UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A PIXEL-BASED FOCUS+CONTEXT TECHNIQUE FOR VISUALIZING VARIATION IN CLASSICAL TEXT

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

BHARATHI ASOKARAJAN
Norman, Oklahoma
2016

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/34669

A PIXEL-BASED FOCUS+CONTEXT TECHNIQUE FOR VISUALIZING VARIATION IN CLASSICAL TEXT

A THESIS APPROVED FOR THE
DEPARTMENT OF ENGINEERING

BY

Dr. Chris Weaver, Chair

Dr. Christan Grant

Dr. Charles Nicholson

© Copyright by BHARATHI ASOKARAJAN 2016
All Rights Reserved.

Acknowledgements

I am always grateful to my advisor Dr. Chris Weaver, for his guidance through this research work and showing me an exciting new world of data visualization. I am thankful to Dr. Sam Huskey, for providing us the Latin data set and discussing the various challenges involved in critical analysis of texts. This indeed helped us formulate the research problem and work towards a solution. I would also like to thank Dr. June Abbas, for her motivation and interest. Thanks to Dr. Ronak Etemadpour from Oklahoma State University, for helping me evaluate the visualization tools. My sincere thanks to Dr. Christan Grant and Dr. Charles Nicolson, for always being curious to know about my research progress and ready to help attitude. Thanks to Dr. Sridhar Radhakrishnan, for encouraging me to pursue my goals academically. Finally, I thank my labmates Sudarshan, Vamshi, Silvia, Emily, and Jeyachandran for being friendly and helpful.

My special thanks to my husband and inspiration, Dr. Mahendran Veeramani for being patient and supportive throughout this experience. To my beloved daughter Saathvika, I would like to express my thanks for being such a nice girl and always cheering me up. I am thankful to my parents, sister, and in-laws for being as enthusiastic as I am with me pursuing research.

This thesis work was supported as a part of the Digital Latin Library project by a grant from the Andrew W. Mellon Foundation.

Table of Contents

1 Introduction **1**

2 Background and Related Work **5**
2.1 Visualization 5
2.2 Interface Schemes For CMV 6
2.3 Pixel-based Visualization 8
2.4 Hierarchical Visualization 11
2.5 Text Visualization 12

3 Textual Variant Analysis - Domain Problem **14**
3.1 Sensemaking in Humanities Research 14
3.2 Vernaculars of Latin Textual Criticism 15
3.3 Variant Analysis Tools 19

4 Pixel-based Hierarchical Focus+Context Visualization **22**
4.1 Design Specification 22
4.1.1 Overview 24
4.1.2 Hierarchical Views 27
4.1.3 Interaction and Query Capabilities 28
4.2 Data Set 32
4.3 Prototype Tool Implementation 34
4.4 Evaluation 35
4.4.1 User Study 36
4.4.2 Experimental Settings 37
4.4.3 Study Metrics 38
4.4.4 Study Results and Discussion 40
4.5 Summary 42

5 TexTile **44**
5.1 Design Specification 45
5.1.1 Tiered Views 47
5.1.2 Navigation 49
5.1.3 Pixel Color 50

5.1.4 Interaction and Query Features 51
5.2 Evaluation 54
5.2.1 Experimental Settings 54
5.2.2 Study Results And Discussion 55
5.3 Summary 64

6 Conclusion **66**

A Evaluation Material - Study A **74**

B Evaluation Material - Study B **80**

C Study Authorization Documents **87**

List of Tables

4.1 Commonly used string distance metrics 25
4.2 Representative tasks used to check usability of the prototype tool, grouped into three kinds of synoptic activity (Interaction, Navigation and Perception). 38

5.1 Quantitative tasks used to assess the usability of TexTile. The task group column indicates the classification of each task based on the three synoptic task groups: Interaction-related(I), Perception-related (P), and Navigation-related (N). 56
5.2 Qualitative tasks used to check usability of TexTile, inspired by Wehrend et al. [58]. Responses were entered on a 1 ("easy") to 5 ("difficult") Likert scale. 59

List of Figures

2.1 Different pixel arrangement techniques. (a) line-by-line (b) column-by-column (c) left-right (d) top-bottom 10
2.2 General layout of hierarchical visualization: (a) Connections in a tree; (b) Enclosure in a nested arrangement. 12

3.1 The first page of Giarratano's 1910 edition of *Calpurnius Siculus* [53], including: (a) poem numbers, (b) line numbers, (c) base text, and (d) critical apparatus. An apparatus entry (f) lists past variants (g) with the one chosen for use in the text, called the *lemma* (e). 17

4.1 The initial prototype tool for analyzing variation in classical Latin texts. The layout consists of: a base text view (indicated by label "BASE TEXT") with a global filter option (A); an "OVERVIEW" of variation between lemmata (B) and witnesses (C); a draggable window to select a range of tokens (D); and a hierarchy of drill-down views that summarize the variation count at the level of pages (E), lines (F), and words (G) called "PAGE VIEW", "LINE VIEW" and "WORD VIEW" respectively. Mousing over a pixel in the overview (red, diagonal arrow) highlights its entire row and column, and also highlights the corresponding text in the base view. 23
4.2 Mapping from Levenshtein's string similarity metric to pixel color. 26
4.3 An example of vertical and horizontal changes to pixel color, as a result of hovering over a pixel in the Overview. Mousing over a pixel for a lemma 'C.' immediately highlights the pixels in its row and column. The base text view is coordinated with mouseover interactions in the Overview. The line in which 'C.' occurs is indicated by a yellow background and all occurrences of 'C.' are drawn in red. 27

