CATview - Supporting the Investigation of Text Genesis of Large Manuscripts by an Overall Interactive Visualization Tool

Marcus Pöckelmann
marcus.poeckelmann@informatik.uni-halle.de
Institute of Computer Science
Martin Luther University Halle-Wittenberg

André Medek né Gießler
andre.medek@informatik.uni-halle.de
Institute of Computer Science
Martin Luther University Halle-Wittenberg

Paul Molitor
paul.molitor@informatik.uni-halle.de
Institute of Computer Science
Martin Luther University Halle-Wittenberg

Jörg Ritter
joerg.ritter@informatik.uni-halle.de
Institute of Computer Science
Martin Luther University Halle-Wittenberg

1. Introduction

Manuscripts that went through many editions being revised, augmented or shortened from time to time are of special interest for historians, sociologists and philologists. Identifying the text passages that have been changed and presenting them in a synoptic view or in critical editions plays a central role in these fields. In order to support the investigation of text genesis appropriate information technology tools should be available. In addition to flexible and effective implementations for comparing variants of a manuscript (e.g. see [1]) the text differences found by such tools should be visualized in an appropriate way allowing scholars effectively navigate and explore the differences. With the *Colored & Aligned Texts view* (*CATview*) we present such an interactive visualization tool in this paper. Starting with a general map of the text witnesses' aligned segments, scholars can search, find and zoom to specific text passages of interest, which are colored according their grade of revision or highlighted as striking depending on the user settings. In collaboration with a synoptic presentation of the text variants *CATview* is an useful add-on for both, web-based editions and frameworks for the editing process.

2. Motivation of the Work

CATview is motivated by an ongoing investigation of Abbé Raynal's "*Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes*", a text about the negative influence of European civilization during the colonization of East and West Indies. The first edition was published in 1770 (Amsterdam). After the manuscript was forbidden, Raynal released expanded editions in 1774 (The Hague) and 1780 (Geneva). A last edition appeared post mortem in 1820, which Book 6 for instance consists of 52.372 words whereas the edition of 1770 features 28.451 words.

The work is part of the project "*SaDA – Semi-automatic Difference Analysis of complex text variants*" [2] funded by the German Federal Ministry of Education and Research (BMBF). The visualization tool presented in this paper is embedded in an overall framework for text comparison that acts according to the following workflow. First, each of the given witnesses of a text is divided into segments, either paragraphs or sentences. Then, fingerprints of the segments are computed allowing an assignment between them, i.e. an alignment of the segments. Finally, the aligned segments are compared in detail and presented in a synoptic manner with apparatus.

CATview is an additional component built on top of this workflow to lighten the navigation within the synopsis and facilitate the investigation of text differences by preparing the collected data in a clear manner.

3. Previous Work

The visualization of text differences and structure with aspects of navigation is already handled by several web-based tools. One example is the navigation bar of Perseus Digital Library [3]. It illustrates the text structure by a set of iconic bars, e.g. for books, sections or verses. The individual bars are divided into segments whose lengths indicate the lengths of the text passages they represent and they are linked to.

The histogram view of Juxta [4] exceeds the pure aspect of navigation. This small pop-up dialog allows scholars exploring 'the overall rate of change across the witnesses' [5] of a manuscript, but only works for the comparison of one specific witness against others.

In Ben Fry's remarkable visualization of “On the Origin of Species” [6] the text structure is abstracted for an overall view and multiple witnesses' revisions are visualized by colors. This is done in a fixed animation. At any point merely the latest revision is shown for each word. The text itself will pop-up as dialog above the mouse cursor without a steady representation. Thus, the text cannot be searched and frequent mouse moving is necessary for reading longer portions. Besides control elements for the animation there is no user interactivity such as zooming or further options to manipulate the presentation.

4. Features of *CATview*

In the first place, *CATview* illustrates the aligned segments in a tabular manner. Each text witness is represented by one row. Their segments are abstractly represented as rectangles. If two segments of different witnesses are aligned by step 2 of the workflow described above, they share the same column. The differences of the aligned segments found in the following step 3 are visualized in an aggregated form by the intensity of the rectangles’ color. To lighten navigation the rectangles are also links that will scroll the synoptic view to the corresponding segment when clicking on a rectangle. On the other hand the current position within the synopsis is represented in *CATview* with a marking. Additional markings to highlight search results and further information with respect to the alignment can be displayed.

