Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Documents

Carlos Monroy, Rajiv Kochumman, Richard Furuta, and Eduardo Urbina

TEES Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA
(979) 845-3839
{cmonroy, rajiv, furuta, e-urbina}@csdl.tamu.edu

Abstract. In this paper we describe ItLv (Interactive Timeline Viewer), a visualization tool currently used to depict the variants obtained in a textual collation. A textual collation is a process in which a base text is compared against several comparison texts to identify differences (variants) among them. The interactive options of ItLv provide different abstractions of a dataset by enabling the presentation and exploration of the relationships that exist within the dataset. Applying ItLv to the dataset resulting from a collation therefore helps understand the relationships among the texts. The example dataset used in this paper is a collation of six early editions of Cervantes' *Don Quixote*.

1 Introduction

We are in the process of creating an Electronic Variorum Edition (EVE) of Miguel de Cervantes y Saavedra's (Spain, 1547-1616) *Don Quixote* [4,5,6]. The EVE will be included in the Cervantes Digital Library (CDL), part of the ongoing Cervantes Project [13]. Our EVE will contain multiple copies of all the early significant editions of this important work, in interlinked facsimile and textual forms. We also are creating critical editions of the work, reflecting the result of scholarly interpretation and emendation of the work, as well as associated commentary, explanation, and other forms of annotation.

An EVE is created with the MVED [5], a collation tool in which a selected base text is compared against several comparison texts to identify the differences (variants) among them. Much of the work of the scholar will involve understanding the variants. Consequently, we apply a visualization tool (ItLv) to depict those variants as well as to provide information about each of them—for example, where in the text it appears, its length in characters, its content, and an image of the original page in the book. Moreover, we can say that ItLv can also be applied to any other set of texts for which a variorum edition needs to be created.

In the context of digital libraries several approaches to provide visualization interfaces have been advanced to represent collections of books, journals, and magazines [7]. Furthermore, there are tools that enable users to search for specific attributes in the items of a collection [8]. Once an item has been found, the user then

needs a browsing mechanism that enables the exploration of the item in further detail [9]. In addition to specific attributes, we are finding that it is useful to apply visualization mechanisms to depict the variants among different editions of the same book as well as among different copies of the same edition.

In terms of comparing different copies of the same text, the Digital Variants Browser (DV Browser) [3] is a system that enables users to "analyze several different versions or writing stages of the same text providing an overview of the entire text material." However, in our case we are not visualizing different stages of the same text, but several different "final versions" of the same text, that is, several editions of the same book.

The DV Browser is based on the Flip Zooming technique [2], which provides focus+context to visualize information. The information is divided into tiles, the tile with the focus is positioned in the center of the display and the rest of the tiles are placed around it. This is a very useful approach for comparing several texts. However, in terms of visualization, it depicts each version of the text as one variant.

Analyzing different versions of software can be seen as a similar process to what we are doing. In this context, Baker and Eick [1], propose a system that enables users to visualize not only the structure of a big software project, but also the evolution of the source code. However, in the case of different versions of software, variants are expected, since new functionality requires new code; whereas, in our analysis, the texts in different editions "should be" the same, since the author has not modified the text. However, due to printing errors, or compositor's preferences variants are introduced between two editions or even between two copies of the same edition; therefore, our focus is to visualize those variants.

Plaisant, et al. [11], describe the use of LifeLines, which is a tool that depicts personal histories using a timeline-based visualization. The use of LineLines helps users to analyze a dataset of discrete events, especially because "LifeLines reduce the chances of missing information, facilitate spotting anomalies and trends, streamline access to details…." These are some reasons why we find a timeline representation useful to analyze variants among several texts. In this paper, we describe our use of ItLv and its functionality in visualizing collation results. Interactive functions allow generation of different representations of a dataset, which allows users to develop a better understanding of the relationships among variances identified in a collation.