4.4 Sequence of interactions involved in performing a pattern recognition task, *"What are the two common variants for words "fraxinea" and "nolit" in poem 1, page 3?* using Giarratano's 1910 critical edition of Calpurnius Siculus: (a) initial visualization state; (b) navigating to page 3, (highlighted in the Page view); (c) selecting line 39 then comparing the patterns of variants of words "fraxinea" and "nolit". By viewing the words locally, we can see that the two common variants are W10 and W16. 30
4.5 A fragment of the TEI – encoded file of the first poem in Giarratano's 1910 critical edition of *Calpurnius Siculus*. Tags such as "wit" and "source" attributes are used to encode witness names. The <lem> tag contains lemmata from the base text. The <rdg> tag is used to encode the text variant of a particular lemma. The tag <l> indicates the line number of a lemma. 33
4.6 The three relational tables — Lemma, Collation, and Variant — with primary and foreign key identifiers that support generation of and querying in the visualizations. 34
4.7 Flow of data in the pixel-based visualization. 35
4.8 Accuracy, confidence, and time taken results for tasks in the three task groups (I, P, N), with mean, error bars, and p-value. A star indicates the best performing task group. 40

5.1 TexTile - A Pixel-based Text Analysis tool with an integrated focus+context technique. The pixels are at the intersection of witnesses and lemmata. The views are aggregated at the levels of pages, lines, and lemmata. (A) The central Lemmata view with focus on a single lemma that can be dragged horizontally and vertically to pan across all four directions. (B) Line views on either side of the central view, with pixel color and variant count aggregated at the level of lines (C) Page views on either side of the central view, with pixel color and variant count aggregated at the level of pages. (D) Full text view, with a summary grand total of variant counts. (E) The list of witnesses that are considered for analysis. (F) user features that can be enabled on demand - text box to select a poem file; show grid; show line/page separators; show variant counts; and pixel height slider. (G) A multiselect witness list box, to select subsets of witnesses for analysis. 46

5.2 The figure depicts the modulus arithmetic for tiers on each side of the TexTile visualization. 'w' indicates the lemma in central focus. L(w) indicates the line number of lemma 'w'. P(L) represents the page number of line 'L'. #w, #L, and #P represent the fixed number of columns allocated for display in each of the tiered views (Lemmata, Lines, and Pages, respectively). These values are currently set as 13, 9, and 5 in TexTile, and were chosen based on typical lemmata per line and lines per page in our target Latin texts. 47
5.3 The five-level color scheme used to map Levenshtein edit distance between two strings into pixel color. 50
5.4 TexTile views at pixel height zoom levels of 3 (top) and 20 (bottom). 51
5.5 TexTile filtered to show a subset of witnesses selected in the multiselect Witness list box at bottom right. 52
5.6 Sequence of interactions involved in performing a pattern recognition task using TexTile. *"How many groups of colored variants do you see for lemma 'Thyrsis' on page 12, line 21?"* (using poem 2 of Giarratano's 1910 critical edition of Calpurnius Siculus). A) Initial visualization state. B) Enabling the page and line separators for guided navigation. C) Identifying the lemma in the central focus view and analyzing degree of similarity between variants based on pixel colors. Interaction and querying identifies three groups of variants. 53
5.7 Accuracy, confidence, and time taken for tasks in the three task groups (I, P, N). The horizontal red line in the bar graph indicates a result of importance that is also discussed in the main text. . . 58
5.8 Qualitative results of nine different domain-independent operations a user might need to execute to analyze data, following Wehrend et al. [58]. The green bar under each bar graph indicates the best performing task. 63
5.9 Comparison of the prototype and TexTile pixel-based text analysis tool with respect to the three performance metrics. The horizontal axis indicates the three task groups(I, P, and N). 64

Abstract

Before the advent of printed texts, text duplication was primarily done by hand copying. As a result, numerous versions of the same text with errors, alterations, and erasures were often created. Classics scholars synthesize these multiple versions of texts to create conjectured reconstructions of the original text. Many scholars continue to use spreadsheets, and sometimes very large sheets of paper, to visually collate variations across known versions. While suitable for collection of data about variations, these approaches are generally poorly suited for analysis of variation above the level of individual words. It is hard to discern patterns of textual variation at the level of lines, pages, or entire text directly, and it requires years of training and hands-on experience to master. Scholars need tools that allow them to apply their deep expertise and express idiosyncratic queries in exploration of patterns of textual variation across scales.

This thesis contributes a novel pixel-based focus+context visualization for analyzing the historical evolution of a text. It offers interactive means to examine and characterize patterns of variation across textual scales. The technique provides a compact representation of variation data sets, allowing scholars to validate the accuracy of textual variants and identify the level of sameness between different copies and reconstructed editions of text. An overview (the context) of the entire text displays the density and distribution of features across pages. A

detail view (the focus) lets scholars narrow down their analysis to points of interest, examine patterns, and assess similarities and differences. The integration of pixel-based representation and focus+context navigation allows users to explore variation across scales while preserving the continuity of experience, like one has when examining a physical manuscript. Interactive features include dynamic filtering on sources, brushing to locate variants in a traditional text layout, and details-on-demand of individual variations and their similarity measures. We conducted a usability evaluation to assess how well the technique supports common, desirable variation discovery and comparison tasks. The integrated representation and navigation style lets scholars see and compare sources above the level of words in a practical way for the first time, expanding the scope of critical analysis of essential historical texts.

Chapter 1

Introduction

Text is a natural form of storing and accessing information. One of the major challenges of analyzing text data is that it is inherently unstructured and fuzzy. The first step employed by text analysts in the process of gaining insights from texts is to transform the raw text data into an intermediate form, such as a semi-structured format like Extensible Markup Language (XML), or a structured form such as relational data. This transformation enables analysts to apply various automated statistical analysis techniques such as clustering, classification, association, and grouping of data. These automated techniques can help greatly in reducing the time needed to identify anomalies and interesting regions in text. However, text is highly context-dependent. It further requires thoughtful examination to confirm the validity of observations made.

Using visualization tools, text can be viewed and examined in context. Visualization provides visual summaries of entire texts and of relationships between portions of texts. Designing scalable, multiple view, and interactive visualization can help users perform an effective analysis of large-scale data like text. Building an effective visualization tool requires a solid understanding of the data and do-

main. A variety of text visualization techniques have been proposed in the past, often the result of collaboration between computer scientists and domain experts. Notable tools include Sequence Surveyor [8] in genomics, TreeJuxtaposer [39] in biology, HotelsViz [57] in historical geography, and Compus [20] in social history.

