sequences can be textually represented, they do not refer to the result of *writing* a text, which our survey focuses on. Two different data types are aligned when video frames of movies are aligned with written text (e.g., subtitles) [22, 79]. Audio-to-text alignment has been used to solve many problems, such as creating training data for automatic speech recognition systems with limited resources [14]. Alignment between texts and images has also been the subject of many research studies. Baraldi et al. [18] developed a semi-supervised approach for aligning commentary texts with miniature illustrations from illuminated manuscripts. Zinger et al. [104] have been working on text-image alignment to align images of words in handwritten lines with their text transcriptions. However, alignment is not limited to textual data formats only. In music, alignment algorithms are tailored to perform audio-to-score alignment linking audio segments of musical performance with their symbolic representation [70]. For analyzing sports data, methods have been developed to align and visualize tracking data collected with different sensors [52], or with other data types, such as humanly defined event data [57].

All these applications share the same principle. Data objects to be compared are segmented into smaller units. If they are represented in different formats, they will be transformed into a middle format; after that, an algorithm (usually dynamic programming algorithm) with a score function will be used to find and align related units. Our survey on text alignment includes related works addressing the alignment of text sources (writings) *and* the visualization of aligned text fragments. This scope definition excludes the alignment of derivatives of texts such as topics [10] or annotations [31, 49]. The alignment of original text fragments poses a significant challenge to the visual representation as the order in which text has been written needs to be preserved.

Survey methodology With the given survey scope, we searched for related tools and publications. As we ourselves are active researchers in this area [54–56, 58, 101–103], we already had a proper body of related works to start with at hand. We additionally consulted the TextVis Browser [62] to extend our collection. We further used Google Scholar to browse visualization, digital humanities and computational linguistics journals and proceedings using related keywords like "alignment visualization" or "collation visualization," which further increased the number of references. Reviewing the related works sections of each paper individually, we traced every cited reference and checked if it fitted our survey scope. In addition, we used the standard Google search to find related translation tools that apply visualization means. The final collection of our survey can be found in Table 1.

[1] https://en.wikipedia.org/wiki/List_of_alignment_visualization_software

Structure of the Survey First, we provide an overview of the three target areas demanding for text alignment visualization in Section 3. What follows is a generalized description of the text alignment process including data and task abstractions in Section 4. Then, we discuss related visualizations that we classified according to the applied techniques in order to highlight the relatedness of the text alignment scenarios and underlying tasks in Section 5. Finally, we discuss design aspects of text alignment visualizations, from which we derive related future challenges in Section 6.

3 Applications of Text Alignment

The following subsections provide comprehensive overviews of text alignment scenarios to support the *collation task* in textual criticism, *text re-use detection & analysis*, and *translation alignment*. We provide details on the scholarly area in which the alignment task is relevant. In order to outline the relatedness as well as the differences among the three text alignment scenarios, we explain the alignment task focusing on Shakespeare's *Hamlet* (see Figure 1).

3.1 Scenario 1: Collation in Textual Criticism

Textual criticism is a field of the humanities concerned with studying how texts have been created, distributed and disseminated. At the end of a text critical process traditionally stands a critical edition of a literary (or non-literary) work that reconstructs the original version of the text based on its text variants. However, the accessibility to numerous digital text variants expedites quantitative comparative analyses of textual variants.

Application Context One of the major sub-tasks of a textual scholar in the traditional text critical process is examining and recording the similarities and the differences among a number of variants of a text. This practice is known as collation [89]. The scholar selects a number of text versions to be compared, and arranges these side by side. Unavailable text transcriptions need to be manually transcribed, and variations among the versions like omissions, additions, substitutions and word or phrase order changes (transpositions) need to be annotated. The more manuscripts are observed in the text critical process, the more complex and laborious the collation task becomes. Automated text alignment takes over this time-consuming, error-prone task by quickly collating a huge number of variants with a high precision. As computers are neither able to analyze and interpret the collation results, nor they are able to assess their quality, visualization frameworks are required to serve scholars with a visual depiction of the automatically determined alignment patterns. Our survey indicates that automated collation and visualization tools support the traditional task of creating a critical edition on the one hand [41, 56], and quantitative studies investigating the collation result of a large quantity of text variants on the other hand [33, 84].

Alignment Specifics Many works in this application scenario address historical text fragments, which poses different challenges to be tackled by automated alignment algorithms. Text variants might be written in different time periods, thus, in different dialects, which makes it harder to find related text passages. This is even more crucial if the original text has been orally transmitted [58]. The written versions might use different diacritics and special characters, and they might contain errors due to inaccurate OCR processes during the digitization process. Detecting textual transpositions, which frequently appear in historical texts, is a big challenge and requires sophisticated approaches. Alignment algorithms tailored to support collation tasks typically work at word level, using dynamic programming algorithms like Needleman-Wunsch or Smith-Waterman to find the best alignment. Exact matching or probabilistic matching approaches (e.g., Levenshtein distance) are applied to determine the similarity of words. Refinements, e.g., ignoring punctuations and diacritics, are applied, and working on lower-cased text increases the quality of automated collation results.

Shakespeare's *Hamlet* & Collation Shakespeare's works are prominent examples for textual criticism. None of his printed works were actually written down by Shakespeare himself, as he wrote his

plays for theater performances rather than publications [39]; Elizabethan playwrights did not care about printed editions of their plays. In order to make profit, Shakespeare's plays were sometimes written down from memory by actors, or reconstructed from notes taken during performances. Acting companies owning his works also produced copies for potential publications and to preserve the integrity of a play. Thus, many different printed editions, so-called *Quartos*, circulated even before the first official printed edition—the *First Folio* (F1)—of his works entitled "Mr. William Shakespeare's Comedies, Histories, & Tragedies" was published in 1623, seven years after his death. For example, the F1 version of *Hamlet* is believed to be the most authoritative edition, although compiled by Shakespeare's fellows after his death. The First Quarto version of *Hamlet* (Q1), probably written down by one of Shakespeare's actors, is seen as less authoritative due to the less beautiful and less detailed language [90]. Figure 1a illustrates the collation of Hamlet's Soliloquy among the Q1 and F1 versions visually. One can see that there is a rough skeleton of authoritative lines mixed with non-authoritative ones. In addition, the Q1 version is much shorter compared to the F1 version. Visual depictions such as this prepare the ground for quantitative comparative studies of textual variation.

3.2 Scenario 2: Text Re-use Detection & Analysis

"At its most basic level, text re-use is a form of text repetition or borrowing." [46] Consequently, this first sentence itself is a textual reuse, which is in general subject to a large variety of applications.

Application Context Re-using fragments of text in form of quotations is very common in literary, political, historical and religious texts. Text re-use is a broad term that includes many stylistic text features such as paraphrases and allusions. Summaries are also forms of text re-use as the summary is derived from the original full text. The translation of a text fragment into another language is considered a cross-lingual text re-use. The field of journalism is a prime example for text re-use as news articles typically contain a huge amount of repetitions. Fetterly et al. [44] found that around one third of web pages are similar to other web pages, and that around one fifth are identical. A thin thread exists between text re-use and plagiarism, similar to the one between borrowing and stealing. In plagiarism, the scholar hides the fact that the text is borrowed and claims ownership, which contrasts to quotations where the scholar mentions that a reused text fragment is borrowed from a specific resource. Visualizations are important to analyze occurring text re-use patterns that can be indicators of plagiarism.

Alignment Specifics Automatic text re-use systems work on a high text hierarchy level as they compute the similarity at sentence or paragraph level. They apply various text similarity measures such as n-gram overlap on character or word level, TD-IDF, relative frequency or query likelihood [21, 37, 68] to assign a score to each compared pair of sentences or paragraphs. Bär et al. [30] proposed a method to overcome the limitations of traditional similarity measures by computing similarity along the three characteristic dimensions content, structure and style, which are inherent to texts. By nature, historical texts contains lots of re-used fragments, and several frameworks have been developed to extract those patterns automatically [29, 86], thereby addressing the problem of faulty data sources and limited language resources.