2 Visualizing the Results of a Textual Collation

ItLv is based on earlier work in our center by Kumar [7] and applies a timeline-like visualization metaphor for the purpose of representing datasets whose elements are ordered by at least some of their attributes. In the EVE application discussed in this paper, we are depicting variants among six editions of *Don Quixote*. *Don Quixote* was first published in Madrid in 1605. The Madrid 1605 edition is called the *princeps* and is used as base text in this collation. In addition to the princeps, the other five editions used are Valencia 1605, Madrid 1605 (Second Edition), Brussels 1607, Madrid 1608, and Madrid 1637. Each of these five editions is compared in turn to the base text, and figure 1 depicts the results of this collation for the first chapter of the

novel. Here, the Y-axis represents the differences between each of the five editions and the princeps. The Y-axis is arranged in chronological order from bottom to top; for example, the bottom-most line represents the differences between the Madrid 1605 princeps and the Valencia 1605 edition, while the top line represents differences between the Madrid 1605 princeps and the Madrid 1637 edition. The X-axis represents the offset in the text; a value of zero represents the beginning of the text, and the larger the value the closer to the end of the text. Each variant is depicted as a rectangle; the height of the rectangle represents the length of the variant in characters.

By clicking on any rectangle a new window will be displayed (figure 2). This new window depicts all the attributes of the variant represented by that rectangle; (a) includes most of the string and numerical attributes, (b) depicts the text of the variant, and (c) provides an image of the page where that variant appears in the original book. This option provides a context and detail for any variant in the collation. (Note that the bold arrow, letters and square brackets in figure 2 are not part of ItLv; they were included to explain the figure.)

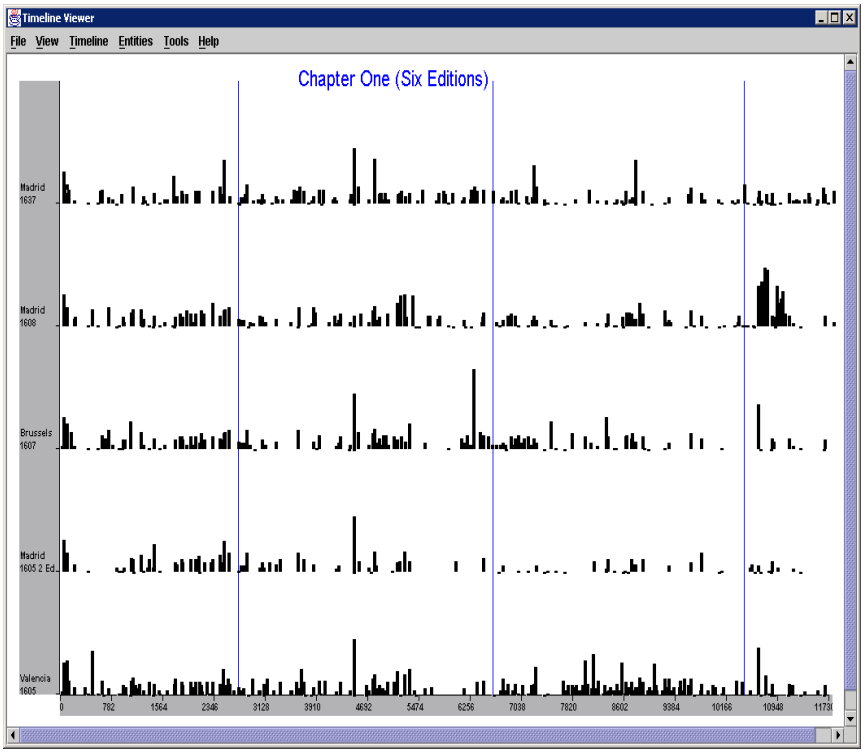

Fig. 1. Variants of the collation among six editions of *Don Quixote*

The results of a collation can be depicted in different ways based on the attribute selected for the Y-axis. This enables the user to analyze the same dataset from different perspectives. Following we discuss two examples. In figure 3, instead of presenting the editions on the Y-axis, as it is in figure 1, a string of five characters containing Y's and N's is displayed. This string represents in what texts a variant was found, for example the string "YYNNY" means that a particular variant was found between the base text (the princeps) and the first, second, and fifth comparison texts; i.e., that corresponds to the editions of Valencia 1605, Madrid 1605 (Second Edition), and Madrid 1637 respectively. Presentation of a variance summarization helps highlight anomalies (e.g., editions with unusual differences from the others) as well as "families" of editions (i.e., editions containing similar patterns of variances suggesting a relationship in their derivation).