Among the various disciplines of humanities, classics is a field of study that involves a wide variety of text analysis activities. The topics covered in the study of classics include literature, art, and the history of languages such as Latin, Greek, and Hebrew. Of the various activities that classics scholars perform today, *philology* remains a central focus [60]. It is more commonly defined as "the study of literary texts and written records, the establishment of their authenticity and their original form, and the determination of their meaning" [4]. Before the introduction of printed texts, text duplication was primarily done through hand copying. As a result, numerous versions of the same text with errors, alterations, and erasures were created over time. Classics scholars synthesize these multiple versions of texts to create conjectured reconstructions of the original text.

Analysis of literary works, like other large scale data analysis, poses multiple challenges to scholars. Scholars who work with large ancient texts often start with open-ended questions, and narrow down their search to focus on interesting patterns and outliers. They analyze further to draw conclusions from these patterns. Other goals are to clearly see a relationship between different versions of text, identify and compare interesting patterns in the recordings of a particular group of scribes, and finally create an edition based on the scholar's knowledge base and interpretation. Some of the basic requirements of a visualization tool to help a scholar in their analysis are capabilities to: summarize large text data; highlight interesting features; interact with text at different scales such as pages, lines and words; filter and query data; and help formulate and verify hypotheses.

Towards this aim, we have developed a novel pixel-based hierarchical focus+context (f+c) visualization technique that represents the different versions of text in parallel and hierarchically. The focus+context design supports displaying multiple levels of text and at the same time provides multiple views of aggregated levels of text structure.

As a part of the Digital Latin Library project (DLL) [2], we are working with scholars of Latin the Department of Classics & Letters at the University of Oklahoma to aid them in building advanced visual analysis tools for analyzing ancient Latin texts. Classical Latin scholars work with literary works that have complex text structure. This work involves comparison of multiple versions of the same text. This work addresses a domain problem from classics and also contemplates recommendations in the visual analytics research development agenda setforth by Thomas and Cook [52]. The pixel-based text analysis tool is aimed at facilitating users in this work through viewing, and exploring data about textual variation.

The chapters of the thesis are organized as follows:

Chapter 2 provides background information on data visualization, visual analytic techniques widely used to explore multi-scale data, and coordination schemes to interact with multiple views.

Chapter 3 describes the domain problem that we are addressing, the terminology and tasks involved in Latin textual criticism, the requirements of a visual analysis tool for classical Latin texts, and the design choices that lead to our focus on pixel-based visualization.

Chapter 4 discusses the design specification and implementation of a multiple view pixel-based focus+context visualization tool for analysis of text variants, the data used for visualization, a user study to evaluate the usability of the multiple view design, and the results and conclusions of the study.

Chapter 5 describes a revised design that integrates focus+context in a single, continuously scrolling view, a user study to evaluate the usability of the revised design compared to the original design, and the results and user feedback from the study.

Chapter 6 concludes with a discussion of the features and factors that improved the overall usability of the pixel-based focus+context design, the benefits and drawbacks of the revised design, and the potential for generalizing the visualization technique to other text variation data sets.

Chapter 2

Background and Related Work

This chapter explains the terminology and provides background on the various techniques used in data visualization.

2.1 Visualization

Visualization is graphical representation of data to help users understand information and find insights with lesser cognitive load. In general, cognitive load refers to the amount of mental resource required to perform a task. The insights gained can lead them to make better decisions. The basic element of any visualization is a *view* and can be defined as an interface that allows users to interact with the data. These views are constructed around visual representations such as scatter plots, parallel coordinate plots, bar charts, line graphs, and isoline or choropleth maps. The choice of visual representation is determined based on data characteristics and attribute relationships. With large amounts of data being collected and stored, viewing attribute relationships using a single view is difficult. Coordinated multiple views (CMV) [46] and multiform [45] visualizations as ex-

plained by Roberts, help data to be presented and viewed using complementary visual forms. This multiple representation of data helps users to view and interpret data in different ways that can lead to effective exploration, analysis, and dissemination of data. To navigate and explore large data using multiple views, understanding the different schemes for coordinating multiple views in a user interface is essential.

2.2 Interface Schemes For CMV

Traditional mechanisms built around a single view use paging, scrolling and panning to navigate information spaces. But these features are inadequate and introduce discontinuity in the exploration when it comes to navigating between multiple views. There are a number of coordination schemes proposed in the past that help to overcome this issue and mentally assimilate the overall structure of the information space and navigate through it. Cockburn et al. [16] have reviewed and categorized interface designs to four major schemes based on the mechanism used to separate and blend views. All of these schemes allow users to rapidly move between detail and context views with varying degrees of continuity.

- **Overview+Detail - Spatial separation.** This method displays the entire data set in one view, namely the *overview*. The results of interaction/selection on this view is displayed in a separate view, the *detail view*. These two views are spatially separated. Actions performed in one view are immediately reflected in the other view. Some of the visualization tools that implement overview+detail include ValueBars [15] and PhotoMesa [30].

- **Zooming - Temporal separation.** This method helps users to traverse

between detailed and contextual views using zooming in (magnifying) or zooming out (demagnifying). Both views occupy the same screen space and different amounts of detail are displayed by zooming in and out. The different zoom levels are achieved through discrete (in a stepped manner) or continuous (fluid) transitions between different data scales using mouse or keyboard. Example visualization systems developed using this technique are Muse by Furnas et al. [23] and a semantic zooming tool by Summers et al. [51].

