Visual Assessment of Alleged Plagiarism Cases

P. Riehmann, M. Potthast, B. Stein & B. Froehlich

Bauhaus-Universität Weimar

Abstract

We developed a visual analysis tool to support the verification, assessment, and presentation of alleged cases of plagiarism. The analysis of a suspicious document typically results in a compilation of categorized "finding spots". The categorization reveals the way in which the suspicious text fragment was created from the source, e.g. by obfuscation, translation, or by shake and paste. We provide a three-level approach for exploring the finding spots in context. The overview shows the relationship of the entire suspicious document to the set of source documents. A glyph-based view reveals the structural and textual differences and similarities of a set of finding spots and their corresponding source text fragments. For further analysis and editing of the finding spot's assessment, the actual text fragments can be embedded side-by-side in the diffline view. The different views are tied together by versatile navigation and selection operations. Our expert reviewers confirm that our tool provides a significant improvement over existing static visualizations for assessing plagiarism cases.

Categories and Subject Descriptors (according to ACM CCS):
H.5.2 [Computer Graphics]: User Interfaces—Graphical user interfaces

1. Introduction

Text reuse is ubiquitous and ever-present. News messages travel from website to website with only slight changes in wording or identical text fragments emerge in a passed version of a bill that have previously been released in documents drawn up by lobbyist groups. While these cases often have little or no consequences for the plagiarizing authors, this is different for student essays or PhD theses accused of plagiarism. In these cases, text passages originating from other authors have been either directly copied or slightly rewritten without properly referring to the original sources which is, in the best case scenario, a lack of scientific thoroughness. Claiming that a given piece of writing has been plagiarized can have severe consequences for those accused. The supporting evidence of such an accusation needs to be presented in a convincing way or it may be refuted, regardless of truth. To ameliorate the situation, we developed an interactive visual analysis tool (Figure 1) which provides effective views and appropriate linking and filtering techniques to explore an alleged case of plagiarism from the entire document down to individual suspicious sections of text (finding spots). An overview provides insight into the distribution of finding spots across the document, their lengths and categorizations, and their relation to sources and authors. Effective filtering, linking, and navigation techniques facili-

tate the process of focusing on different aspects of the case, such as a certain source, plagiarism category, or the largest finding spots. The selected finding spots are presented as a list of difflines, a glyph-based abstraction for revealing the inner structure of a finding spot. They serve as intermediate representation between overviews and actual text by encoding the modifications that turned the source text fragment into a finding spot by explicitly highlighting the copy-and-paste sequences. For drilling down a finding spot, the actual text fragments can be opened below as textual view. Therefore, along with our set of expert functions, each finding spot can be considered in detail and, if needed, be reassessed or altered and, eventually, the assessor must approve or reject it from the list of suspicious fragments.

The specific motivation for this work stems from our professional experience as developers of the text reuse search engine Picapica [Pot] and as initiators and organizers of an annual international competition on plagiarism detection, called PAN [PSRS]. In this context, we are also in contact with experts and members of the German anti-plagiarizing community. Discussions with our colleagues and experts, as well as a review of available tools, revealed that most plagiarism search engines present their results as running text containing color-highlighted word sequences at positions where text has been reused whereas different colors hint at differ-
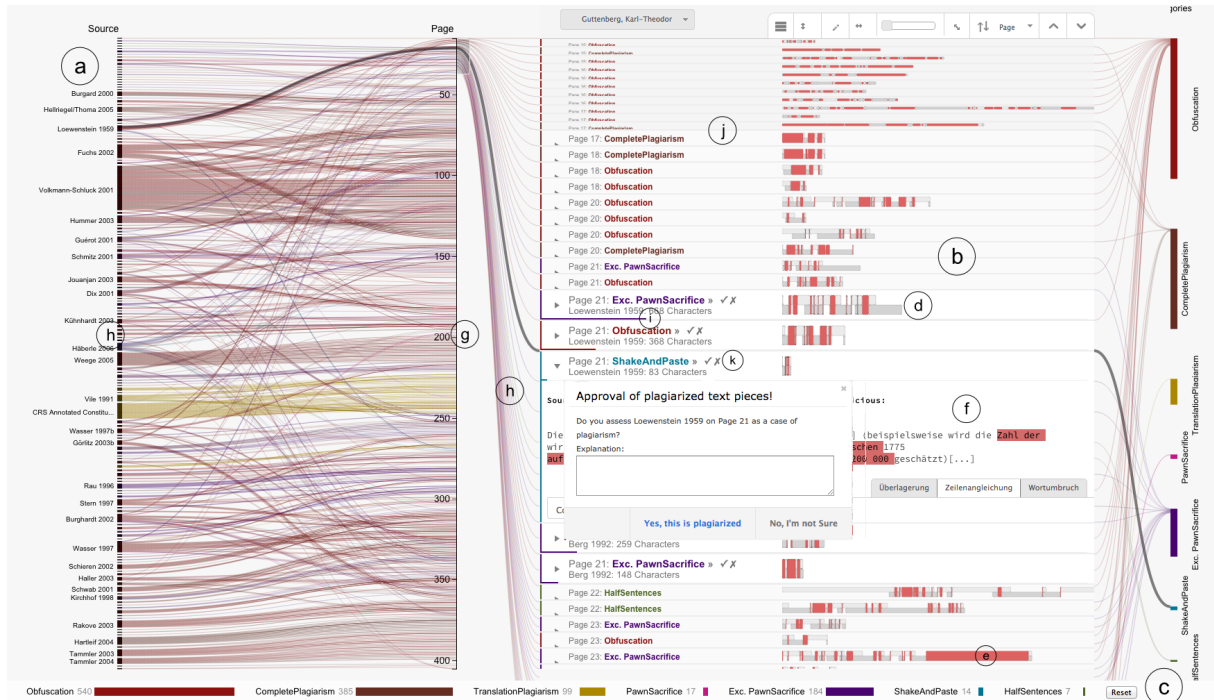
Figure 1: *Our visual tool for assessing cases of plagiarism displays the types of plagiarism found on the bottom (c), a list of difflines (b) (glyph-based visualization of the finding spots (d)) in the center, and an overview on the left (a). Copy-and-pasted passages are marked in red (e). A finding spot can be opened for a side-by-side comparison of suspicious and original text fragments. The overview reveals the distribution of finding spots across the document (g) and their relationship to the sources (h). The overview supports brushing and selection to define a subset of finding spots to be displayed in the diffline view.*