Figure 4 shows the same results as in figure 3. On the left is the same display as figure 3's with some clusters highlighted. It is also possible to observe certain clusters of variants for the following combinations of texts: "YNNNN" with a high concentration of variants at the end of page two and in two thirds of page three. For "NNNNY" there are two clusters of variants, one in pages one and the beginning of page two, and the other in page three. On the right, a sorted list shows the number of variants for all the 31 possible combinations of texts in the collation. As the number of texts increases, the number of combinations in which variants can appear increases also, therefore, using this attribute can be useful only for a small number of texts, e.g., less than seven. An alternative method to overcome this problem can be creating categories based on the number of texts in which the variants appear. On each screen, only those variants present in the same number of texts would be depicted regardless of the text they appear on. Thus, one screen will depict the first category, i.e., those variants that appear on only one text; another screen will depict those variants that appear on two texts (the second category), and so on.

Figure 5 shows a different presentation of the results of the collation. This time, the results are grouped by the length of the variant in characters (depicted on the Y-axis). In figure 1, the size of the rectangles is proportional to the length of the variant. Therefore shorter variants have smaller rectangles, whereas larger variants have bigger rectangles. Nevertheless, selecting the length of the variant as the Y-axis attribute and depicting the rectangles proportional to the length of the variant would be redundant, that is the reason all the rectangles have the same size. In this context, small variants are represented by those rectangles depicted at the bottom of figure 5, whereas large variants are represented by those rectangles depicted at the top of figure 5.

ItLv provides an option to represent the number of variants in each element of a category both as a table and a graph (figure 6), the user can sort any of the columns either in ascending or descending order. In this example, the table and graph shows the number of variants for each size. These values are taken from the results of the collation depicted in figure 5. In this particular case, variants of length 3, 1, and 0 characters are on the three top cases. Variances of length zero mean that a variant was present in the base text but not in the comparison text, therefore a value of length zero is given.