Shakespeare's *Hamlet* & Text Re-use Shakespeare is one of the most quoted authors [66]. Prominent representatives for frequent usage of *Hamlet*-quotations are Edgar Allan Poe, Charles Dickens and Walter Scott [93]. Around 2,500 Shakespeare quotes have been found in the works of the essayist William Hazlitt, of which 500 are attributed to *Hamlet*. Quantitative text analysis methods are tailored to discover re-used text fragments automatically [29]. They take an entire text collection consisting of a source text, e.g., *Hamlet*, and reference texts that potentially contain quotations, e.g., the works of William Hazlitt. The source text is split into fragments like sentences, and for each of them similar or equal patterns are recognized in the reference texts. The project HyperHamlet [74] offers to browse through around 9,000 different quotations of Hamlet. Figure 1b illustrates three different variants of "to be or not to be" re-used in works of Samuel Taylor Coleridge. If quotations are made explicit, they are rather easy to find, but many works also contain unintended text re-use [66]. This also pictures the relation to plagiarism, for which re-used text passages are not made explicit on purpose. Using plagiarism software, McCarthy and Schlueter [63] discovered that Shakepeare's works themselves are strongly influenced by previous works.

3.3 Scenario 3: Translation Alignment

As opposed to the previous scenarios, translation alignments exclusively entail cross-lingual relations between words and sentences. Many people get in touch with visual depictions of multi-lingual translations, offered to the broad public in the form of online-translators.

Application Context Aligned multilingual translations are valuable resources for professional translators, and second languages learners as they offer, in contrast to dictionaries, valid cross-lingual relations not only on word-, but also on phrase- and sentence level [61]. Contrastive linguistics, a field that systematically compares languages in order to investigate similarities and differences among them, is another section that demands multilingual translations.

Alignment Specifics The alignment between a base text and its translations can be at word, sentence or paragraph level, for which several statistical approaches have been developed [92]. However, translation alignment can be done also manually using online editors to create accurate training data [65, 67, 101], which is considered a main component for statistical machine translation systems [26] and probabilistic bilingual lexica [17]. Applying sophisticated visualization approaches to show the translation alignment is essential as they help analyzing translation equivalents, exploring their usage contexts, and disambiguating word senses. The first visualization of aligned bilingual texts was introduced by the *Loeb Classical Library*,[2] which published important works of ancient Greek and Latin literature in bilingual format. The original texts were placed on the left-hand page, whereas the literal translations were placed on the right-hand page.

Shakespeare's *Hamlet* & Translation Alignment Due to Shakespeare's popularity, his works have been translated into more than 100 languages; Hamlet is available in more than 75 different languages. This high diversity of different translations is a valuable resource to support translation alignments. The hierarchical structure of Shakespearean plays makes it easy to align acts, scenes as well as the ordered speeches of characters. For a more granular alignment, translations in well-resourced languages can be triangulated with a translation in an under-resourced language [38]. Figure 1c exemplifies this with the *to be, or not to be* line in six different languages amongst which are the well-resourced languages English, German, Spanish, Italian and Russian, plus Uzbek considered an under-resourced language. It is likely that only few parallel texts between German and Uzbek exist, on the other hand, a higher number of parallel texts between Russian and Uzbek as well as German and Russian are available. Thus, Russian can be utilized to align the German with the Uzbek Hamlet translation, moreover, to increase cross-lingual resources valuable for a diversity of purposes [7, 35].

4 Text Alignment Process

Text alignment is the process of comparing two or more texts aiming to extract similar patterns and link them, or to identify differences among texts. The input sources for a text alignment process are usually text files organized in parallel format. In the following, let $T_1, ..., T_n$ be a set of mono- or multilingual text documents to be aligned.

With the *Gothenburg model*,[3] developers of Collatex [40] and Juxta [59] have defined a baseline model for the automated alignment of textual editions widely applied in the digital humanities. Though it has a specific focus on text variants, it can be generalized well to all related text alignment scenarios discussed in this paper. The model describes text edition alignment in five successive modules, from which we derive a three-step model for the visual analysis of text alignment in general as illustrated in Figure 2.

[2] https://www.loebclassics.com/

[3] https://wiki.tei-c.org/index.php/Textual_Variance

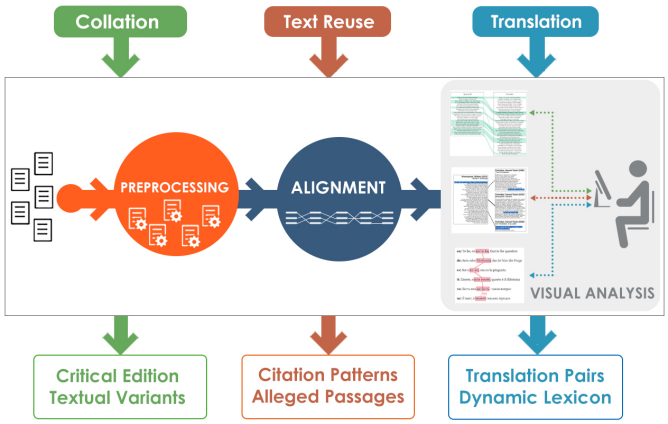


Fig. 2. Visual analysis of text alignment for three scenarios.

Pre-processing: In order to prepare the text sources $T_1, ..., T_n$ for the alignment process, they need to be split into tokens to facilitate measuring the similarity between text fragments. Typically, a text T_i is split into words that form textual units $\{t_{i1}, ..., t_{im}\}$ for which the similarity is computed, and that are later on aligned and visualized. Textual units can be chapters, paragraphs, sentences, lines or single words. Furthermore, textual units might be hierarchically structured to support aligning texts on different hierarchy levels. A further pre-processing step is text normalization aiming to reduce alignment errors. It can include transforming all tokens into lowercase characters, removing punctuation characters, or stemming. In addition, for some scenarios it is beneficial to mask stopwords, if a stopword list is available. This can help focusing on the rather meaningful parts of the input texts in the subsequent alignment computation.

Alignment: The alignment process can be seen as a black box that takes the text units of all documents $T_1, ..., T_n$ and returns a list of aligned text units. Typically, dynamic programming algorithms Needleman-Wunsch [71] or Smith-Waterman [85] are used to perform the alignment in conjunction with score functions and refinement criteria. Hidden Markov Models are also performant in this context [43]. However, in some cases the alignment is also performed manually [19]. For each pair of texts T_i and T_j, the alignment process delivers K tuples of aligned text units:

$$ALIGN(T_i, T_j) \longrightarrow \{_{k=1}^{K}(t_{ix}^k, t_{jy}^k)\}$$

This result is achieved in two main phases. The first phase includes calculating a similarity score $S(t_{ix}^k, t_{jy}^k)$ between all possible combinations of textual units. The similarity criteria to evaluate the relatedness between compared units depends on the underlying task. The second phase is an optimization in which the best combination of similar units is chosen:

$$OptimalAlignment = max(\Pi_{k=1}^{K} S(t_{ix}^k, t_{jy}^k)).$$

Visual Analysis: The alignment procedure delivers sets of tuples representing related text units. In order to convey a concrete impression of how $T_1, ..., T_n$ are related, alignment visualization solutions play an important role. They provide a visual depiction of alignment patterns to domain scholars, allowing them to analyze similarities and differences among the compared texts. The visual analysis often also allows for scholarly feedback, i.e., automatically determined alignments are examined and might be modified according to the knowledge of the scholar. The visualizations are typically tailored dependent on the underlying user task that demands for visually analyzing textual alignments. Dependent on the scenario, Figure 2 illustrates the resources that can be gained from this process.