ent source documents. A few tools provide basic overviews with only very limited interaction capabilities. Such solutions may suffice if short texts have to be analyzed. However, they do not scale gracefully with text length, nor with complexity of a plagiarism case. In such cases, a lot of information concerning different aspects of work and practice needs to be considered by experts or discussed in a council charged to audit a suspicious case. Besides answering questions about the overall characteristics of the suspicious document, the assessment of each individual finding spot remains crucial. The aforementioned solutions mostly provide scrollable page-based textual views for browsing the entire document. However, scrolling a 400 page document interrupted by reading and comparing each finding spot to the related source is a tedious task.

The central contributions of our plagiarism analysis tool include a three-tiered approach for exploring alleged cases of plagiarism, new overview paradigms for navigating and selecting subsets in a suspicious document, difflines as effective glyph-based abstractions of differences and similarities between two text fragments, and the support for fluid and coherent interaction between the different levels of detail. As an initial data set, we chose the most elaborate collections of suspicious PhD theses, GuttenPlag [Mis14a] and

VroniPlag [Mis14b]. Reviews with our plagiarism experts confirm that our tool can effectively support their workflow and provides a significant improvement over existing static visualizations for assessing plagiarism cases, especially regarding time savings during the assessment process and in visually supporting councils and committees in forming an opinion about a plagiarism case.

2. Anti-Plagiarism Community

In Germany, a very active and self-organized anti-plagiarism community is committed to finding and documenting cases of plagiarism in PhD theses. The members document their results in public wikis such as GuttenPlag [Mis14a] and VroniPlag [Mis14b]. They scrutinize documents that have been suspicious to one or several members for various reasons. It is an ongoing process which typically takes months or even years since all community members are volunteers. Each finding spot is documented, compared with the work it was allegedly taken from, and published with specific information such as position in the suspicious document, position within the original document, original author, etc. A single source or even multiple documents of the same author(s) are often used repeatedly. Eventually, the finding spots are categorized as different types of plagiarism. The most common

categories defined by the community are described below, as are their colors used in our system.

- ■ **(Almost) Complete Plagiarism**: a section largely produced by copy-and-paste.
- ■ **Obfuscation**: a text passage which is more or less paraphrased, often by simply substituting words with synonyms or inserting/deleting select words here and there.
- ■ **Pawn Sacrifice**: text from a cited source is used but is referred to somewhere else in the document.
- ■ **Exacerbated Pawn Sacrifice**: text is copied straight from a source and a correct reference is cited, but the reference is introduced with "'likewise . . . '" suggesting that there is a similar statement but not equal text.
- ■ **Shake and Paste**: longer text sections, typically paragraphs, are taken and mixed from different sources.
- ■ **Half Sentence Mending**: short sentences or sentence fragments from a source have been used.
- ■ **Translation Plagiarism**: a text translated from a foreign language source which was more or less rephrased.

The barcode visualization [Mis14b] is the most common visualization utilized by members of the anti-plagiarism community. It provides an overview of a suspicious document and is used to demonstrate the current status of an ongoing investigation. The horizontal barcode shows the pages of a document as vertical stripes which indicate whether one or more finding spots occur on a particular page. A five-level color scale defines the amount of suspicious fragments per page. The depiction is usually just a static image, but some can show the detection of finding spots over time in an animation. Another non-interactive visualization [Use14] of the GuttenPlag community employs a page-based view of the entire document. Each finding spot is shown in a color that corresponds to the author of the source document. However, in Guttenberg's case, with nearly 400 pages and 138 different authors, the colors are too similar to allow an unambiguous assignment to an original author. Nevertheless, such a visualization provides a solid overview of the amount of text taken from others and it works well for minimal sources.

3. Related Work

A different kind of text reuse, documented by the Lobbyplag website [Ope14], reveals changes in regulation drafts for the General Data Protection Regulation (GDPR) of the European Union. It allows a comparison of changes within the committee amendments and relates them to lobby proposals about the same topic that might contain similar content and wording. A horizontal barcode spanning the entire page serves as overview and navigation tool. The amendments and lobby proposals are shown as a side-by-side comparison without visually linking the texts in any manner. Only text changed in the amendments and proposals is highlighted in red (removed) and in green (newly inserted).

In her book, Weber-Wulf [Web14] gives an overview of the current situation of plagiarism and its detection. More than 50 plagiarism detection systems can be found, some offered as commercial products such as Turnitin [iPa14]), Ephorus (now merged with Turnitin), and Urkund [Pri15], and some merely small open source tools. Since 2004, almost all of the available systems have been repeatedly evaluated with respect to their detection quality and fitness for purpose, the results of which have been published at [Web]. Since 2008 these evaluations also assess usability. In this regard, few systems achieve more than 70% of the available points (both on an objective and a subjective scale), so that most are rated "poor" or even "unacceptable" [WMTZ]. We surveyed the available systems with regard to their visualizations employed: none of the systems individually visualize findings and only few provide abstract overviews of their findings, which usually boil down to tables that give numbers of findings alongside document names.