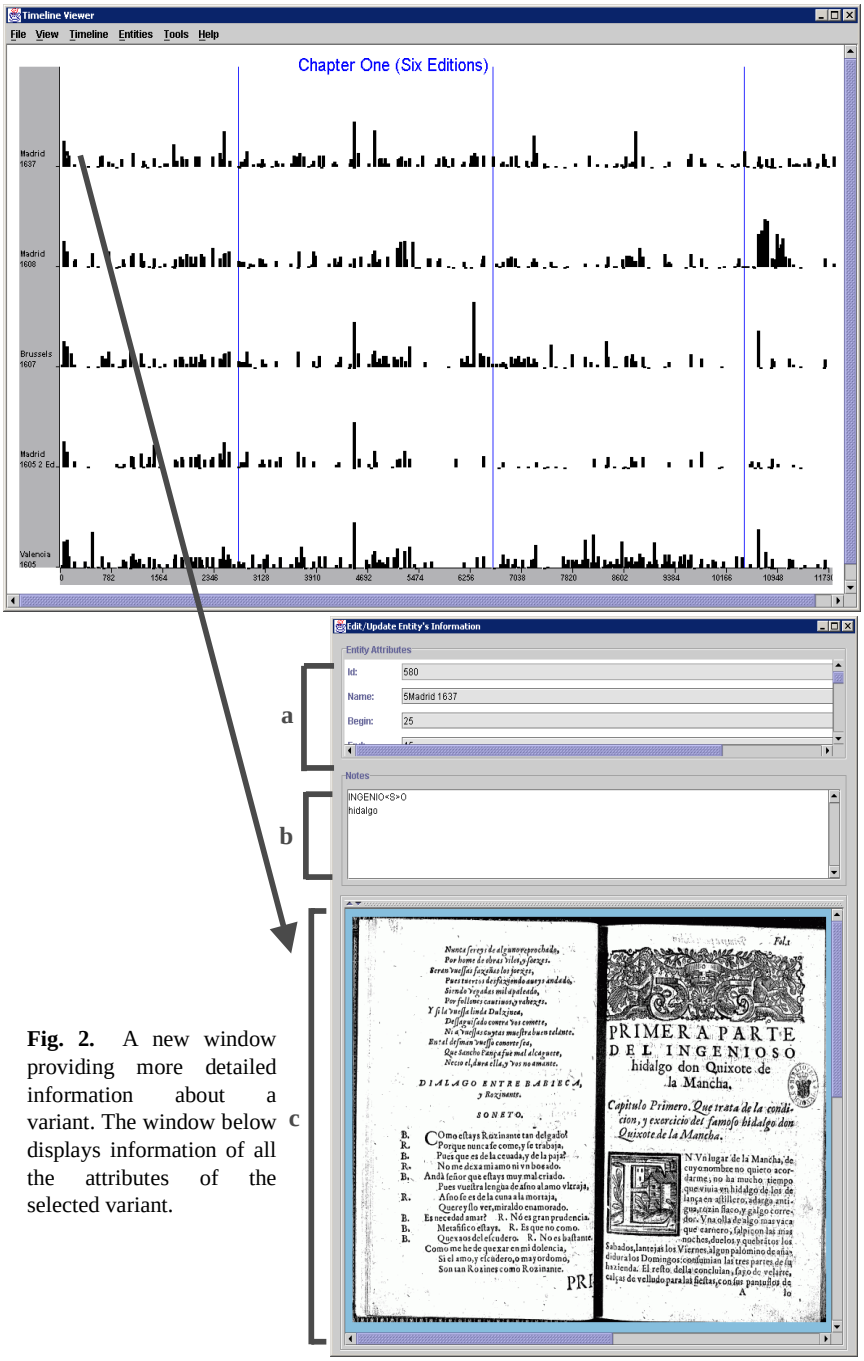

Fig. 2. A new window providing more detailed information about a variant. The window below displays information of all the attributes of the selected variant.