- **Focus+Context - Seamless focus within context.** This techniques enables the user to investigate specific details of the data while at the same time provides an overview of the entire dataset. This method is very similar to overview+detail, but minimizes the seam between views by placing the focus within contextual views. Distortion techniques, such as suggested by Furnas et al. [22] and Sarkar et al. [48] are employed to achieve a balance between local detail and global context. The basic idea of distortion techniques is to show portions of the data in focus with a high level of detail, while all other detail surrounding the focal area are shown at a lower level of detail.

  One of the major advantages of focus+context views is that, all views of different data scales are presented in a single coherent display. As a result the memory load associated with assimilating the different views present in the tool is highly reduced, as compared to the other three techniques. Example focus+context visualization systems include the Perspective Wall [35] to visualize document file systems, the Table Lens [44] for analyzing tabular information and TreeJuxtaposer [39] for supporting comparison across

hierarchical datasets.

- **Cue-based techniques.** This is a subset of focus+context techniques. Users can select information in terms of their interest, usually by assigning a certain visual cue to them. With this technique, information that satisfies certain search criteria is displayed in focus, while the remainder surrounding it are blurred. This technique allows visual emphasize of certain areas, that might require user attention. This can lead to faster navigation through a data set. Examples of cue-based approaches include semantic depth of field by Kosara et al. [32] and Baudisch et al.'s Fishnet [12].

2.3 Pixel-based Visualization

Pixel-based (pixel-oriented) visualization is a visual analytic technique popularized by Keim [27]. It efficiently uses screen space for visualizing large amounts of data. In this technique, each data point is mapped to a single pixel or a region of pixels. In general, the pixel shape is either a rectangle or a square. The pixel color is derived from a color map that represents the range of the data value being represented. These properties of the technique help to produce a compact representation of the entire data set. The pixel-based visualization technique is widely used in applications and research that involves large multivariate datasets.

Notable pixel-based visualizations include Literature Fingerprinting [28], which visualizes two English novels, "The Iron Heel" and "Children of the Frost", by Jack London. The authors used a pixel-based technique to perform authorship attribution analysis on the two novels. In this analysis, the authors mapped pixel color to vocabulary richness. The final pattern reveals the structure and difference in vocabulary richness used in the two novels. A second example is the Informa-

tion Mural [26], which visualizes the number of sunspots recorded daily from 1850 – 1993 using pixels. In this visualization, the color is mapped to data density, in order to understand the overall pattern in daily sunspots recorded. They were able to map a very large data set on the available screen space, without averaging. A third example is OnSet [47], which visualizes chemical compositions with binary data characteristics. Unlike other pixel-based visualizations, OnSet uses the presence or absence of pixels to represent an element present or absent in a chemical composition.

Some of the important design considerations of a pixel-based technique are how to map data points to pixel color; arrange pixels; group pixels; and support interaction. Based on one's requirements and analytic needs, these design choices can vary widely.

- **Pixel Color Mapping.** By mapping data attributes to colors and arranging them near to one another, patterns and trends in data becomes visible. Since hue has advantages over gray scale, identification and differentiation between ranges of data points can be made easier, as studied by Levkowitz et al. [34]. To study the impact of colors on usability of pixel-based visualization systems, Oelke et al. [42] used visual boosting techniques such as a halo effect and distortion of individual pixels. In text data, color is a visual attribute that is used to display the features of textual components and can be applied in many different ways. Such features include word count, the part of speech of words, vocabulary richness, and sound patterns in poems.

- **Pixel Arrangement.** Since this technique uses a pixel or a small region of pixels per data value, it allows us to visualize as large an amount of data as possible on the available display. If each data value is represented by one

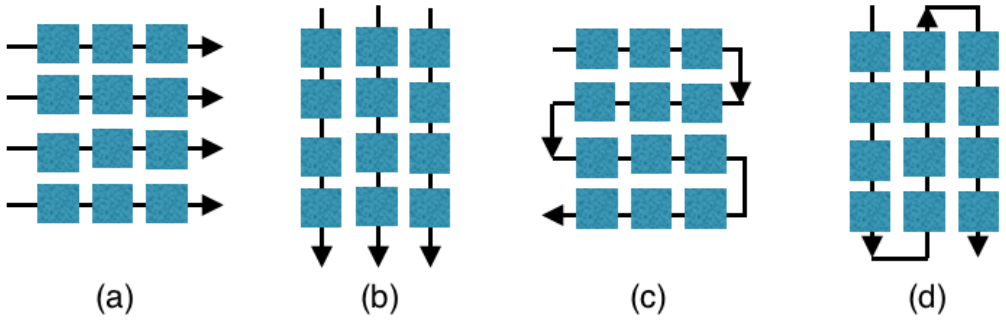


Figure 2.1: Different pixel arrangement techniques. (a) line-by-line (b) column-by-column (c) left-right (d) top-bottom

pixel, an important question is how to arrange pixels on the screen to help users compare data values. Keim et al. [29] suggest arranging data points line-by-line or column-by-column (as in Figure 2.1a and 2.1b, respectively) or in a recursive pattern. Recursive arrangement can be done left-right or top-bottom (as in Figure 2.1c and 2.1d, respectively). Pixels can also be arranged using two other methods: following the natural order of a data point's occurrence, or by using a query-dependent arrangement. In the former method, the natural occurrence of data is used to arrange the pixels from top to bottom or right to left. Whereas in the latter, the user can filter and visualize only the data that is relevant in the context of a specific query. This method aims to reduce users being overwhelmed by large amounts of data. Both methods help to preserve the distance between data objects.

- **Pixel Segmentation.** In pixel-based visualization, data pixels are usually drawn inside a rectangular window. Though this strategy allows for efficient screen usage, it also leads to dispersal of the pixels belonging to one category or group throughout the entire window. This results in difficulty

in detecting overall patterns, clusters, and correlations in the data. One solution to overcome this issue, is to draw segments radially, displaying one dimension per segment in the circle [27]. For this there needs to be an automated technique to order the dimensions in such a way that the correlation and pattern in the different dimensions is apparent.