Gipp and Meuschke [GMB*13] developed a visualization based on an underlying citation-based plagiarism detection algorithm. The documents are also arranged side-by-side with overview bars in-between representing the entire document. References are shown as dots in each overview bar and identical citations are connected by a curved line (see Citeplag website [Sci14] for examples). They also published an interesting survey about the state of the art in detecting academic plagiarism [MG13]. The paper of Jänicke [JGBS14] offers several visualizations of textual differences and commonalities of different English Bible translations, such as Text Re-use Grid, text-centered visualizations, and Sentence Alignment Flows, which strongly resemble the Wordgraph metaphor [RGP*12].

The visualization of regular diff algorithms is also related to the depiction of plagiarism. windiff:2014 [Mic14], an older tool for comparing different revisions of source code, provides vertical bars beside the text views which show differences of code revisions by coloring variations and identifying moved parts. Contrary to our approach, it does not focus on the equal parts by particularly aligning the changed parts alongside the remaining ones. Unfortunately, this approach is barely applicable to continuous text that is not explicitly wrapped, such as source code. Chevalier et al. [CDBF10] propose a different approach by utilizing an animation technique for smooth transitions between text revisions. Another topic related to certain aspects of our approach is the visual tracking of changes made during consecutive revisions or edits of single text documents, which is exemplified in HistoryFlow by Viegas [VWD04], the Wikidashboard by Suh [SCKP08], or the Chromogram by Wattenberg [WVH07]. An interesting approach, also supporting the navigation between several levels of abstraction while exploring large texts, was provided by Koch [KJW*14].

4. Design Process and Visual Concept

The annual PAN [PSRS] competition on plagiarism detection, which we organize, and our own text reuse search engine Picapica [Pot] focus on the automatic retrieval of plagiarism. Visualization was not a necessity when plagiarism detectors were evaluated in the past (see for example [PHB*14]). Nevertheless, participants of the competition and customers of our Picapica service alike frequently ask for solutions that save work time when reviewing plagiarism cases.

The development of Picapica, as well as that of our first visualizations, was advised by the German anti-plagiarism community. Their process of *manually* analyzing a suspicious PhD thesis can be summarized as follows: after a suspicion has been raised, the document in question is scanned for further dubious text spots, usually by manual retrieval. For each so-called finding spot, a corresponding text fragment from a potential source document is listed. In addition, the finding spots are classified with respect to the perceived way in which the suspicious text fragment has been derived from its source (e.g., by obfuscation, mending the sentence fragments of the original, or simply by copying and pasting).

The rationale for identifying as many finding spots in a suspicious document as possible is due to the fact, that, in practice, a single, short plagiarized text passage is considered insufficient evidence to make a case against the document's author: for example, the author might claim a mishap. Therefore, a complete analysis of a suspicious document is a strict necessity to support and defend plagiarism allegations. For instance, when councils need to form an opinion about a plagiarism case, a lot of information concerning different aspects about work and practice needs to be considered by experts or discussed in the council. Based on the identified finding spots, they have to answer those questions which are critical to a thorough assessment of an entire suspicious document: How are the finding spots distributed among the pages of the entire document? Which categories of plagiarism are present in the document and which of them are most frequent? How many sources were used? Which sources are used most for paraphrasing text and to what extent? Which sources appear in which category and how often? What is the average length of the finding spots or, more specifically, what is the distribution of their lengths? Besides the consideration of these general characteristics of the suspicious document, the assessment and reassessment, presentation, and discussion of individual finding spots is an important part of the process.

For an effective support of this process and to answer the aforementioned questions in a convincing way, we derived the following key elements of our visualization system:

- An overview is needed to support group decision processes in order to gain insight into the distribution of finding spots across the document, their lengths and categorizations, and their relation to sources and authors.

- The most important requirement, saving time in forming an opinion about a list of finding spots, is facilitated by introducing a compact glyph-based representation which demonstrates the relationship between source text and finding spot. This intermediate representation visually emphasizes the copy-and-paste fragments of a finding spot and therefore simplifies reaching a consensus about a finding spot without looking at the text.
- The actual text fragments—source text and finding spot—are sometimes still necessary and can be opened below a diffline as a side-by-side or merged view.
- Effective filtering techniques facilitate the process of focusing on different aspects of the case, such as a certain source, plagiarism category, or the largest finding spots in order to verify the claim of plagiarism or to convince council members with respect to a given case.

Our visual plagiarism analysis tool is aimed at people who typically do not have any experience in advanced information visualization and need to focus on the analytical task. Thus it is clearly structured and only consists of the category view on the bottom, the overview on the left, and the main view in the center which shows the list of finding spots visualized as difflines. The overview and the main view are linked and we provide appropriate navigation techniques to explore the entire document down to individual finding spots.

4.1. Visualizing All Finding Spots at Once

The overview visualizations enable users to interactively explore different aspects of the structure of the suspicious document. Our graph-like view relates the pages where finding spots occur and the extent of finding spots, as well as the different finding spot categories to the source documents from which they were allegedly taken. The overviews are exchanged according to the overall sorting order (by page within suspicious document, by text length of the finding spot, by plagiarism category, or by source document). A crossing minimization is applied to improve the aesthetics.

The individual overviews also allow the users to navigate the entire document and to filter the finding spots based on the aforementioned features. For continuous features, a range-based filter is provided: a particular subset of the pages or a set of really short finding spots can be selected. Discrete values are filtered by directly selecting their visual representations. Filtering of finding spots works consistently across all views and defines the finding spots that are contained in the diffline list. The currently viewable detail of the list is emphasized. The finding spots reveal their position or ranges within the overview by connecting the vertical positions to the respective finding spot entries with paths crossing the gap between both views (see Figure 1(h)). The existence and controls of these paths are being adjusted on the fly while scrolling, filtering, or reordering the list, so it is always clear which subsets (in the categorical/source views)

or which ranges (in the page/length view) of difflines can be seen at the moment. The category bar shown at the bottom provides information about the different kinds of plagiarism and their respective numbers occurring in the document under investigation. It also allows the selection of a subset of categories. Additionally, if enough horizontal space is available, the category bar can be integrated into the right-hand side by connecting the finding spots with their categories.