Table 1. Surveyed Tools & Visualization Techniques

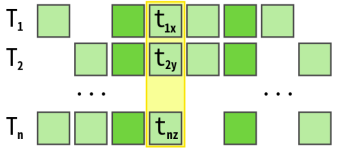
Task	Tool	Sequence Heat Map	Grid Heat Map	Aligned Barcodes	Text Heat Map	Side-by-side	Tabular View	Variant Graph
Collation	TexTile [15]	✓						
	ViTA [6]		✓	✓				
	NMerge [80]							✓
	TRAVIZ [56]							✓
	DRVG [54]	✓						✓
	Storylines [84]							✓
	WordGraph [76]							✓
	Word Tree [96]							✓
	Stemmaweb [13]						✓	✓
	CollateX [40]						✓	✓
	PyCoviz [72]						✓	
	eComparatio [28]					✓	✓	
	iAligner [103]	✓				✓	✓	
	LERA [82]	✓				✓		
	CATView [73]	✓				✓		
	JuxtaCommon [59]				✓	✓		
	EVT [78]				✓	✓		
	Versioning Machine [81]					✓		
	ShakerVis [47]					✓		
	Variance Viewer [5]					✓		
	Iteal [58]			✓		✓		✓
	Baumann et al. [19]			✓		✓		
	Alignment maps [34]			✓		✓		
Text Reuse	PicaPica [77]			✓		✓		
	Text Re-Use Browser [55]		✓	✓		✓		
	Kitab Project [2]			✓		✓		
	GuttenPlag [16]			✓	✓			
	Chinese Text [88]				✓			
	Text Re-Use Grid [55]		✓					
	PatternGramm [75]	✓						
	News Auditor [20]					✓		
Translation	Ugarit [101]					✓		
	Alpheios [12]					✓	✓	
	Japanese-Chinese [36]		✓					
	Linguee [3]					✓		
	Glosbe [1]					✓		
	Reverso Context [4]					✓		
	Divan Hafez [102]					✓		
	Ducat [87]					✓		
	Translation Arrays [32]					✓		

5 Text Alignment Visualizations

This section discusses the different visualization techniques supporting the three application scenarios described in Section 3. A classification of visualization techniques embodied in the 40 alignment tools preparing the basis for this survey is given in Table 1.

5.1 Sequence-aligned Heat Maps

Sequence-aligned heat maps serve the purpose of conveying a summarized overview of alignment patterns among different source texts, and they are typically used to navigate to text units of interest.



Visual Representation Each text T_i from the corpus $T_1, \ldots, T_n$ is represented by a sequence of colored rectangles. A rectangle stands for a text unit, which is typically a sentence, paragraph, chapter, or even a larger text fragment. The text units are ordered as they appear in the text, and aligned text units are placed on the same vertical axis. Vertical sections with only one cell indicate that for the corresponding text unit no match among the other texts exists, whereas a gap in sequence i denotes that the corresponding text T_i does not contain a variant to matching text units of other texts. Sequence-aligned heat maps are usually combined with other text alignment visualization that show details on demand, i.e., that allow for close reading the compared texts. The grid size plays an important role, the more text units and the higher the variation, the less space can be reserved for a vertical section.

Applications Sequence-aligned heat maps are often applied to navigate through different text variants to support the *collation* task [73, 103]. The framework LERA [82] applies a coloring scheme to indicate

different degrees of variation; the lighter the color of a rectangle, the more similar the corresponding segment is to the other parallel segments (see Figure 7). Two works employ the sequence-aligned heat map to different text hierarchy levels [15, 54]. For example, TexTile allows users to dynamically inspect variation among sequences generated for the entire text-, page-, line- and word-level as shown in Figure 3. A five-level univariate color scheme is applied to reflect similarity or distance between a variant and its parallel segment in the reference text. Next to collation, sequence-aligned heat maps have also been used to visualize detected plagiarised text passages among a base text and a set of documents using the Composite Categorical Patterngram (CCP) [75]. Different colors are used to distinguish between base file (blue dots), frequent patterns (green), and infrequent matched patterns (red).

Strengths & Weaknesses Being a textual abstraction, sequence-aligned heat maps make alignment patterns on word, sentence, and paragraph level perceivable, but they need to be combined with a text-based representation to read aligned fragments. The visual representation works well for less-varying texts; however, transposed fragments increase the number of columns and white space.

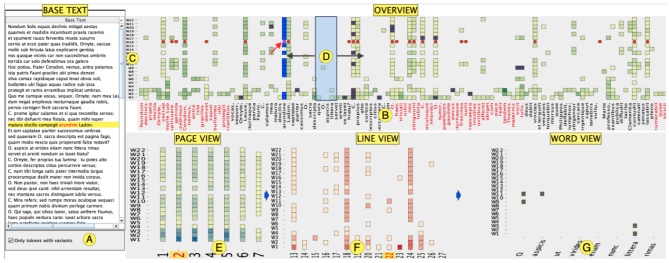
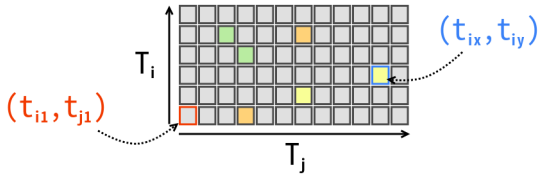


Fig. 3. TexTile variation visualization at different text hierarchy levels [15].

5.2 Grid-based Heat Maps

Grid-based heat maps illustrate alignment patterns among two texts of a corpus, which allows drawing conclusions on the type of textual variation among them.



Visual Representation Each cell of the two-dimensional grid represents a juxtaposition of two text units t_{ix} and t_{jy} of two arbitrary texts T_i and T_j of the text corpus $T_1, \ldots, T_n$. Cells might be evenly sized, but they can also visually reflect the lengths of the corresponding text units, e.g., by rectangular cells. Typically, color is used to inform on particular features of the relationship between t_{ix} and t_{jy}, e.g., alignment types or matching scores. Typically, clicking a cell opens a closer look at the juxtaposed text units.

Applications Grid-based heat maps are applied to different text hierarchy levels. In two applications, a cell, or a dot, represents a match among two text variants on word level, so that diagonal patterns illustrate similarity in text flow [6, 36]. The Text Re-use Browser [55] uses a dot plot matrix similarly to highlight patterns of repetitive or systematic text re-use. A dot represents a pair of similar sentences, and the color of the dot reflects the matching score. The Text Re-use Grid [55] provides alignment information for an entire document collection. A cell stands for a pair of processed documents, and the color informs on frequency and type of occurring text re-use patterns. Figure 4 provides examples for both Text Re-use Browser and Grid.

Strengths & Weaknesses A limitation of this technique is that only two text sets can be chosen for comparison. However, the design is applicable to all text hierarchy levels—even alignments on document level are supported. The resultant matrix typically contains a large amount of white space and redundant information. However, the arrangement of glyphs makes alignment patterns salient.

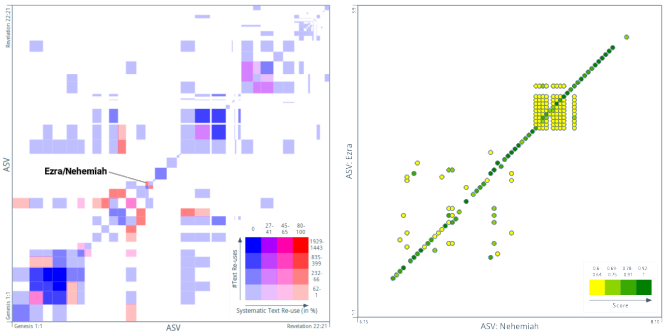
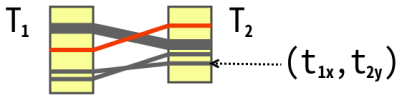


Fig. 4. Text Re-use Grid and Text Re-use Browser illustrate text re-use patterns at different text scales [55].

5.3 Aligned Barcodes

GuttenPlag [16] proposed a *Barcode* visualization to highlight the amount of plagiarised text per page using colors. In contrast to such a singular barcode that only visualizes how passages of one text are reused, aligned barcode visualizations communicate the relations to reference documents by displaying alignment patterns.



Visual Representation Aligned barcodes are typically restricted to illustrating the alignment patterns found among two texts. The sizes of the text boxes representing T_1 and T_2 represent the lengths of the texts or the number of text units. Each alignment tuple (t_{1x}, t_{2y}), typically on the level of a sentence, is drawn in the form of a line string connecting the relative positions of the related text fragments in T_1 and T_2. Displaying all alignment tuples that way results in alignment patterns being visible. Next to line strings, rectangular areas can be used to illustrate the alignment of entire paragraphs [34]. Also, color can be used to highlight different types of alignments [58].