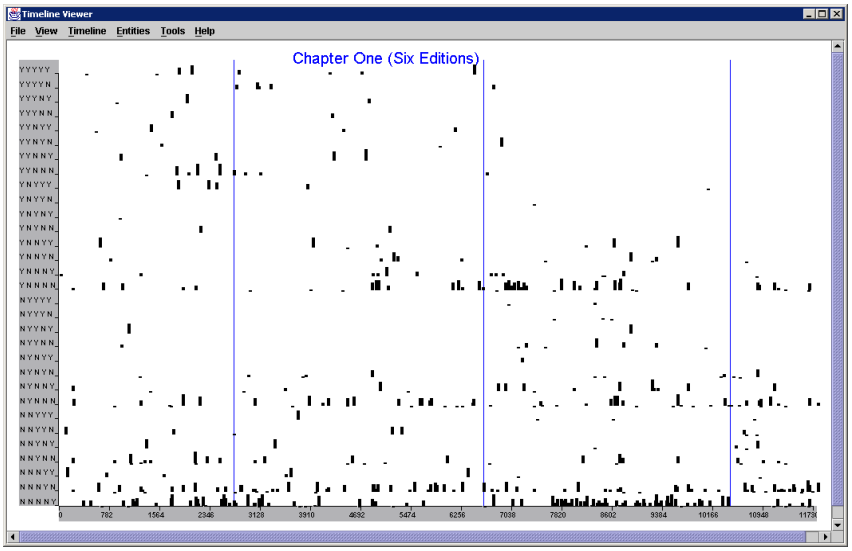

Fig. 3. A visualization of the variants categorized by the texts they appear on

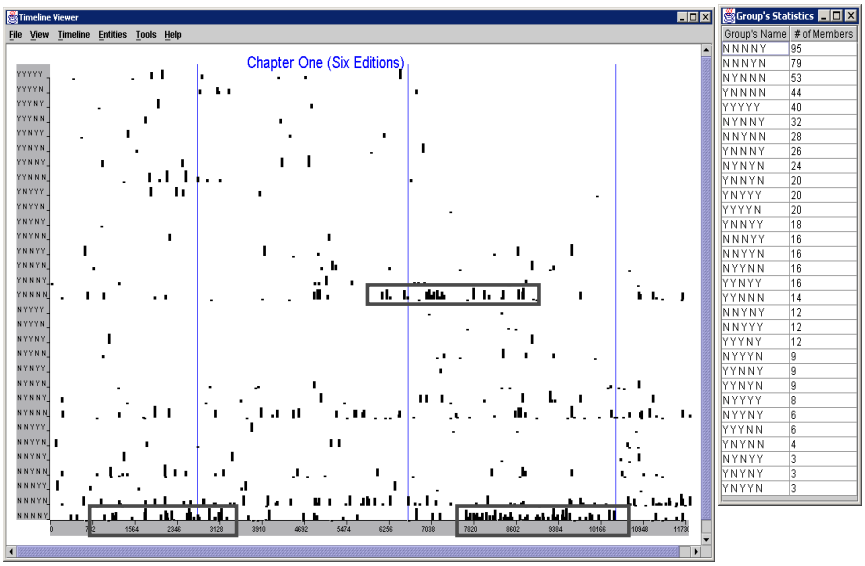

Fig. 4. A visualization of the variants categorized by the texts they appear on and some clusters of variants. The three horizontal rectangles were added manually to highlight three clusters of variants