4.2. Finding Spots and Difflines

The finding spot entries with their difflines are arranged in a tabular layout within the main view in order to enable the comparison of plagiarism patterns of several finding spots. Each finding spot is represented as a horizontal entry in which all of its essential information is shown (Figure 2). Our central goal when designing the diffline was to visually convey information about the structure of the finding spot and its differences to the source without being forced to read the text itself. Our analysis of the finding spots of available cases revealed that, across all the different plagiarism categories (except translation plagiarism), there is a lot of direct copy-and-paste occurring. The frequencies and patterns seemed somewhat different, but it was difficult to judge by solely comparing two text fragments side-by-side. This observation led to the idea to provide a visual diff representation that expresses how a text changed between a finding spot and its source. With an appropriate glyph alphabet we are able to present the changes in a visual manner:

1. Identical fragments (copy and paste)
2. Modification of text fragment resulting in fewer, equal, or more characters
3. Insertion or removal of characters at a certain position (boundary cases of the above)

Different glyph alphabets were designed. Figure 5 depicts three designs that were both promising and unique enough to be tested by users during our pilot phase (see Section 6). As a general rule, all diffline designs represent the source document above the suspicious document, following a left/upper=source → right/lower=plagiarism rule, which is consistent with the other views. The text length of the finding spot is usually encoded as the length of its diffline (see Figure 1(d)). For some tasks however, e.g. in order to facilitate the search and comparison of multiple diffline patterns, it makes more sense to use the entire horizontal space (like in Figure 2(c)). In such cases, we encode the actual length of the finding spot separately as a horizontal bar drawn in the category color below each list entry, whereas the longest finding spot of the document is used for normalization (see Figure 2(b) and also Figure 1(i)).

Our example cases usually contain more finding spots than can be displayed with all relevant information (category, page number, title of possible source, etc) on a regular screen. To see the entire picture and to avoid unnecessary scrolling, we provide means to semantically scale all
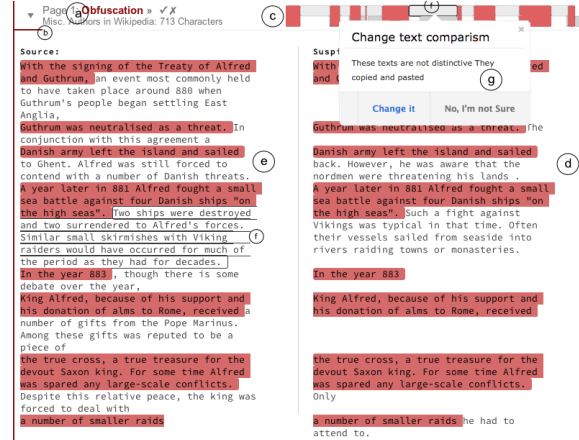

Figure 2: *The visual representation of a finding spot shows the essential information on the top left (a) (position and length of the suspicious fragment where length is indicated as a thin horizontal line directly below (b)), plagiarism category (also shown on the left as a vertical line with the color of the category), and the name of the potential original document. The diffline visualization is shown at the top (c). The finding spot is opened and the suspicious text fragment (d), as well as the potential original (e), are shown directly below. The textual view is based upon a particular wrapping intended for easier recognition of the differences and commonalities of both texts. Hovering above a glyph or a text element will highlight both (f) to simplify the mental match.*

list entries up or down by hiding or showing less important information, changing font sizes, and adjusting the size of difflines. At the lowest detail level, the diffline is minimized by sliding the upper and the lower part of a diffline on top of each other so that the copy-and-paste structure remains legible, whereas details about which kind of changes occurred are omitted. The different level of details can also be combined in a Focus and Context view (see Figure 1(j)) where the list entries in the center are given more vertical space to provide additional information while the remaining entries become smaller towards the top and the bottom margin.

4.3. The Textual Views

Although a diffline reveals lots of information about a finding spot and its alleged source, both must be accessible and comparable in a textual form, too. The textual views can be opened on demand and are embedded in the diffline list directly below their respective diffline. We support three different approaches for comparing text fragments. The first approach resembles the depiction of tracked changes in a word processor (Figure 3). The second approach enables fading in and out the differences between original and suspicious text (Figure 4). Both approaches eventually present variations of a respective text fragment embedded in a single running text,

which might be ideal for reading purposes whereas for the diffline approach it seems more promising comparing texts side-by-side with an appropriate wrapping (Figure 2). The wrapping should facilitate the detection of differences and commonalities between texts at a glance and direct the user to the location where reading in detail might be most relevant. In our layout, the identical parts in both texts serve as the skeleton, which is vertically aligned across both texts and highlighted by their background. Modified text blocks are vertically filled so that the corresponding copy-and-paste sections remain aligned. A monospace font with equal character width facilitates judgment regarding how much text has been removed and added or if a substitution of equal length occurred. The color ▬, representing the copy-and-paste sections of the difflines, is used as the background to visually link the structure of the layout and the diffline glyphs (Figure 2(c),(e),(d)). While the copy-and-paste sections, in general, start at the beginning of a line, the modified text blocks in between can also start on a new line or simply at the end of the copy-and-paste section. The latter results in a more compact layout which is useful if the frequency of copy-and-paste sections and modifications is high, e.g., for the plagiarism category Half Sentence Mending.
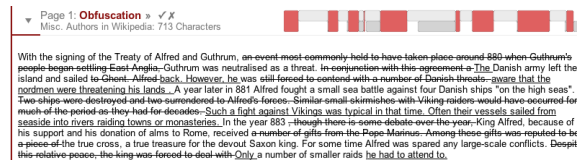