2.4 Hierarchical Visualization

In many data sets, there exists a natural hierarchical structure. Examples include, hierarchy in text (book, chapters, pages, lines, words, characters), geography (country, state, county, area), and time (years, months, weeks, days, hours, minutes, seconds). Hierarchical visualization helps to integrate macro-scale observations of large data set and micro-scale observations of individual data points.

As described by Card et al. [13] and Heer et al. [24], the hierarchical structure in data can be expressed using *Connections* such as branching in trees, edges in graphs, and splits in dendrograms. These representations use adjacency as a visual property to express the hierarchical structure in data. *Enclosures* use nested views to depict parent-child relationship through containment rather than adjacency. Examples of visualizations that exploit enclosure are treemaps and nested circle packing layouts.

A disadvantage of trees is that they require considerable amounts of empty space to visually layout relationships in data as branchings. As the tree gets large, accessing nodes in a dense tree layout can also be an issue. Some of the techniques to overcome this include interactive distortion, overview+detail, and dynamic querying. In contrast, Enclosure nested views fill the space, and the entire area (or length) of the container is used to represent a quantitative

value. They offer better readability, require less complex interaction, and provide multiple navigation features for exploring data at different levels of granularity. Enclosure is particularly useful for identifying large values in a data set. The general structure of hierarchical visualization designs are depicted in Figure 2.2. Some examples of hierarchical visualization techniques those are described by Stasko et al. [50], Mackinlay et al. [35], Shneiderman (tree-maps) [49], and Fruchterman et al. [21].

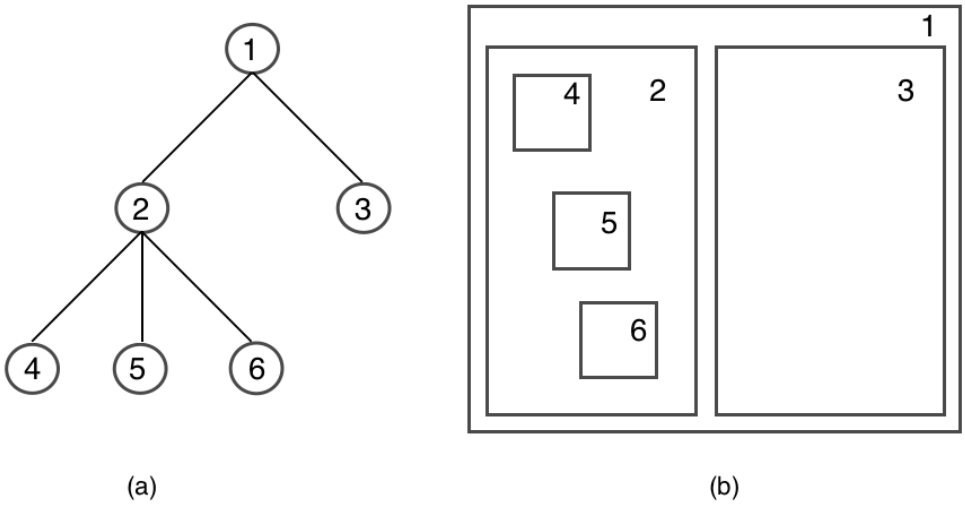


Figure 2.2: General layout of hierarchical visualization: (a) Connections in a tree; (b) Enclosure in a nested arrangement.

2.5 Text Visualization

For text analysis tasks, fully automated methods may not be efficient, since text is context dependent and a deep understanding of text is essential. The actual research happens on the text rather than the numbers. Visualization tools developed for text analysis purpose should provide direct access to the text from

any visualization. Many such text analysis tools exists, including those described by Wattenberg et al. [55], Collins et al. [17], and Vuillemot et al. [54].

In the case of text data with an innate hierarchy, one can apply effective aggregation of data characteristics at the different levels of text granularity. Exploring a text at multiple levels of hierarchy calls for a seamless navigation mechanism, one that allows users to move continuously between different levels with low effort. This is achieved in VarifocalReader [31] using a navigation mechanism called SmoothScroll [59], which is a combination of overview+detail and focus+context techniques. Visual abstraction of text, by providing an overview, can support users in gaining a general understanding of the information that a text conveys. However, working with abstracted information is often not enough to solve a given task, and may require access to finer levels of detail in the text hierarchy.

With our pixel-based focus+context technique, we aim at smoothing out the context switches that one has to make while analyzing single, large text documents. Our technique displays intermediate levels present in the text and provides interaction and navigation features navigate within and across different text scales. Like many data analysis approaches, visual text analysis can gain from a combination of multiple visual analysis techniques. We combine pixel-based display with focus+context representation and interaction.

Chapter 3

Textual Variant Analysis - Domain Problem

3.1 Sensemaking in Humanities Research

The process of humanities scholars, when analyzing text documents, is comparable to the sensemaking process described by Pirolli and Card [43]. The entire model of sensemaking consists of a set of subprocesses, that are performed iteratively as per the analytic needs.

1. The process starts with the collection of *External data sources*, such as raw text files, that are the source for the analysis performed.

2. The second stage is to populate a *Shoebox*, which is the process of selecting a subset of data, from the previous step, that is relevant to analysis. This is an iterative process which is performed as long as is required.

3. The third stage is to develop *Schemas*, which involves re-representation of the data in a convenient form that allows hypotheses formulation.

4. The next stage is *Hypotheses*, which involves the actual development of hypotheses from the information gathered in the previous processes.

5. Finally the process completes with *Presentation*, which is the showcase of the analysis results.

In proposing the sensemaking model, Piroli and Card suggest applying technology at various stages to improve the overall sensemaking process.

The exploratory nature of the tasks performed by a classics scholar is comparable to the above sensemaking process. Scholars choose a literature work to study, select a subset of editions relevant for their research, formulate hypotheses, draw conclusions regarding the hypotheses, and publish findings with enough evidence and a narrative to support their claim.