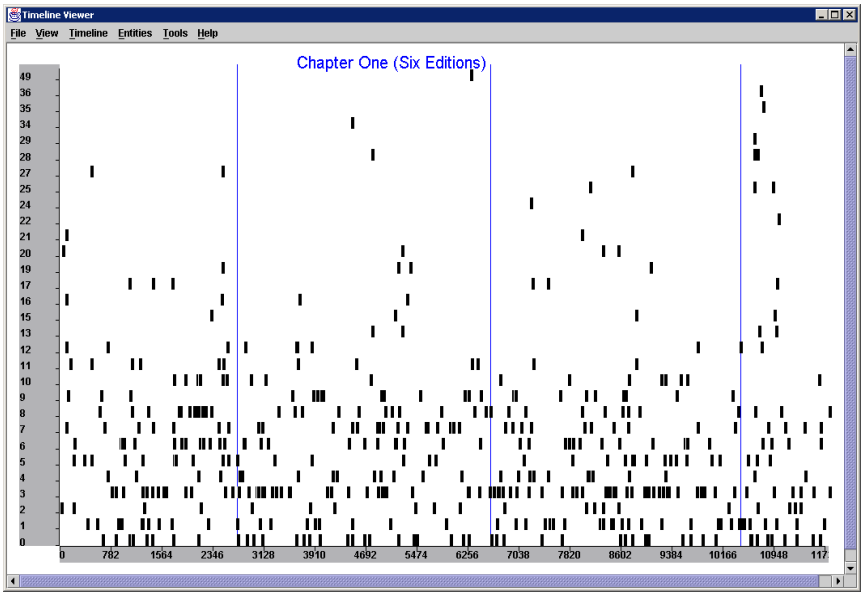

Fig. 5 Variants categorized by length in characters

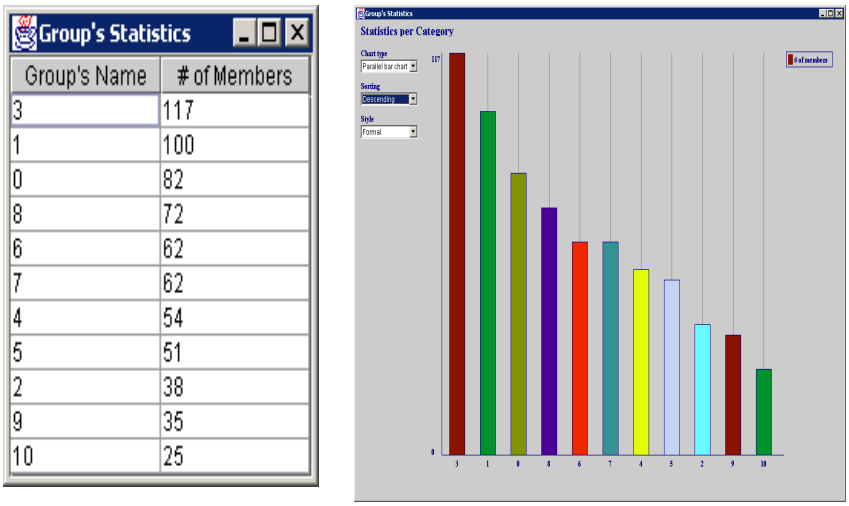

Fig. 6 A table and a graph displaying the number of variants of each size in the collation

In figure 7, the use of a pop-up window is presented. This pop-up window is activated when the mouse cursor is positioned over one of the variants. The information displayed in the pop-up window can be set in advance by the user, and modified at any time. In this example, the attributes selected are the offset where the variant begins in the text, the content of the variant, and the image of the page in the original book where the variant was found. Two lines, one horizontal and the other vertical show the context of the variant. All variants with some similarities are highlighted in a different color, in this particular case the variant is between the word "que" and the abbreviation q̃ , which we encode in our transcriptions as "<q>".

Manipulating the dataset to explore different subsets can be achieved by using the filter option. This option enables users to filter variants based on a logical condition applied to any of the attributes. For example a user can specify that only entries containing a particular substring of characters should be displayed in order to observe whether that substring appears with some frequency across the texts. Figure 8 depicts those variants that include the abbreviation q̃. This shows that this abbreviation was used in page one and in part of page two. Therefore, we can raise the hypothesis that this abbreviation was the preference of the compositor in charge of these two pages.

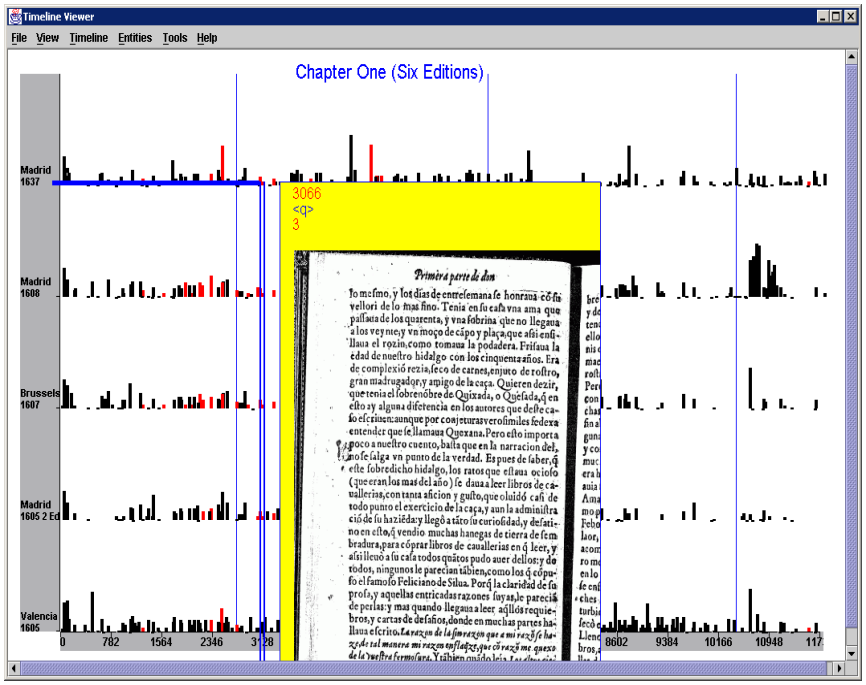

Fig. 7 A pop up window displaying some attributes of a variant: a) offset, b) the variant, c) the length in characters, and d) the image of the text