Figure 3: *The classic approach for text comparison uses striking out or underlining to reveal removed and inserted words, respectively. The sample text was taken and adapted from [Mis14c].*

Since our system is intended to support an assessor's workflow, this sometimes means supporting less spectacular and more common interactions which can nevertheless be crucial for improving workflow. Each finding spot shows a set of icons (only at a certain level of detail), such as icons for approving (Figure 1(k) and 4(b) approved) or rejecting the finding spot, which will then be removed from the list. Another icon enables re-assigning finding spots to other plagiarism types should their current type not be suitable, e.g., for not containing enough copy-and-pasted pieces to be considered complete plagiarism. More importantly, each glyph of the diffline (see Figure 2 (c)), as well as each element in the text view, can be altered (by animated transitions of shape and color). For example, if corresponding text fragments are marked as equal (e.g., by mistake of another member of the community), but instead contain many changes, they can be re-assessed. Conversely, if the the diff algorithm differentiates between text pieces which are, in fact, nearly the same, they can be combined (Figure 2 (g)).
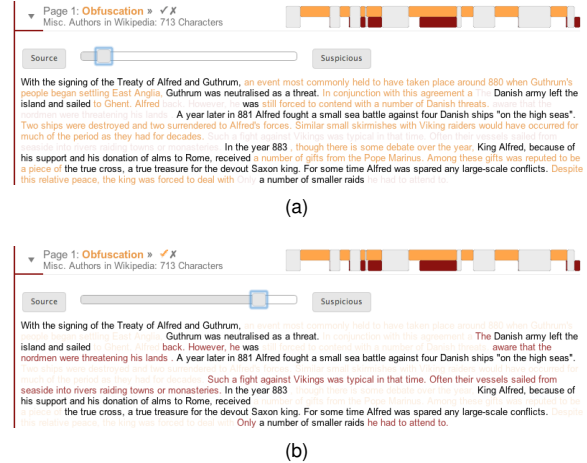

(a)

(b)

Figure 4: *The diff blending morphs between the two texts by adjusting the transparency values of the respective markups of the changed text. (a) The focus is on the source. (b) The focus is on the suspicious text. Although gaps are created by focusing on the source (a) or on the finding spot (b), the text can be read surprisingly well. Moving the slider from one stop to the other creates an animation-like behavior that draws attention to the changes, especially when using our alternative color scheme.*

4.4. Color Model

Although the difflines were designed to reveal the structure of a finding spot even without color, an appropriate color coding facilitates the process of recognizing and interpreting the glyphs. The default color coding displays identical word sequences in difflines and the text views in red ▬ (see Figures 2, 3). The color red emphasizes the fraction and frequency of copy-and-paste actions that were used to assemble a finding spot. It also aids in visually matching copy-and-paste fragments in the diffline with the structured text view. We experienced that pure red looks aggressive and unpleasant on most displays. Words that appear in only one of the aligned texts were shown in different gray levels (▬ and ▬). An alternative color scheme (Figure 4) aims to draw attention to the modified parts which might have to be analyzed further. The visual impression is inverse to the first scheme. A neutral gray tone ▬ is used for the copy-and-paste passages. A shade of gold-orange ▬ is introduced for word sequences that only appear in the alleged original work. It is supposed to express originality and positive character. Text fragments which are only contained in the suspicious work are shown in the category color, which hints at how this modified text segment has been created. For example, if the plagiarism category is obfuscation and the text fragment in the finding spot and in the original are of approximately equal length, the finding spot is probably a paraphrased version of the original, which merits closer inspection.

We chose to assign colors to the categories (usually less

than seven per case) since mapping each source document to an individual color was not appropriate due to the large number of sources in some cases (compare [Use14]). Colors like the Tableau 20 color scheme [Ger] were tested but rejected for looking far too positive. Subsequently, the colors were selected by hand to evoke at least a neutral look, or ideally, a negative impression that seems more appropriate in this context. We derived ▬ and ▬ from ▬ for (Almost) Complete Plagiarism and Obfuscation. Pawn Sacrifice ▬ and Exacerbated Pawn Sacrifice ▬ use different, but familiar, tones to emphasize their commonality, as well as Shake and Paste ▬ and Half Sentence Mending ▬. Translation Plagiarism ▬ uses a hue which is not related to all others. Although our color scheme narrows down the color space, this was never experienced as an issue, both in our expert reviews and during our lab demonstrations: our color scheme maintains a reasonable level of discrimination.

5. Data Preprocessing and Implementation Details

The alleged cases of plagiarism are publicly available at the aforementioned wikis [Mis14a] and [Mis14b]. We acquired the underlying data via the Wikia-API. Since the cases have been entirely manually annotated with very limited templating support from Wikia's Wiki software, many inconsistencies with regard to naming schemes, tags, typos, encodings, etc. remain. All of these issues cause little disruption to the Wikis since the Wiki software handles them gracefully, but they foreclosed our attempts to process the raw data automatically. We therefore systematically reviewed the plagiarism cases and semi-automatically removed inconsistencies by hand, sometimes using Python scripts. As a result of roughly 180 hours of student work, a total of 41 plagiarism cases containing nearly 6100 finding spots (with an average of 6200 words per spot) that link to over 950 sources are now available in a consistent JSON format.

Our prototype is entirely web-based and both its logic and presentation layer are executed at client side, whereas the server only delivers the web page along with the required script files. The JSON files of the finding spots are dynamically prefetched during scrolling and filtering operations before the respective difflines come into view. The system has been developed and tested with recent versions of the Chrome web browser. Four JavaScript libraries were used: jQuery for accessing the DOM-Elements more conveniently, low-level methods of D3 for structuring and wrapping the drawing operations, the google-diff-match-patch library to determine text changes between a finding spot and its original, and Backbone.js for MVC support.