There are various stages in a scholar's research that involves tedious work such as scanning, assessing, and selecting texts for evaluating hypotheses. Developing sophisticated text analysis tools can bring improvements to these stages of scholarly research. In this chapter, we discuss one especial need in classics, namely support for a time consuming process called *Textual Criticism*. We discuss next how data analytics technology can help ease this complex process.

3.2 Vernaculars of Latin Textual Criticism

For Classics scholars, who perform critical analysis of ancient texts, there is a strong motivation to compare and contrast different versions of a single text. Over the years, the original manuscripts of numerous Latin works of literature have been either lost, or subjected to alterations due to scribal errors and erasures. What remains today are numerous copies of the original text. When comparing

these different versions of original text, the observed differences are called *variant readings*, or simply *variants* or *readings*. Understanding these variants is key to the research of scholars and historians. Variations in text can substantially alter the meaning and interpretation of these historically significant texts. One of the major tasks of classics scholars is to curate versions and reconstruct an entire text. The resulting text is a *critical edition* (or just *edition*), that closely approximates the original, with reference to supporting resources and other evidence. This process is known as *Textual Criticism*.

Digitization of texts, increased insistence on "born-digital" works in the scholarly community, and improved accessibility to works of significant importance to scholars have all helped facilitate the textual criticism process, which is otherwise tedious. Recently, increasing collaboration between humanists and computer scientists is leading to development of new methods, systems, and specialized tools for performing critical analysis of texts more efficiently.

A Critical Edition

Before discussing the various tasks a scholar needs to perform while critically analyzing a printed edition, we introduce important terms and ideas. The first step in the creation of a modern edition is manuscript *collation*. To collate a manuscript is to observe and record everything which may be of use towards determining what stood in the source [6]. Based on the collation, an editor issues various emendations to the text and publishes these as footnotes which constitute the *critical* part of *critical edition*.

An important reason for noting all known variants in critical editions is that editors do not want the readers to be unduly influenced by seeing just one variant. This might lead them to reading the reconstructed edition as if it were authori-

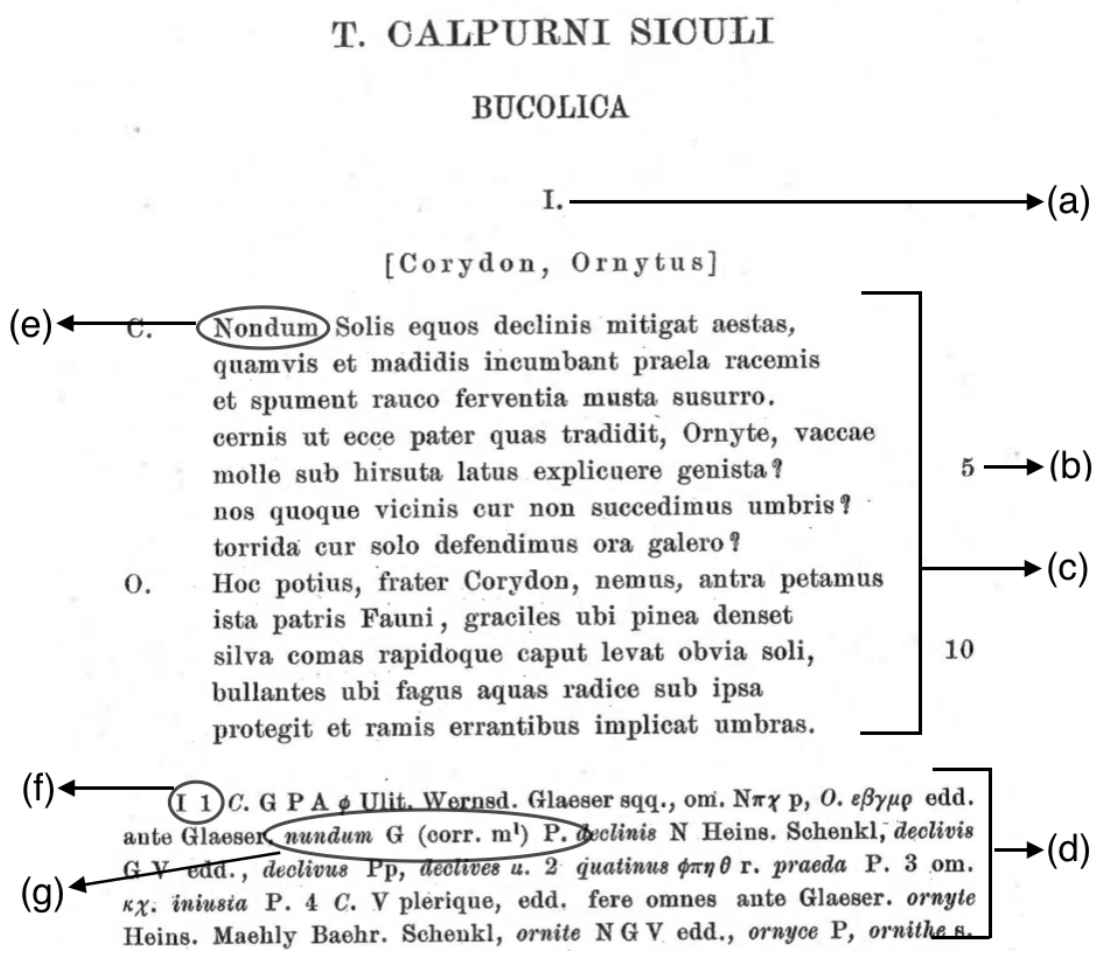


Figure 3.1: The first page of Giarratano's 1910 edition of *Calpurnius Siculus* [53], including: (a) poem numbers, (b) line numbers, (c) base text, and (d) critical apparatus. An apparatus entry (f) lists past variants (g) with the one chosen for use in the text, called the *lemma* (e).

tative original. Instead, by listing all important variants, readers are encouraged to engage in an interpretative reading.