4.1 Illustrating the Alignment

The presentation of the alignment in a tabular manner helps reviewing the overall structure to see patterns of revisions (Figure 1). It is also useful to evaluate the alignment in purpose to improve the underlying algorithms as it calls attention, e.g. to falsely aligned segments. To see more details the user can zoom in and out with the mouse wheel and slide the currently selected excerpt by dragging it with the mouse (Figure 2). Thereby, the option to change the order of the rows, i.e. the witnesses, and a consecutively numbering of the columns lighten the orientation. This basic functionality allows to easily identify relevant portions of text that have been added or removed during the authors' revisions. Figure 2 illustrates this feature: one paragraph was added to witness H20 (which denotes the 1820 edition) between the aligned paragraphs at columns 59 and 61, where paragraphs of all four witnesses are aligned.

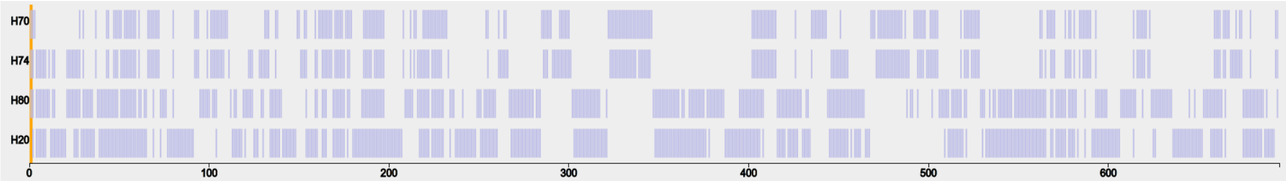

Figure 1: *CATviews*' general map of aligned paragraphs for the four witnesses of the *Histoire des deux Indes* (book 6).

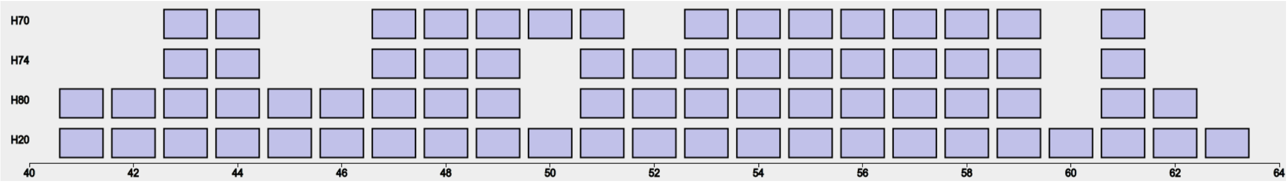

Figure 2: A smaller excerpt of the alignment seen in Figure 1 by using the built-in zoom.

4.2 Coloring of Rectangles

The intensity of colors assigned to individual rectangles denotes the degree of similarity between the corresponding aligned segments and can be determined for instance by the ratio of the segments' number of differences to the amount of text. The color ranges from a standard light blue, which indicates similarity, up to a strong dark blue for rather extensive revisions of the text. This additional information effectively helps locating hot spots of revisions. Figure 3 and 4 show *CATview* with enabled colors for two levels of zooming.

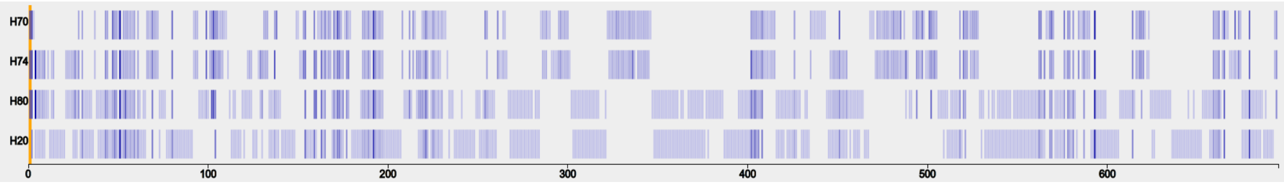

Figure 3: The general map of Figure 1 automatically enriched by colors for the paragraphs, based on the aggregated information gathered by a detailed difference analysis.

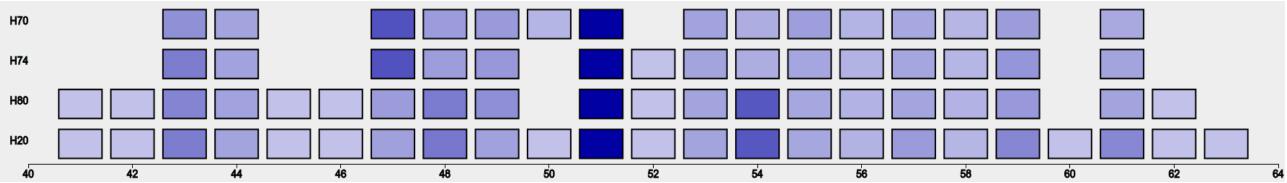

Figure 4: An excerpt of Figure 3 on a higher zooming level: columns in light blue express a strong similarity between the aligned paragraphs, whereas the dark blue at the columns 47, 51 and 54 indicate a strong difference. In column 54 a major revision first appeared in witness of 1780 with siglum H80.