Applications Aligned barcodes often appear as overview visualizations in addition to side-by-side views in collation [6, 58, 59] or translation scenarios [34]. They direct the user to interesting patterns of consecutive alignment units or transposed passages, and the aligned barcode can be used to navigate through the texts; an example of this scenario is shown in Figure 8. Aligned barcodes are also used to analyze text re-use patterns [2]. The Text Re-use Browser uses color to indicate the matching score of an aligned text fragment, and the occurring pattern informs on repetitive or systematic text re-use. Picapica [77] applies an aligned barcode to visualize plagiarism using T_1 for the observed text and the box area for T_2 for the entire corpus of reference texts, which makes all plagiarised fragments explorable.

Strengths & Weaknesses Aligned barcodes are text alignment abstractions. That makes them applicable to long texts, but they require always a linked text-based representation to expedite knowledge discovery. They are limited to comparing two text variants, typically on line, sentence or paragraph level. This makes patterns of subsequent aligned text fragments as well as transpositions easy to spot. Nevertheless, aligned barcodes can be hard to read with an increasing number of transpositions that occlude due to crossing lines.

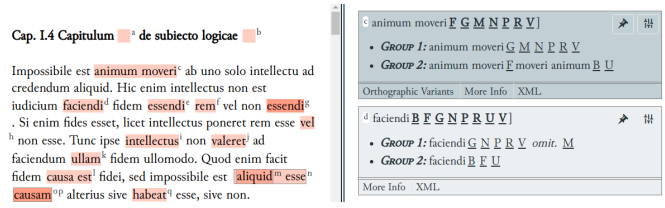


Fig. 5. Edition Visualization Technology (EVT) text-oriented heat map visualization. Variance positions are highlighted and the critical apparatus is placed on the right side [78].

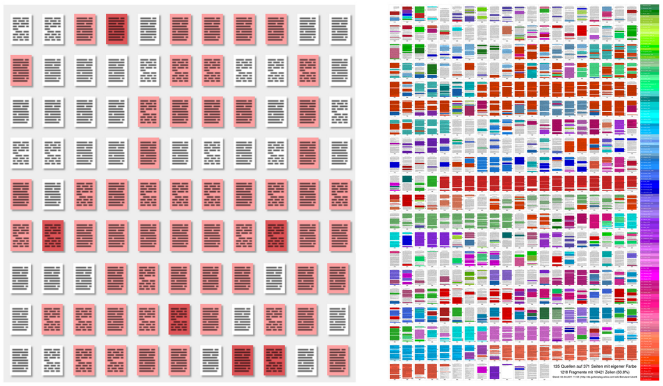
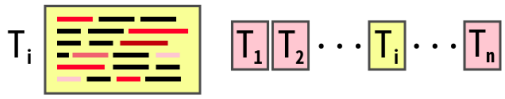


Fig. 6. GuttenPlag text-oriented heat map visualization. Plagiarism positions are highlighted with different degrees of red color reflecting the amount of plagiarised text [16]; images licensed under CC-BY-SA.

5.4 Text-oriented Heat Maps

Yet having numerous source texts in the underlying corpus, text-oriented heat maps focus on a singular text overlaid with alignment information.



Visual Representation The focused text T_i is selected from the corpus $T_1, \ldots, T_n$ and the text (flow) of T_i gets magnified. Typically, words or phrases receive a colored background indicating how often they appear in the other texts of the corpus, or how strongly they vary. Single hue color maps [51] are used to differentiate between low and high frequency.

Applications JuXtacommons [59] and Edition Visualization Technology [78] use this technique to illustrate textual variations among multiple witnesses. They both utilize different degrees of saturation to display different levels of variance from the base witness T_i, the more saturated, the more witnesses differ (see Figure 5). Similarly, text-oriented heat maps can display how often a text fragment has been re-used [88]. GuttenPlag [16] applies this technique to highlight the page-wise amount of plagiarised text passages; pages colored in saturated red contain a large amount of re-used text fragments, as shown in Figure 6.

Strengths & Weaknesses The visual output is always generated for an individual text in focus. Variance to the other text editions are summarized and projected on the focused text, so that individual relations to other texts are only conceivable through detailed views. Moreover, relations among other text editions remain hidden. However, text-oriented heat maps produce an excellent picture of variation on word, phrase and paragraph level, and the focused text can be easily followed.

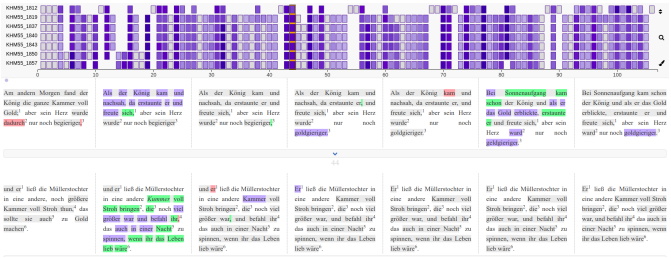


Fig. 7. LERA side-by-side view combined with a sequence-aligned heat map to facilitate navigation through the texts [82].

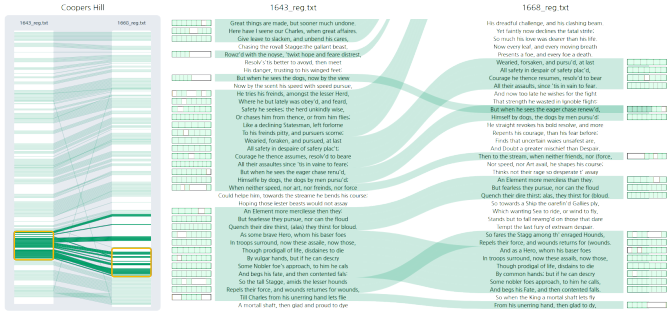
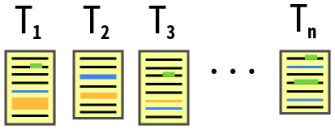


Fig. 8. Iteal combines aligned barcodes with a side-by-side view [58].

5.5 Side-by-side Views

Side-by-side visualizations are the most commonly used technique to display aligned texts at different levels. It locates the compared texts alongside each other and aligned units, which can be paragraphs, lines, sentences, phrases, or words, are highlighted. A navigation bar or aligned barcodes are usually associated with the side-by-side model to facilitate the navigation through the aligned units in case of long texts.



Visual Representation Side-by-side views reserve a rectangular space to display the text of different aligned editions $T_1, \ldots, T_n$ next to each other. Alignment tuples are not explicitly drawn in the form of connections. Typically, words, or their backgrounds, are colored to indicate—dependent on the use case—matches or variance among editions. The user needs to scan the parallel texts to identify positions of variance. Interaction means like mouseover are used to relate words with each other; thus, individual alignments can be shown on demand. Different colors are used to indicate different types of alignments.

Applications Side-by-side views are the most commonly used technique to visualize *collation* results, very often limited to two text variants being compared [5, 28, 58, 59, 103]. Versioning Machine [81] is one of the earliest tools proposed to visualize more than two TEI-encoded texts; the first version was launched in 2002. While it provides hover functionality to relate passages simultaneously, a coloring scheme to differentiate alignment types is not implemented. LERA is also capable of juxtaposing more than two variants [82]. It colors inserted text fragments in green, substitutions in blue, and deletions in red. Figure 7 shows an example of seven juxtaposed variants of the German fairytale *Rumpelstilzchen*. Edition Visualization Technology also uses color to highlight distinct typology, but further apply color saturation to indicate the variability of a tradition [78]. As for *collation*, side-by-side views also support analyzing text re-use patterns between two arbitrary texts of a corpus [2, 55]. The color of text in those scenarios only indicates the re-used fragments. Side-by-side views are the means of

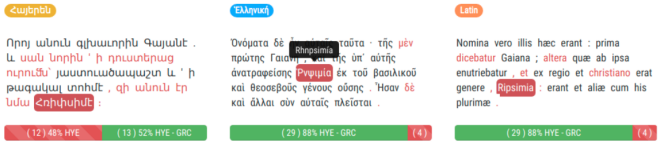


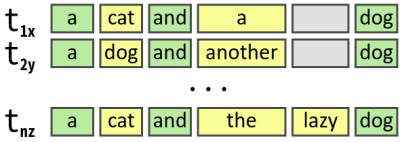
Fig. 9. Ugarit side-by-side translation visualization, aligned words in Armenian, Ancient Greek and Latin text fragments are highlighted [101].

choice to visualize translation alignments. Tools such as Ugarit [101], Linguee [3], Glosbe [1] and Reverso Context [4] provide bilingual *KeyWord-in-Context (KWIC)* search functionality. The results of a keyword search are the translation and juxtaposed contexts in source and target language, in which the keyword appears; the word and its translations visually stick out (see Figure 9). Such word- or phrase-level alignments are also used in academic contexts [12, 87, 102]. Ugarit uses color to discriminate aligned from unaligned tokens [101], and hover functionality is used to highlight related tokens. Cheesman et al. [34] apply side-by-side views to explore translation alignments among different translations of Shakespeare's *Othello*, by juxtaposing two texts, but also proposing a layout for base text and two translations.