Structure of a Printed Edition

The structure of a modern printed Latin edition comprises a *base text* and a *critical apparatus* in the form of footnotes (see Figure 3.1c and 3.1d, respectively). The critical apparatus is widely used by editors to discuss variants found in a

variety of sources and to present their justification for the choice of variants to include in the base text, based on editorial conjectures. A *lemma* (see Figure 3.1e), is the word or portion of the base text, to which a note in the apparatus refers. The plural of lemma is *lemmata*. A typical critical apparatus as shown in Figure 3.1d, consists of a set of lemmata that each describe a variant, the name of the hands/editions/witnesses the variant linked to, and the type of variant (for example see Figure 3.1g). Common types of variants include lexical, syntactic, and orthographic (spelling) corrections. Apparati follow well-established scholarly conventions but are dense, complex, and can also contain a variety of edition and editor specific idiosyncratic information that can be hard to decipher.

Common tasks that a scholar performs in creating a printed edition include:

- comparing numerous variant-witness combinations;
- finding interesting patterns in variants, recorded by witnesses;
- analyzing evolution of a variant from one version to another;
- making sense of the overall distribution of variants in an edition; and
- creating the edition by composing various emendations.

Another important activity that scholars perform is alternating between distant and close reading of texts. *Close reading* requires readers of a text to thoughtfully read and reread a text, and comprehend well to find interesting patterns. This method is mostly guided by experience. *Distant reading* [37] focuses on looking at texts as data. Although distant reading reduces the complexity and time-consuming nature of close reading, it also precludes the interpretation of texts through direct examination. Text analysis tools that provide easier transi-

tion between close reading and distant reading would increase the advantages of both of these important scholarly approaches.

3.3 Variant Analysis Tools

In Digital Humanities, some notable tools for comparing and visualizing variants are Juxta [3], Interactive TimeLine Viewer [36], CollateX [1], and TRAViz [25].

Juxta is a tool for comparing and collating multiple witnesses to a single textual work using variant density heat maps and histograms. Though this tool helps to recognize patterns and density of different metrics of the text, they lack features for querying a text at different levels of granularity such as pages, lines, and words.

The Interactive TimeLine Viewer uses simple graphical representation such as histograms, to display word lengths in different versions of text. Comparing histograms to identify similarity between versions is a hard and tedious endeavor. Moreover, it is not scalable with increasing number of versions, which can be 50 or more in modern critical editions.

CollateX and TRAViz are graph-based variant comparison tools. Both are scalable to a large number of versions, but this generally results in a very complex visual display. These tools provide very less access to critical apparatus data associated with the base text that also increases the complexity in comparing the editions.

Alan Galey's Visualizing Variation project [7] proposes a series of browser-based visualization prototypes for viewing critical editions and Variorums. The project aims primarily to aid users with close reading of texts and is less interested with distant reading techniques.

We collaborated with classics scholars at The University of Oklahoma's Classics & Letters department to understand the various processes involved in textual criticism, and the types of information that are present in critical apparatus of an edition. The critical apparatus of a primary text is typically recorded at the bottom of each page. As discussed earlier, this portion of the edition is highly detailed due to the large amount and wide variety of information present. This includes location (line number, page number, chapter) of a word/part of the primary text to which the annotated notes refers, important variants of the same word used in previous editions, and the witnesses (editors, scribes, emending hands) who annotated the variants.

A scholar's goal in analyzing textual variants are to clearly see a relationship between the primary text and variants, identify and compare interesting patterns in the recordings of particular scribes, and often also to create a new edition based on individual expertise and conjectures. Consequently, the major design objectives for our visualization tool are fivefold:

1. A compact display of variant distribution among witnesses over the length of the text.

2. A navigation mechanism to traverse levels of text structure.

3. Effective use of color encoding to facilitate pattern recognition and similarity comparison between versions.

4. Interactive controls that displays details-on-demand and query capabilities to view desired portions of data.

5. An aesthetically pleasing representation of the hierarchical organization of large texts.

Overall, the goal is to identify interesting patterns and examine their idiosyncrasies. With these requirements taken into consideration, the following chapters describe an initial design of a variant analysis tool and the results of a user study conducted to assess this prototype and identify improvements. This is followed by a more refined version that implements features to overcome the shortcomings of its predecessor, then a final evaluation to assess the usability of the result.

Chapter 4

Pixel-based Hierarchical Focus+Context Visualization

To perform the tasks discussed in the previous chapter, we designed a pixel-based text analysis tool, as shown in Figure 4.1. The tool consists of three main views: a Base Text view, an Overview and a Detail view. The detail view has three sub views, that display the text at three different levels of aggregation, namely pages, lines, and words. This chapter describes the design features and tool implementation. Towards the end, we describe a user study conducted to evaluate the tool and analyze the results.

4.1 Design Specification

The Base Text view displays a Latin text in its conventional form. The Overview displays the distribution of lemma sequentially, maintaining the reading order of the text. The three hierarchical views - Page, Line and Word views display presence and count of lemma aggregated at their respective text scale. The