4.3 Benefits of Additional Markings

Another feature of *CATview* is to connect rectangles of different witnesses with a line if the corresponding segments are supposed to be similar but can not be aligned due to a conflicting data situation. This situation occurs if the alignment is blocked by an assignment of other segments, e.g. in case of transposed or merged segments. The additional lines lighten the identification of such cases for a closer look.

Furthermore, the tool can display search results by highlighting single rectangles or full columns that contain the search phrase with a colored background. This feature helps effectively estimating the distribution of a subject within all witnesses. Figure 5 illustrates the described markings.

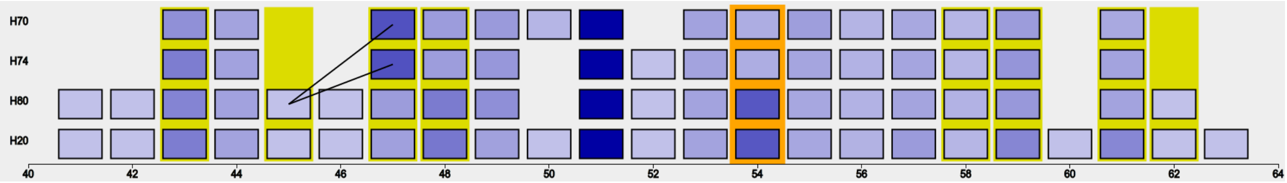

Figure 5: Additional markings in *CATview:* the bar at column 54 indicates the current scroll position in the corresponding synoptic view. Search results for the key word “Colomb” (French for “Columbus”) are highlighted by a colored background (columns 43, 45, 47, 48, 58, 59, 61 and 62). The lines from 45 to 47 show paragraphs supposed to be similar that could not be aligned as in H80 Raynal split paragraph 47 of H74 into two paragraphs (45 and 47) and added an additional paragraph in between.

4.4 Other Planned Features

There are further features currently under discussion. The height of rectangles could present the size of the corresponding segments, displaying the original page numbers for each witness on higher zooming levels could lighten the orientation and multiple aligned segments could be aggregated on lower zooming levels, to name a few.

4.5 Technical Remarks

CATview is designed for web applications and implemented in JavaScript as a Singleton Object. Thereby functionality of the JavaScript libraries D3.js [7] and jQuery [8] is used to generate and manipulate the SVG-Image, which contains the graphical elements described above. A publication of the finalized tool as open source is planned.

Funding

This research was funded by the German Federal Ministry of Education and Research (BMBF) [grant number 01UG1247] as part of the project “Semi-automatische Differenzanalyse von komplexen Textvarianten” under the direction of Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter and Prof. Dr. Hans-Joachim Solms.

References

[1] **Medek, A. and Pöckelmann, M. and Bremer, Th. and Solms, H.J. and Molitor, P. and Ritter. J.** (2015). Differenzanalyse komplexer Textvarianten: Diskussion und Werkzeuge. *Datenbank-Spektrum, Themenheft "Informationsmanagement für Digital Humanities",* Ed. Henrich, A. and Heyer, G.

[2] "SaDA - Semi-automatische Differenzanalyse von komplexen Textvarianten". http://www.informatik.uni-halle.de/sada (accessed October 28, 2014)

[3] "Perseus Digital Library". http://www.perseus.tufts.edu (accessed October 28, 2014)

[4] "juxta - Compare · Collate · Discover". http://www.juxtasoftware.org (accessed October 28, 2014)

[5] "juxta commons". http://www.juxtacommons.org (accessed October 28, 2014)

[6] **Fry, B**. (2009). On the Origin of Species: The Preservation of Favoured Traces. http://www.benfry.com/traces (accessed February 17, 2015)

[7] "D3 - Data-Driven Documents". http://d3js.org (October 28, 2014)

[8] "jQuery - write less, do more.". http://jquery.com (October 28, 2014)

Summary

In this paper we present an interactive visualization tool, an useful add-on for web-based editions and frameworks for the editing process, to effectively navigate and explore the differences of multiple text variants in a graphical overview. Despite its relative simplicity, our tool named *Colored & Aligned Texts view* (*CATview*) has universal applications in the field of text comparison according to the available data. Starting with a general map of the text witnesses' aligned segments, scholars can search, find and zoom to specific text passages of interest, which are colored corresponding to the grade of revision or highlighted as striking depending on the user settings. In collaboration with a synoptic representation of the text *CATview* effectively supports scholars in the investigation of differences between several variants or the full genesis of a text as well as improving the underlying algorithms for text comparison.