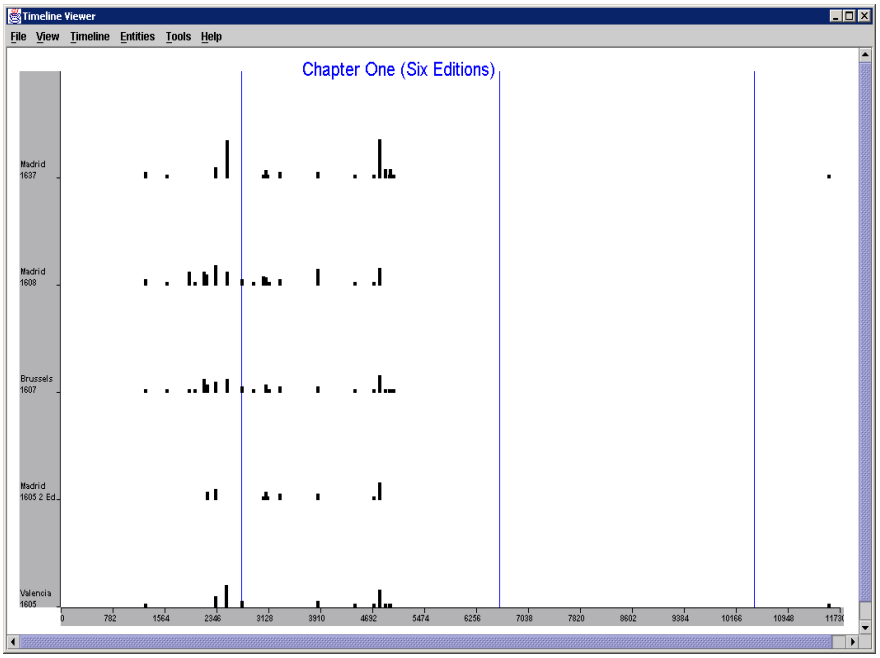

Fig 8 Depicting only those variants including the abbreviation q̃

As pointed out earlier, the ability to remove and add variants to the display is a very useful option provided by the ItLv. Another scenario could be for the user to remove those variants belonging to the category of punctuations, e.g., commas, semicolons, and so forth, prior to analyzing the results of a collation. By using the filter option, the user is able to hide those variants and focus on the rest. The variants can also be displayed again by removing the filter that has been applied.

When depicting the results of a collation, all variants are displayed in relation to the offset in the base text. Let us take texts A and B as an example, and let us assume that text A is the base text and text B is the comparison text. If a whole paragraph in text B appears in a different page than it appears in text A, those variants will be depicted relative to their offset in the base text (text A) not their offset in the comparison text (text B).

We have also used the ItLv to validate the results of the collation [12]. This enables users to analyze a dataset in which false variants have been either removed completely or minimized as much as possible from the display. In this context, we are using ItLv to investigate hypotheses about the origins of *Don Quixote's* early editions and their interrelationships.

3 Conclusion and Future Work

The main goal of this paper is to present the use of a visualization tool (ItLv) to depict the results of collating several copies of one text. Using ItLv's interactive features enables scholars to explore the similarities and differences among the texts in further detail. In the present case, ItLv is primarily for the use of domain experts in textual analysis, since it is applied to textual editing, and specifically to the analysis of textual variants in the creation of a variorum edition, as well as to present in visual format the results of comparing the texts included in the preparation of the edition.

ItLv as a visualization tool provides a flexible interface to visualize the differences among versions of a text, for example several copies of the same edition of a book or among different editions of a book. Interactive manipulations of the display provide both context and detail for the variants, for example by linking a variant to the image of the page in which it appears. Different visualizations of the same dataset provide different abstractions that can enable users to recognize and further understand the differences and similarities among texts.

We continue to explore the application of ItLv to visualization of the differences found within a collection of documents. Through this exploration, we hope to gain a better understanding of the changes that have been made among the editions, as well as to gather evidence of relationships between editions—for example, evidence about which earlier editions were consulted by the publishers of the later editions.