Strengths & Weaknesses A major constraint for side-by-side views is the limited screen width, which makes them applicable for a limited number of text variants. Users might also get distracted since they have to move the eyes between parallel texts to identify positions of variance. However, in side-by-side views it is easy to read larger text fragments of a variant. Further, they allow inspecting variance on different text hierarchy levels, i.e., word, sentence, and paragraph.

5.6 Tabular Views

Tabular views visualize word-level variations for small text fragments such as sentences or short paragraphs.



Visual Representation Represented in the form of a table, the rows represent different aligned text units $t_{1x}, t_{2y}, \ldots, t_{nz}$. The columns stand for aligned tokens. Typically, columns are colored in order to show similarity or variation. Cells might be empty if a text unit does not contain a variant reading for a token.

Applications Tabular views are widely used to visualize text variants at word level [28, 72, 103]. An example of turning six variants into a tabular view with CollateX [40] is shown in Figure 10. The *Interlinear Text* view offered by Alpheios [11] employs the method to visualize translation alignments. Alpheios displays each word in the source text combined with its translation displayed under it while assigning a red color to unaligned tokens.

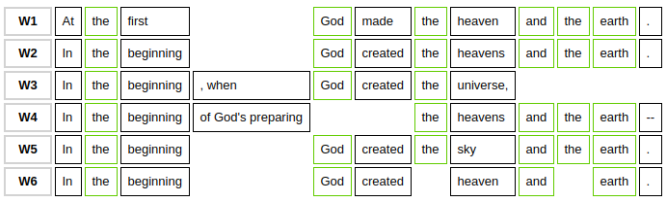
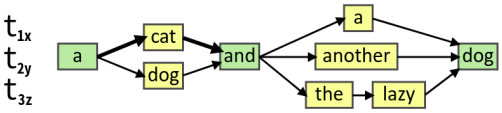


Fig. 10. CollateX tabular view visualization for six different English translations of the first Bible verse. Aligned tokens are placed at the same column green color reflects a total agreement among text variants [40].

Strengths & Weaknesses Using tabular views, it is easy to spot aligned patterns on word and phrase level. In addition, a text variant, for which an entire line is reserved, can be easily followed. Though this representation works well for less-varying text fragments–although similar patterns induce increased space coverage through redundant information–, transposed passages quickly disperse the view as the number of columns and the amount of white space increase. In general, the screen size limits the number of text variants that can be compared. Further, tabular views only work at word level and produce readable representations for short sentences.

5.7 Text Variant Graphs

Similar to tabular views, text variant graphs illustrate word-level variations for small text fragments, but, employing a graph structure to represent variation enables visualizing additional alignment features.



Visual Representation Schmidt and Colomb [80] were the first using a graph to visualize textual variance for multi-version texts. They proposed the *Variant Graph*, which is a directed acyclic graph (DAG) with nodes typically representing words or tokens of text units being aligned. An edge in the graph stands for a pair of subsequent tokens in any of the text variants. Thus, an edge represents a version, and each text variant contributing to the graph defines a path in the variant graph. The size of nodes may reflect how often a word appears among the editions. Likewise, the thickness of an edge can display the number of variants including the corresponding word pair. Edges can be labeled or colored to inform on the corresponding variants.

Applications Among all visualization techniques discussed, the text variant graph is the only one specifically tailored to communicate a *collation* result of small text fragments. CollateX [40] adopted the concept to illustrate the results of their collation alignment algorithm. As edges are labeled, the graphs typically have a large width, even for short text fragments. Transpositions are shown with dashed lines. Stemmaweb [13] extended the idea proposing an interactive version of the graph to support stemma analysis. TRAViz [56] applies diverse visual features to communicate features typical for variant graphs aiming to make them easier to digest (see Figure 11). This includes sizing nodes according to frequency, coloring instead of labeling edges, and a sophisticated graph drawing algorithm that aligns related variants vertically. WordTree [96] and WordGraph [76] employ similar visual cues, but they are not particularly designed to support the collation task. Silvia et al. [84] applied the storyline metaphor to visualize variations in classical texts. The model uses a force-directed layout algorithm to generate a topologically representative layout, which enables users to read and recognize text patterns with reasonable speed and accuracy; an example is shown in Figure 12.

Strengths & Weaknesses In contrast to tabular views, variant graphs reduce redundant information by merging equal or similar tokens. This strategy makes it easy to spot likewise divergent and similar sections among the variants. The visual output eases to create critical editions as text variant graphs can be seen as automated collation results. The variant graph design's backside is that individual variants are not salient, typically highlighted after user interaction. As tabular views, variant graphs currently only work on word level, at best with short, less-varying text fragments.

Fig. 11. TRAViz graph showing variance among six English translations of the first Bible verse [56].

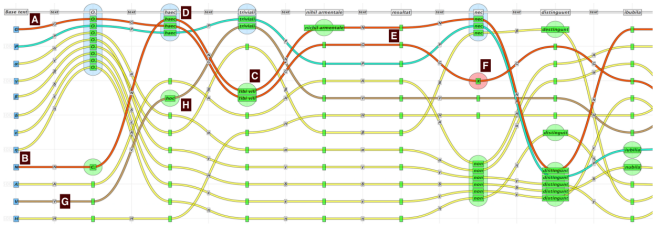


Fig. 12. Storyline-based text variant graph [84].

5.8 Miscellaneous Techniques

Next to the techniques mentioned above, a few related works favor other methods to support the visual exploration of text alignment patterns. The Chinese Text Project [88] uses a network graph to visualize the textual relationships in pre-Qin and Han texts. Each node represents a document in the corpus and edges indicate a similarity between documents. Edge thickness reflects the amount shared (re-used) text fragments between two nodes. ShakerVis [47], developed to explore segment variations among German *Othello* translations, applies a parallel coordinates plot and linked scatter plots to illustrate how different translations relate to each other.

5.9 User Interactivity

Alignment tools typically provide means of interaction to configure the alignment computation and to manipulate and analyze the result's appearance. We distinguish pre- and post-alignment interaction.

Pre-alignment User Interactions While users can typically upload their own texts to be aligned, these interactions are rather configurations that influence how alignment algorithms will operate. For instance, users are enabled to adjust if the alignment algorithm will be sensitive to punctuation, capitalization, or diacritics or to ignore non-alphabetical tokens or line breaks [59, 82, 103]. Other tools allow users to choose between exact or fuzzy word matching based on the Levenshtein distance [56, 58, 103]. Further, some tools are equipped with various alignment algorithms the user can choose from [6, 40, 56].

Post-alignment User Interactions Such interactions relate to tasks users can perform once the alignment is computed and visualized. Post-alignment interactions span the entire range of Brehmer's task taxonomy [25]. The central task for which alignment visualizations are used is to *compare* different text versions. A variety of user interactions are offered for this purpose, e.g., *search*ing the alignment output or *filter*ing the results [59, 82, 101]. Standard mouse interactions are typically reserved to highlight specific alignment patterns and to indicate similarity [12, 56, 59, 81, 101], or to demand detailed information or statistics [55, 59, 78, 82, 101]. Some tools offer alignment visualization at different scales alongside with multifaceted means to *browse* and *explore* the alignment result. Typically, individual alignments can be *select*ed, and a fine-granular visual encoding is shown [12, 28, 55, 58, 59, 82]. In other tools, visual components can be rearranged by the user, e.g., by moving adjacent panels triggering an update of the visualized alignment pattern [81], or by switching between horizontal and vertical tabular views [73]. As alignment results might not meet users' expectations, some tools allow to *manipulate* the alignment interactively. For instance, in variant graph visualizations, it is allowed to split and merge nodes [13, 56]. In text-oriented heat map tools, users are able to *change* the base witness [59, 78]. Juxtacommons [59] supports users to *annotate* notes to variance positions in the heat map. Lastly, many tools support the *output* task by offering to export or download an alignment result for further usage, mostly in XML or JSON format [5, 12, 13, 59, 82, 101]. Some tools let users share their visualization via URL [59], or the visual output can be explicitly downloaded as image [13].