6. Diffline Design Decisions

Several glyph alphabets were designed to express what changes might have occurred between two texts. Three designs that were most promising were chosen based on their respective features (see Figure 5): (1) the rectangular difflines because of their simplicity, (2) the trapezoidal diffline due to their seemingly expressive glyph alphabet, and (3) the condensed difflines because of their compactness. A pilot study was conducted to obtain feedback about their general usability, their comprehensibility, and which of them should serve as default. We chose a between-group design with 18 participants. Our rationale for doing so was due to the fact that being briefed in two or more diffline alphabets causes confusion: similar visual elements were used across the alphabets, and a strong learning effect occured from performing the same task consecutively, albeit with different alphabets.

Each participant accomplished three different tasks. Prior to these tasks, the participants were thoroughly briefed about the characteristics of the particular diffline used in his/her group by exploring two different example difflines along with their corresponding finding spots consisting of the source and suspicious text fragments. For the first task, the glyphs of four difflines had to be assigned to their matching text pieces. These difflines varied in word length and structure (approximately 14-21 glyphs per diffline, a representative number). As for the second assignment, the participants were supposed to visually examine another four difflines— glyph by glyph and without accompanying text—and to explain what changes could have possibly occurred. Finally, the participants answered a questionnaire about how difficult they found the assignments, how useful they found the glyph alphabets, and what general improvements they propose.

The pilot study indicates that the rectangular difflines were appreciated most (mean of 1.3 on a 6 point Likert scale), whereas the other approaches were judged as being less comprehensible (mean of 2.3). Some difficulties occurred while interpreting and orienting the short rectangles for the condensed difflines and comprehending the meaning of the orientation of the triangles in the trapezoidal ones. The rectangular version was most easily understood and resulted in no errors when interpreting glyphs, while an average of 2 and 4.3 errors were made for the trapezoidal and condensed difflines, respectively. Although nearly half of the participants recommended the usage of colors as a very helpful improvement, the results of rectangular difflines show that gray levels (no colors at all) are sufficient for the specific tasks of this study. Altogether, we chose the rectangular version as default one and used color to highlight copy-and-paste fragments in a finding spot.

7. Expert Reviews, Feedback and Findings

After the main functionality of the system was developed (overview, difflines, textual views, basic interaction), we reviewed the system with three external experts. One writes plagiarism assessments for a living and the others are very active in the German anti-plagiarism community. Even though they have been very active in the community for
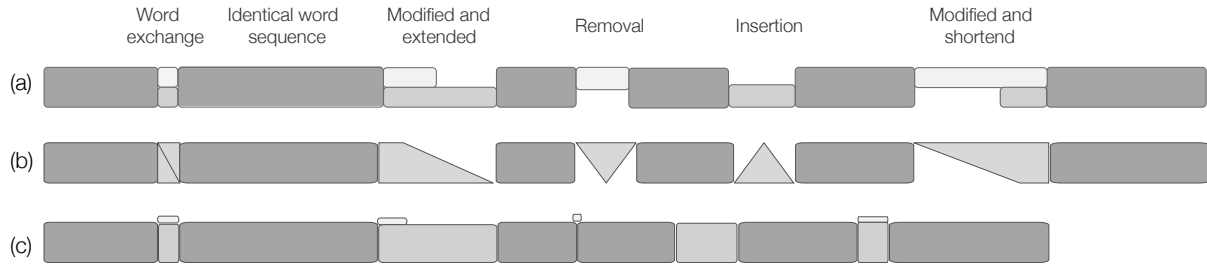
Figure 5: *The various difflines composed of different glyph alphabets. All of them represent the same information. The top area represents the original document. The lower part represents the suspicious fragment.*
*(a) **Rectangular diffline**: Only rectangular shapes of various length and height are used as glyphs. The copy-and-paste sections, depicted as a double height rectangle, are apparent. The rectangles of word sequences that have been modified are shown one above the other in order to make them visually comparable regarding their changes in length. The rectangular version depicts each remaining, removed, or inserted word sequence in its relative length. The accumulated length of the diffline is therefore longer than the representation of either text.*
*(b) **Trapezoidal diffline**: Trapezoidal and triangular glyphs are employed to illustrate the differences in length of modified sections. Triangles represent newly inserted or completely removed text. A glyph that is composed of two triangles shows modifications of similar length. The idea behind this glyph alphabet was to reduce the overall number of visual items. Precisely one item for each kind of event is drawn in order to make the recognition more straightforward.*
*(c) **Condensed diffline**: This diffline is aligned to the length of the suspicious text and consists of rectangular representations for each section. The darker rectangles show identical text. The light gray rectangles show newly inserted text. The small line-shaped glyphs atop the other rectangles provide hints of textual changes. If the line is as long as the rectangle below, it implies that the original text fragment has been at least as long or even longer than the suspicious one.*

many years, they have only used static visualization thus far. Therefore, they enjoyed the general interactivity of the system and its different views of a case. Every one of the experts immediately tried to locate particular finding spots they were familiar with. In this regard, filtering and exploring by sources seems what interests them most, especially identifying and filtering by the finding spots of these sources that were used most in the suspicious document. Their favorite features are:

- Having the ability to see all finding spots at once
- Being able to trace the finding spots back to their sources without, for example, recalling a particular color coding (like in [Use14]).
- Being able to recognize relationships between the distribution across the entire document and particular categories (e.g. Figure 6).
- Having easy access to small sets or individual finding spots via fluent interactions and filtering capabilities is a clear advancement over the existing visualizations.