A preliminary use of ItLv showed good results as a visualization tool to explore the similarities and differences, and to identify patterns for a set of texts. However, when the number of chapters increases, we expect challenges such as how to identify patterns present in two non-contiguous chapters, especially if the chapters are far apart from one another. Along with the number of chapters is the need to provide a mechanism that enables the user to navigate through all the screens.

Presently, the image of the page where a variant appears is displayed, but there is no any indication of where in the image that variant appears. We are currently exploring the use of an additional visual aid to indicate the region in the image where the variant appears; this will help the user to find the variant in the image with more accuracy.

It is important to point out that in this paper we report the initial results of the use of ItLv. At this stage we have not performed extensive user testing or usability experiments. Presently, the preliminary experiments include small chapters only. We therefore, expect to start user testing soon in order to evaluate the user's experiences after incorporating longer chapters.

4 Acknowledgements

This material is based upon work supported by the National Science Foundation under grant no. IIS-0081420. Support for this study was also provided in part by the Interdisciplinary Research Initiative Program, administered by the Office of the Vice President for Research, Texas A&M University.

5 References

1. Baker, M., Eick, S., "Space-Filling Software Visualization." Journal of Visual Languages and Computing. Vol. 6, No. 2, June 1995. pp. 119-133.
2. Bjork, S. "Hierarchical Flip Zooming: Enabling Parallel Exploration of Hierarchical Visualizations" Proceedings of the Working Conference on Advanced Visual Interfaces. Palermo, Italy 2000, pp. 232-237.
3. Bjork, S., Holmquist, L., "Exploring the Literary Web: The Digital Variants Browser." Proceedings of Literature, Philology and Computers, Edinburgh, UK, 1998.
4. Furuta, R., Kalasapur, S., Kochumman, R., Urbina, E., Vivancos-Pérez, R., "The Cervantes Project: Steps to a Customizable and Interlinked On-line Electronic Variorum Edition Supporting Scholarship", Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 2001, pp. 71-82.
5. Furuta, R., Hu, S., Kalasapur, S., Kochumman, R, Urbina, E., Vivancos-Pérez, R., "Towards an Electronic Variorum Edition of Don Quixote", Proceedings of the first ACM/IEEE-CS joint conference on Digital Libraries, Roanoke, Virginia, 2001, pp. 444-445.
6. Kochumman, R., Monroy, C., Furuta, R., Goenka, A., Urbina, E., Melgoza, E. "Towards an Electronic Variorum Edition of Cervantes' Don Quixote: Visualizations that support preparation", Proceedings of the second ACM/IEEE-CS joint conference on Digital Libraries, Portland, Oregon, July 2002, pp. 199-200.
7. Kumar, V., Furuta, R., Allen, R., "Metadata Visualization for Digital Libraries: Interactive Timeline Editing and Review", Proceedings of the third ACM conference on Digital Libraries, Pittsburgh, Pennsylvania, May 1998, pp. 126-123.
8. Maayan, G., Feitelson, D., "Hierarchical Indexing and Document Matching in BoW", Proceedings of the first ACM/IEEE-CS joint conference on Digital Libraries, Roanoke, Virginia, 2001, pp. 259-267.
9. Marshall, C., Price, M., Golovchinsky, G., Schilit, B., "Designing e-Books for Legal Research", Proceedings of the first ACM/IEEE-CS joint conference on Digital Libraries, Roanoke, Virginia, 2001, pp. 41-48.
10. Monroy, C., Kochumman, R., Furuta, R., Urbina, E., Melgoza, E., and Goenka, A., "Visualization of Variants in Textual Collations to Analyze the Evolution of Literary Works in The Cervantes Project", Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, September 2002, pp. 638-653.
11. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B. "LifeLines: visualizing personal histories", in Proceedings of CHI'96, Vancouver, BC, Canada, April 14-18, 1996, pp. 221-227.
12. "The Cervantes Project", Center for the Study of Digital Libraries, Texas A&M University. http://www.csdl.tamu.edu/cervantes