6 Design Implications

Visualization of textual alignment plays an important role to facilitate understanding and interpreting the relation between compared texts. We studied 40 tools and classified them according to the task and visualization techniques they offer to analyze occurring alignment patterns (see Table 1). Concerning different features of an alignment scenario, listed in Table 2, we classified the visualization techniques according to applicability (green fields), potential applicability (yellow fields), and inapplicability (red fields indicate future challenges). Respective fields are colored grey if the technique is by design not applicable to the alignment task. For adopting one of the discussed techniques or innovating a new visualization model for textual alignment, one should consider the factors listed below.

Alignment Task The selection of the appropriate visualization technique is affected mainly by the underlying task. In general, collations focus on varying text fragments and require highlighting variance patterns. For instance, variant graphs and text-oriented heat maps merge common parts of the compared texts easing to spot differences. However, text re-use and translation alignment rather focus on common patterns, which should be reflected in the visualization output, for example, by connecting related passages with lines. One can also consider switching between highlighting similarity or dissimilarity [54]. In order to indicate an alignment relation, the best practice is to use the same color for related units [5, 59, 81]. Another strategy is to merge common text units into one visual representation [56, 59, 78]. Other techniques use connecting paths between related text passages [2, 6, 34, 40, 58, 77].

Number of Text Variants In most of the scenarios we observed, the number of compared text variants is below ten. Only heat map visualization techniques can handle a larger number of variants. Two main strategies apply. On the one hand, a single variant is focused, and text-oriented heat maps are used for visualization. On the other hand, a heat map is used to generate an overview, and a subset of variants can be analyzed in more detail using a text-based visualization mode.

Text Hierarchy Levels Texts inhere different structures, in which fragments are organized, e.g., page and line or chapter and sentence. Before tailoring a new or choosing among existing solutions, one should clarify to what extent alignment patterns should be explorable at different scales. Appropriate visual encodings need to be selected in such a case. For instance, *iteal* uses aligned barcodes to inspect occurring patterns within the whole document, a side-by-side view for line-level alignments in sections, and variant graphs for analyzing word-level alignments for lines [58]. For navigation purposes, Shneiderman's information-seeking mantra [83] *"Overview first, zoom and filter, then details-on-demand"* should be implemented, i.e., different visualization techniques should be linked as coordinated views to enable users exploring the corpus smoothly.

Language All discussed visualization techniques are applicable to a monolingual corpus. Although most scenarios deal with left-to-right Latin scripts, tools are also applicable to right-to-left scripts [12, 53, 101, 102] or writing systems without word boundaries [36]. However, not all visualization techniques are applicable to multilingual scenarios.

Text Stability Less-varying texts give users more freedom to select an appropriate visualization technique. However, the more transpositions occur, the less meaningful the visualized results also get for the techniques showing green fields in Table 2. When dealing with unstable texts, we recommend to apply various means of interaction, i.e., allowing to filter alignments according to a certain matching score or to hide all transposed alignments.

7 Future Challenges

Drawing from our own experience in developing text alignment visualizations and the open challenges highlighted as red fields in Table 2, we derived numerous future directions listed below.

Table 2. Applicability of Visualization Techniques to Corpus & Task-related Features

Corpus & Task-related Features		Sequence HM	Grid HM	Aligned Barcodes	Text HM	Side by Side	Tabular View	Variant Graph
Tasks	Collation	✓	✓	✓	✓	✓	✓	✓
	Text Re-use	✓	✓	✓	✓	✓	(✓)	(✓)
	Translation Alignment	(✓)	✓	✓		✓	✓	
Languages	Monolingual	✓	✓	✓	✓	✓	✓	✓
	Multilingual	(✓)	✓	✓		✓	✓	
Alignment Levels	Word	✓	✓	✓	✓	✓	✓	✓
	Sentence/Passage	✓	✓	✓	✓	✓	(✓)	
	Document		✓					
Corpus	scalable to n texts	≈ 50	≈ 50	2	No Limit	≈ 10	≈ 20	≈ 20
	unstable texts		✓	✓	✓	✓		

The Need for a Modular System Drawing from the overview in Table 2, one can see that for each of the possible features already at least one visualization exists that can be applied. Moreover, our observations when preparing the survey exposed a high number of redundant tools solving one and the same issue. This indicates that research among the visualization, NLP, and digital humanities communities is not well known. However, a modular system applicable to any related text alignment scenario would be an appropriate strategy to prevent this issue. Based on the text alignment process in Figure 2, a modular system could first allow choosing among different means of pre-processing. Second, it could offer to change the alignment algorithm as this is already done by CollateX [40]. Finally, the user of such a system could toggle between different modes of visualization, gaining on-demand changes of perspective on the same matter.

The Need for Scalable Solutions Though visualizations seemingly serve any upcoming demand, Table 2 also indicates that only one scalable solution (text-oriented heat maps) exists. Other visualization techniques are limited to tens of texts to be compared. With the steady growth of textual data, the need to develop novel scalable text alignment visualization techniques represents a significant future challenge. For example, the Bible is known to exist in more than 450 different English translations. None of the listed approaches supporting collation is ready to visualize variation among them appropriately. Likewise, visualizing text re-use patterns in a large-scale text corpus in an efficient, clear, and user-friendly way would be desirable. Going back to the *Hamlet* example, it would be highly interesting to develop a technique capable of visually conveying the spread of Hamlet quotations offered ready-to-use by the HyperHamlet project [74].

The Need for Document-based Visualizations Table 2 shows us that only one standard visualization technique, the grid-based heat map, is ready to convey alignment relations on document-level visually. All other techniques are per default incapable of visualizing such information. The Chinese Text Project [88] uses a standard tool to generate a network graph also capable of showing relations among documents. However, such solutions require future extensions to convey alignment patterns across texts meaningfully. For example, graph drawings could be tailored to cluster text editions or include temporal dependencies inherent in text corpora.

Variant Graph Opportunities The frequent usage of the variant graph to support collation tasks indicates its value as opposed to other representations like the tabular view. However, Table 2 shows that its current design is not applicable to the entire variety of textual features a corpus might enclose. For example, the structure of a variant graph might be valuable to support manual alignment scenarios. Further, variant graphs could also be applied to analyze sentence sequences across text editions. Especially for variant graphs, defined as being directed acyclic, an inclusion of loops to overcome the issue when dealing with unstable texts could be a beneficial extension. Lastly, variant graphs are, by default, not limited in how many text variants they can bear. However, we set the number in the table to twenty as almost no examples with a higher number of text editions exist. Developing smart methods of clustering recurring phrases among the editions that are not crucial for analyses, and visualizing them could deliver insightful views of larger text bodies. All listed extensions of the variant graph require modifications in node and edge representation as well as a revision of the interaction scheme.

Exploring New Data Sets and Alignment Levels Alignment of *OCR output* has been employed to perform various tasks [24, 97, 100]. Visualizing this kind of alignment can be challenging, especially if we consider relating output characters with their actual positions on the scanned pages. *Movies subtitles* also provide a great source of multilingual parallel texts. The alignment of movies subtitles at sentence or word level has been the subject of numerous studies [8, 91, 94]. Visualizing those alignments in a scalable, interactive form combined with search functionality could be a valuable resource for professional translators and translation studies researchers.

Technique Transfer to Related Domains Alignment is not limited to text only; it has been employed in other domains. The current visualization techniques could be transferred and used to visualize the alignment between other data types. Proven to be effective for text-based scenarios, a transfer to other domains could deliver new, interesting perspectives. For instance, variant graphs could be applied to visualize alignments among protein sequences, side-by-side views can be employed to alignments among musical scores, or sequence-aligned heat maps can be applied to visualize audio signal alignments.