They were particularly fond of the diffline idea, which they found clear and legible for the intended task of providing a visual pre-assessment to decide which spots are more ambiguous and should be further investigated in detail with the help of the textual view. Two of the experts liked the special wrapping of the text view using the equal parts as a skeleton, whereas the third was more fond of the text blending method. Another experience during the reviews was that, after becoming more familiar with the prototype, the experts started exploring and comparing different cases and

discussing their peculiarities (Figure 6 contrasts cases with different properties). They further suggested introducing an ordering in decreasing length of finding spots grouped by most used sources in order to speed up the review process: if larger fragments are confirmed to be plagiarism, smaller ones can be postponed. In this regard, they preferred an absolute encoding of the length of a finding spot and recommended introducing a possibility to adjust a length threshold to filter finding spots that are too short to be of use.

During lab tours, our prototype became one of the most discussed exhibits. Our guests are usually surprised by the severity of some of the cases (which is made apparent by the overviews) and the pettiness of others, whereas both have received comparable media attention. We originally expected that the difflines reveal distinctly different patterns between categories, e.g. more frequent text modifications in the category Obfuscation or that the difflines look quite similar with only very few modified text fragments for the category Complete Plagiarism. In some cases, one can see quite consistent patterns but, unfortunately, quite often the categories show a wide spectrum of copy-and-paste patterns which leads to interesting questions, such as: are the categories defined by the community itself not discriminative enough? Are they too fuzzy in description, or was the community sloppy in ensuring a consistent categorization?

8. Conclusions and Future Work

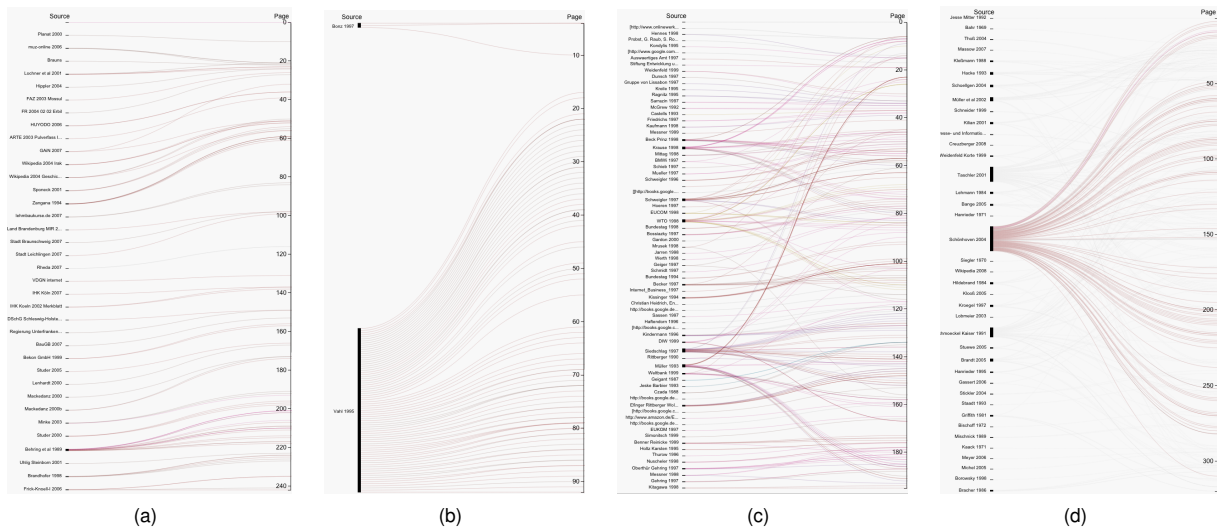We present a new approach for the interactive visual analysis of alleged cases of plagiarism. Our interface is based

Figure 6: *Different methods of plagiarism. Each case differs in length, number of finding spots and sources, categories, and distribution of finding spots. (a) The very few crossings indicate that this suspect worked in a linear way, integrating source by source after another. (b) Only few sources suffice if they can be exploited extensively. (c) This short document employed surprisingly many sources. (d) The suspicious document utilizes a main source across all pages (selected), which indicates that the overall structure of the original work was employed and filled in with other sources (not selected).*

on three levels of abstraction. Our overview displays provide information regarding the structure of the document, the specifics of the finding spots, and how they are related to the original works and authors. The list of difflines provides a compact overview of finding spots, reveals plagiarism patterns by visually encoding the differences and similarities of two text fragments, and directs attention toward further analysis. To this end, a textual side-by-side comparison of original and finding spot can be shown to enable their direct comparison. Our prototype provides effective means to navigate and filter the finding spots and enables direct interaction between finding spot, original, and their diffline. As our study shows, users became quickly proficient with our system and were able to correctly interpret difflines. Furthermore, the reviews by our plagiarism experts confirm that our tool is far more effective than existing static and non-static visualizations. Therefore, we believe that a visual analysis tool like ours will play an important role to verify plagiarism allegations in an effective manner and to convincingly present the evidence to councils or even to the general public.

Further development of natural language processing technologies will possibly lead to automatic categorization of finding spots which is potentially more precise than the community members are today. We are also working on detection algorithms for obfuscation techniques employed in paraphrased text, such as utilizing slightly different words with the same stem, converting verbs into nouns and vice versa, or using synonyms. The difflines should be extended to express and reveal passages that were created with such modifications. However, a particular challenge is the un-

certainty that comes with a machine-generated categorization. Another aspect that should be addressed in the future is the visualization of nonlinear paraphrasing where particular word sequences are shuffled or rearranged in order to mimic autonomous reasoning and deducing. Although barely existing in our manually categorized cases (even in the Shake and Paste and Half Sentence Mending categories), we are certain that such less obvious approaches are used in more cleverly plagiarized documents.