8 Conclusion

In the course of the last years, we have been ourselves involved in developing visualization techniques to support a diversity of text alignment scenarios. However, the overview of related tools and scenarios, in which related text alignment visualizations are applied, and what visual features they choose to convey alignment-related features, has been an interesting exercise. The resulting survey provides a comprehensive overview of how text alignment visualizations currently support domain-related tasks. We discussed their advantages and limitations, proposed design implications and we spotted a number of future challenges that could bring forth new visualizations providing more versatile perspectives to scholars to analyze alignment patterns.

REFERENCES

[1] Glosbe online dictionary. https://glosbe.com, Accessed: 2020-04-10.

[2] Kitab project. http://kitab-project.org/, Accessed: 2020-04-10.

[3] Linguee. https://www.linguee.com/, Accessed: 2020-04-10.

[4] Revesrsocontext contextual dictionary. https://context.reverso.net/, Accessed: 2020-04-10.

[5] Variance viewer. http://variance-viewer.informatik.uni-wuerzburg.de/, Accessed: 2020-04-10.

[6] A. Abdul-Rahman, G. Roe, M. Olsen, C. Gladstone, R. Whaling, N. Cronk, R. Morrissey, and M. Chen. Constructive visual analytics for text similarity detection. *Computer Graphics Forum*, 36(1):237–248, 2017.

[7] J. Ács. Pivot-based multilingual dictionary building using wiktionary. 2014.

[8] F. Al-Obaidli, S. Cox, and P. Nakov. Bi-text alignment of movie subtitles for spoken english-arabic statistical machine translation. *CoRR*, abs/1609.01188, 2016.

[9] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2392–2401, 2011.

[10] E. Alexander and M. Gleicher. Task-driven Comparison of Topic Models. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):320–329, 2015.

[11] B. Almas and M.-C. Beaulieu. Developing a new integrated editing platform for source documents in classics. *Literary and Linguistic Computing*, 28, 10 2013.

[12] B. Almas and M. Berti. Perseids collaborative platform for annotating text re-uses of fragmentary authors. 09 2013.

[13] T. L. Andrews and C. Macé. Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *LLC*, 28:504–521, 2013.

[14] X. Anguera, J. Luque, and C. Gracia. Audio-to-text alignment for speech recognition with very limited resources. 09 2014.

[15] B. Asokarajan, R. Etemadpour, J. Abbas, S. J. Huskey, and C. Weaver. Textile: A pixel-based focus+context tool for analyzing variants across multiple text scales. In *Eurographics Conference on Visualization, Short Papers, Barcelona, Spain, 12-16 June 2017*, pp. 49–53, 2017.

[16] J. B. B. Gipp, N. Meuschke. Comparative evaluation of text- and citation-based plagiarism detection approaches using guttenplag. In *11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2011.

[17] D. Bamman. Building a dynamic lexicon from a digital library. pp. 11–20, 01 2008.

[18] L. Baraldi, M. Cornia, C. Grana, and R. Cucchiara. Aligning text and document illustrations: Towards visually explainable digital humanities. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1097–1102, Aug 2018.

[19] M. Baumann, M. John, H. Pflüger, C. Herberichs, G. Viehhauser, W. Knopki, and T. Ertl. An Interactive Visualization for the Analysis of Annotated Text Variance in the Legendary Der Heiligen Leben, Redaktion. In *LEVIA 2019: Leipzig Symposium on Visualization in Applications*, 2019.

[20] M. Behrisch, M. Krstajić, T. Schreck, and D. Keim. The news auditor: Visual exploration of clusters of stories. pp. 61–65, 01 2012.

[21] M. Bendersky and W. B. Croft. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, p. 262–271. Association for Computing Machinery, New York, NY, USA, 2009.

[22] P. Bojanowski, R. Lagugie, E. Grave, F. R. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. *CoRR*, abs/1505.06027, 2015.

[23] C. S. Bond and A. W. Schüttelkopf. *ALINE*: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallographica Section D*, 65(5):510–512, May 2009.

[24] F. Boschetti, M. Romanello, A. Babeu, and D. Bamman. Improving ocr accuracy for classical critical editions. pp. 156–167, 09 2009.

[25] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[26] P. F. Brown, J. Cocke, S. A. Della-Pietra, V. J. Della-Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. Rossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85, 1990.

[27] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pp. 169–176. Association for Computational Linguistics, Stroudsburg, PA, USA, 1991.

[28] O. Bräckel, H. Kahl, F. Meins, and C. Schubert. *eComparatio – a Software Tool for Automatic Text Comparison*, pp. 221–238. 08 2019.

[29] M. Büchler, P. R. Burns, M. Müller, E. Franzini, and G. Franzini. *Towards a Historical Text Re-use Detection*, pp. 221–238. Springer International Publishing, Cham, 2014.

[30] D. Bär, T. Zesch, and I. Gurevych. Text reuse detection using a composition of text similarity measures. 12 2012.

[31] M. F. Cheema, S. Jänicke, and G. Scheuermann. Annotatevis: Combining traditional close reading with visual text analysis. In *Workshop on Visualization for the Digital Humanities, IEEE VIS*, 2016.

[32] T. Cheesman. Translation Sorting: Eddy and Viv in Translation Arrays. *B. Wiggin (ed.), Un/Translatables*, 19(1):121–142, 2012.

[33] T. Cheesman and A. A. I. R. Roos. Version Variation Visualization (VVV): Case Studies on the Hebrew Haggadah in English. *Journal of Data Mining and Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, July 2017.

[34] T. Cheesman, S. Thiel, K. Flanagan, Z. Geng, A. Ehrmann, R. S. Laramee, J. Hope, and D. M. Berry. Translation arrays: Exploring cultural heritage texts across languages. In *Proceedings of the Digital Humanities 2012*, 2012.

[35] Y. Chen, A. Eisele, and M. Kay. Improving statistical machine translation efficiency by triangulation. In *LREC*, 2008.

[36] C. Chu, T. Nakazawa, and S. Kurohashi. Japanese-chinese phrase alignment using common chinese characters information. In *In Proceedings of MT Summit XIII*. Citeseer, 2011.

[37] P. Clough, R. Gaizauskas, S. S. L. Piao, and Y. Wilks. Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 152–159. Association for Computational Linguistics, USA, 2002.

[38] T. Cohn and M. Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 728–735, 2007.

[39] M. J. Cummings. How Shakespeare's Plays Found Print, 2003. http://www.shakespearestudyguide.com/Folio.html (Retrieved 2017-03-09).

[40] R. Dekker and G. Middell. Computer-supported collation with collatex. managing textual variance in an environment with varying requirements. In *Supporting Digital Humanities 2011*, November 2011.

[41] M. Driscoll and E. Pierazzo. *Digital Scholarly Editing: Theories and Practices*. Digital Humanities Series. Open Book Publishers, 2016.

[42] L. Dwika and K. Schulz. Significant word-based text alignment for text reuse detection. 01 2018.

[43] S. Eddy. Multiple alignment using hidden markov models. 12 1995.

[44] D. Fetterly, M. Manasse, and M. Najork. On the evolution of clusters of near-duplicate web pages. In *Proceedings of the First Conference on Latin American Web Congress*, LA-WEB '03, p. 37. IEEE Computer Society, USA, 2003.

[45] P. Forner, J. Karlgren, C. W. hacker (eds, M. Potthast, T. Gollub, M. Hagen, J. Graßegger, J. Kiesel, M. Michel, A. Oberländer, A. Barrón-cedeño, P. Gupta, P. Rosso, and B. Stein. Overview of the 4th international competition on plagiarism detection, 2012.

[46] G. Franzini, E. Franzini, and M. Büchler. Historical text reuse: What is it?, 2016. http://www.etrap.eu/historical-text-re-use/, Accessed: 2020-04-24.

[47] Z. Geng, T. Cheesman, R. S. Laramee, K. Flanagan, and S. Thiel. Shakervis: Visual analysis of segment variation of german translations of shakespeare's othello. *Information Visualization*, 2013.

[48] B. Gipp. Citation-based Plagiarism Detection. In *Citation-based plagiarism detection*, pp. 57–88. Springer, 2014.

[49] P. Goffin, W. Willett, J. Fekete, and P. Isenberg. Exploring the placement and design of word-scale visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2291–2300, 2014.

[50] R. Haentjens Dekker, D. van Hulle, G. Middell, V. Neyt, and J. van Zundert. Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Literary and Linguistic Computing*, 30(3):452–470, 03 2014.

[51] M. Harrower and C. A. Brewer. Colorbrewer.org: an online tool for

selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[52] A. Hiemann, T. Kautz, J. Waschke, and M. Hlawitschka. Visual Information Fusion of Spatio-Temporal Multi-Modal Data in Sports. In *VisWeek 2019 Posters*, 2019.

[53] S. Jänicke, T. Efer, M. Büchler, and G. Scheuermann. Designing Close and Distant Reading Visualizations for Text Re-use. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pp. 153–171. Springer, 2014.

[54] S. Jänicke and A. Geßner. A Distant Reading Visualization for Variant Graphs. In *Conference Abstracts of the Digital Humanities 2015*, 2015.

[55] S. Jänicke, A. Geßner, M. Büchler, and G. Scheuermann. Visualizations for text re-use. In *IVAPP '14: Proceedings of the 5th International Conference on Information Visualization Theory and Application*, 2014.

[56] S. Jänicke, A. Geßner, G. Franzini, M. Terras, S. Mahony, and G. Scheuermann. TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities*, 30(suppl_1):i83–i99, 10 2015.

[57] S. Jänicke, P. Kaur, P. Kuźmicki, and J. Schmidt. Participatory Visualization Design as an Approach to Minimize the Gap between Research and Application. In *The Gap between Visualization Research and Visualization Software (VisGap)*. The Eurographics Association, 2020.

[58] S. Jänicke and D. J. Wrisley. Interactive visual alignment of medieval text versions. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 127–138, 2017.

[59] Juxtacommons. A user guide to juxta commons, 2019. `http://juxtacommons.org/guide`, Accessed: 2019-09-10.

[60] K. Katoh, J. Rozewicki, and K. D. Yamada. Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, 20(4):1160–1166, 2019.

[61] M. Kay and M. Röscheisen. Text-translation Alignment. *Computational linguistics*, 19(1):121–142, 1993.

[62] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 117–121, 2015.

[63] R. Lewis. 'A Brief Discourse of Rebellion and Rebels' by George North: A Newly Uncovered Manuscript Source for Shakespeare's Plays. By Dennis McCarthy and June Schlueter. *The Library*, 19(4):514–520, 12 2018.

[64] A. Lopez. Word-based alignment, phrase-based translation: What's the link. In *In Proc. of AMTA*, pp. 90–99, 2006.

[65] J. Lundborg, T. Marek, M. Mettler, and M. Volk. Using the stockholm treealigner. 01 2007.

[66] J. Maxwell and K. Rumbold. *Shakespeare and Quotation*. Cambridge University Press, 2018.

[67] I. D. Melamed. Manual annotation of translational equivalence: The blinker project. *ArXiv*, cmp-lg/9805005, 1998.

[68] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, p. 517–524. Association for Computing Machinery, New York, NY, USA, 2005.

[69] D. W. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2nd ed., 2004.

[70] T. Nakamura, E. Nakamura, and S. Sagayama. Real-time audio-to-score alignment of music performances containing errors and arbitrary repeats and skips. *CoRR*, abs/1512.07748, 2015.

[71] S. Needleman and C. Wunsch. 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, 3 1970.

[72] E. Nury. Visualizing collation results. *Variants*, pp. 75–94, 03 2019.

[73] M. Pöckelmann, A. Medek, P. Molitor, and J. Ritter. Catview - supporting the investigation of text genesis of large manuscripts by an overall interactive visualization tool. In *Proceedings of the Digital Humanities 2015*, 2015.

[74] S. Quassdorf. Hyperhamlet: A database of quotations from and allusions to shakespeare's most famous tragedy. October 2019.

[75] R. L. Ribler and M. Abrams. Using visualization to detect plagiarism in computer science classes. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pp. 173–. IEEE Computer Society, Washington, DC, USA, 2000.

[76] P. Riehmann, H. Gruendl, M. Potthast, M. Trenkmann, B. Stein, and B. Froehlich. Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *IEEE transactions on visualization and computer graphics*, 18, 03 2012.

[77] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich. Visual assessment of alleged plagiarism cases. *Computer Graphics Forum*, 34, 06 2015.

[78] R. Rosselli Del Turco and C. Di Pietro. Between innovation and conservation: The narrow path of user interface design for digital scholarly editions. pp. 133–163, 12 2018.

[79] K. P. Sankar, C. V. Jawahar, and A. Zisserman. Subtitle-free movie to script alignment. In *BMVC*, 2009.

[80] D. Schmidt and R. Colomb. A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67:497–514, 06 2009.

[81] S. Schreibman. The versioning machine. *Literary and Linguistic Computing*, 18:101–107, 04 2003.

[82] S. Schütz and M. Pöckelmann. Lera - explorative analyse komplexer textvarianten in editionsphilologie und diskursanalyse. In *Proceedings of the Digital Humanities im deutschsprachigen Raum, DHd 2016*, 2016.

[83] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, Sep. 1996.

[84] S. Silvia, R. Etemadpour, J. Abbas, S. Huskey, and C. Weaver. Visualizing variation in classical text with force directed storylines. In *Proceedings of the Workshop on Visualization for the Digital Humanities*. IEEE, Baltimore, MD, October 2016.

[85] T. F. Smite and M. S. Waterman. Identification of common molecular subsequences.

[86] D. A. Smith, R. Cordell, E. M. Dillon, N. Stramp, and J. Wilkerson. Detecting and modeling local text reuse. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, p. 183–192. IEEE Press, 2014.

[87] N. Smith and C. Blackwell. Ducat citation alignment 1.3.0. `https://eumaeus.github.io/uva_cex_ducat/`, Accessed: 2020-04-10.

[88] D. Sturgeon. Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, 33(3):670–684, 11 2017.

[89] G. Tanselle. *A Rationale of Textual Criticism*. Literatura (E-libro). University of Pennsylvania Press, Incorporated, 2010.

[90] S. E. Team. Textual Criticism Analysis - King Lear by William Shakespeare, 2008. `http://www.shmoop.com/textual-criticism/king-lear-analysis.html` (Retrieved 2017-03-07).

[91] J. Tiedemann. Improved sentence alignment for movie subtitles. In *In Proceedings of RANLP, Borovets*, 2007.

[92] J. Tiedemann. *Bitext Alignment*, vol. 4. 05 2011.

[93] R. H. Trillini. *Casual Shakespeare: Three Centuries of Verbal Echoes*. Routledge, 2018.

[94] A. Tsiartas, P. K. Ghosh, P. G. Georgiou, and S. Narayanan. Context-driven automatic bilingual movie subtitle alignment. In *Proceedings of Interspeech 2009*. Brighton, UK, Sept. 2009.

[95] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 01 2009.

[96] M. Wattenberg and F. B. Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, Nov 2008.

[97] D. Wemhoener, I. Z. Yalniz, and R. Manmatha. Creating an improved version using noisy ocr from multiple editions. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 160–164, 2013.

[98] S. Werner. Welcome to the collation, 2011. `https://collation.folger.edu/2011/08/welcome-to-the-collation/`, Accessed: 2019-10-10.

[99] W. Williams and C. Abbott. *An Introduction to Bibliographical and Textual Studies*. Modern Language Association of America, 1985.

[100] I. Yalniz and R. Manmatha. A fast alignment scheme for automatic ocr evaluation of books. pp. 754 – 758, 10 2011.

[101] T. Yousef. Ugarit: Translation alignment visualization. OSF Preprints, Mar 2020.

[102] T. Yousef and M. Foradi. Word alignment of divan-e-hafez. `http://divan-hafez.com/`, Accessed: 2020-04-10.

[103] T. Yousef, C. Palladino, and G. Crane. Intra-language text alignment using ialigner. In *Proceedings, 7th International Conference of Digital Archives and Digital Humanities*. National Taiwan University, 12 2016.

[104] S. Zinger, J. Nerbonne, and L. Schomaker. Text-image alignment for historical handwritten documents. vol. 7247, pp. 1–10, 01 2009.