Although our current tool contains some capabilities for group reviews, such as approving finding spots or changing their categorization, other operations to manage complex alleged plagiarism cases are needed: foremost proper user management, as well as an additional top level view, in which several suspicious cases can be depicted at once in an effort to compare them regarding their topics, methods of plagiarism, or shared sources. Once such capabilities are available, further tests involving the community and an integration with our Picapica software are intended.

References

[CDBF10] CHEVALIER F., DRAGICEVIC P., BEZERIANOS A., FEKETE J.: Using Text Animated Transitions to Support Navigation in Document Histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 683–692. 3

[Ger] GERRARD C.: Tableau Colors. http://public.tableausoftware.com/profile/chris.gerrard#!/vizhome/TableauColors/ColorPaletteswithRGBValues. [Online; accessed 2015-02-13]. 7

[GMB*13] GIPP B., MEUSCHKE N., BREITINGER C., LIPINSKI M., NUERNBERGER A.: Demonstration of Citation Pattern Analysis for Plagiarism Detection. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, UK, Jul. 28 - Aug. 1 2013), ACM. 3

[iPa14] IPARADIGMS LLC: TurnItIn. http://turnitin.com/, 2014. [Online; accessed 1-December-2014]. 3

[JGBS14] JÄNICKE S., GESSNER A., BÜCHLER M., SCHEUERMANN G.: Visualizations for text re-use. In *IVAPP 14: Proceedings of the 5th International Conference on Information Visualization Theory and Application* (2014), SCITEPRESS. 3

[KJW*14] KOCH S., JOHN M., WORNER M., MULLER A., ERTL T.: VarifocalReader – In-Depth Visual Analysis of Large Text Documents. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1723–1732. 3

[MG13] MEUSCHKE N., GIPP B.: State of the Art in Detecting Academic Plagiarism. *International Journal for Educational Integrity 9*, 1 (Jun. 2013), 50–71. 3

[Mic14] MICROSOFT SUPPORT: How to Use the Windiff.exe Utility. http://support.microsoft.com/KB/159214, 2014. [Online; accessed 2014-12-01]. 3

[Mis14a] MISC. ANONYMUS AUTHORS: GuttenPlag - kollaborative Plagiatsdokumentation. http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki, 2014. [Online; accessed 1-December-2014]. 2, 7

[Mis14b] MISC. ANONYMUS AUTHORS: VroniPlag Wiki - kollaborative Plagiatsdokumentation (Eine kritische Auseinandersetzung mit Hochschulschriften). http://de.vroniplag.wikia.com/wiki/Home, 2014. [Online; accessed 1-December-2014]. 2, 3, 7

[Mis14c] MISC. AUTHORS OF WIKIPEDIA: Alfred the Great. http://en.wikipedia.org/wiki/Alfred_the_Great, 2014. [Online; accessed 1-December-2014]. 6

[Ope14] OPENDATACITY (DATENFREUNDE UG) AND VEREIN EUROPE-V-FACEBOOK.ORG: Lobbyplag.eu. http://lobbyplag.eu/, 2014. [Online; accessed 1-December-2014]. 3

[PHB*14] POTTHAST M., HAGEN M., BEYER A., BUSSE M., TIPPMANN M., ROSSO P., STEIN B.: Overview of the 6th International Competition on Plagiarism Detection. In *Working Notes Papers of the CLEF 2014 Evaluation Labs* (Sept. 2014), Cappellato L., Ferro N., Halvey M., Kraaij W., (Eds.), CEUR Workshop Proceedings, CLEF and CEUR-WS.org. 4

[Pot] POTTHAST M.: Picapica. http://www.picapica.org. [Online; accessed 2015-02-13]. 1, 4

[Pri15] PRIO INFOCENTER AB: Urkund. http://www.urkund.com, 2015. [Online; accessed 5-February-2015]. 3

[PSRS] POTTHAST M., STEIN B., ROSSO P., STAMATATOS E.: PAN Website. http://pan.webis.de. [Online; accessed 2015-02-14]. 1, 4

[RGP*12] RIEHMANN P., GRUENDL H., POTTHAST M., TRENKMANN M., STEIN B., FROEHLICH B.: WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK's Wildcard Search. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (Sept. 2012), 1411–1423. doi:http://doi.ieeecomputersociety.org/10.1109/TVCG.2012.96. 3

[Sci14] SCIPLORE: CitePlag demonstrates Citation-based Plagiarism Detection (CbPD). http://citeplag.org/ and http://sciplore.org/, 2014. [Online; accessed 1-December-2014]. 3

[SCKP08] SUH B., CHI E., KITTUR A., PENDLETON B.: Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with Wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1037–1040. 3

[Use14] USER8 (PSEUDONYM IN GUTTENPLAG WIKI): Herausragende Quellen. http://de.guttenplag.wikia.com/wiki/Visualisierungen and http://de.guttenplag.wikia.com/wiki/Herausragende_Quellen, 2014. [Online; accessed 1-December-2014]. 3, 7, 8

[VWD04] VIÉGAS F., WATTENBERG M., DAVE K.: Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2004), CHI '04, ACM, pp. 575–582. 3

[Web] WEBER-WULFF D.: Tests of Plagiarism Software. http://plagiat.htw-berlin.de/software-en/. [Online; accessed 2015-02-13]. 3

[Web14] WEBER-WULFF D.: *False Feathers: A Perspective on Academic Plagiarism*. Springer, Berlin, 2014. 3

[WMTZ] WEBER-WULFF D., MÖER C., TOURAS J., ZINCKE E.: Plagiarism Detection Software Test 2013. http://plagiat.htw-berlin.de/software-en/test2013/report-2013/. [Online; accessed 2015-03-01]. 3

[WVH07] WATTENBERG M., VIÉGAS F., HOLLENBACH K.: Visualizing Activity on Wikipedia with Chromograms. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II* (Berlin, Heidelberg, 2007), INTERACT'07, Springer-Verlag, pp. 272–287. 3