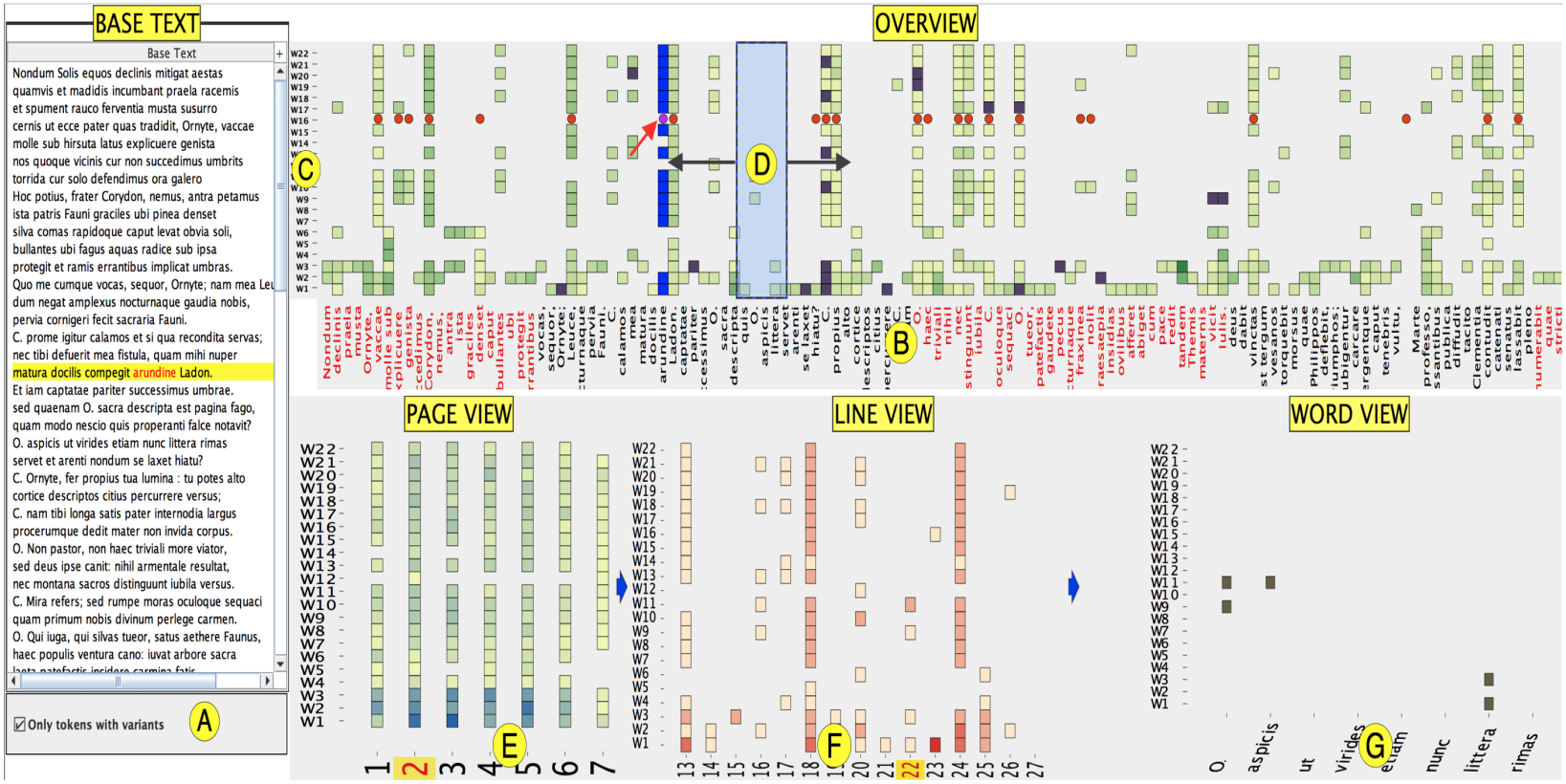


Figure 4.1: The initial prototype tool for analyzing variation in classical Latin texts. The layout consists of: a base text view (indicated by label "BASE TEXT") with a global filter option (A); an "OVERVIEW" of variation between lemmata (B) and witnesses (C); a draggable window to select a range of tokens (D); and a hierarchy of drill-down views that summarize the variation count at the level of pages (E), lines (F), and words (G) called "PAGE VIEW", "LINE VIEW" and "WORD VIEW" respectively. Mousing over a pixel in the overview (red, diagonal arrow) highlights its entire row and column, and also highlights the corresponding text in the base view.

following sections explain the intricacies of each of these views.

4.1.1 Overview

In the Overview, the horizontal axis displays lemmata in the same order of occurrence as in the text. The font color of tokens in the horizontal axis alternates between black and red to indicate the start and end of a new page. During analysis, this feature is helpful to keep track of the location of a token that is currently of interest. The vertical axis displays the list of witnesses (W1 to W22) that are present in the critical apparatus. In the central area of the graphical representation, at the intersection of the tokens and witnesses, are the variant readings, which are indicated by the presence of a colored pixel. Looking at the set of pixels vertically for a lemma, one can find the witnesses for which variants are present. Looking at the set of pixels horizontally for a witness, one can find the lemmata for which a variant is present.

Pixel Color Mapping

Pixel color mapping is used to encode the similarity between a lemma and its variants. In the Overview, the color of a pixel is mapped from a similarity measurement between the token in the base text and its corresponding variant reading for each witness. Some of the common measure to calculate similarity between two strings are Levenshtein distance, Hamming distance, episode distance, and longest common sequence distance. The definition for each of these metrics is provided in Table 4.1. In our case, the similarity is derived using the Levenshtein edit distance formula [33]. This distance metric considers three operations, namely insertions, deletions, and substitutions required to convert one string into the

Metric name	Definition
Levenshtein distance[33]	The minimal number of insertions, deletions and substitutions to make two strings equal.
Hamming distance[38]	The minimum number of substitutions required to change one string into the other.
episode distance[18]	The minimum number of insertions required to change one string into the other.
longest common sequence distance[40]	The length of the longest pairing of characters that can be made between both strings, such that the pairings respect the order of the letters. Only insertion and deletion of characters are allowed.

Table 4.1: Commonly used string distance metrics

other string. This measure is generally suitable to compare two strings of arbitrary lengths, and that is the scenario in our case. For instance, as shown in Figure 3.1, *Nondum* ("not yet") in base text has a variant reading *nundum* (a scribal error). The edit distance is calculated as two, since there are two substitution operations required to transform one string to another: the first character 'N' has to be substituted by 'n' and the second character 'o' has to be substituted by 'u'.

To map the edit distance to pixel color, we normalize distance value over the entire text. Then the range of distances is split and mapped to a five tier univariate color scheme, as shown in Figure 4.2. Bright pixels indicate high lexicographic similarities (smaller edit distances). Dark colored pixels represent higher edit distances, and indicates low lexicographic similarity with the lemma. To represent an omission variant type, we draw the corresponding pixel in purple (variants that are excluded from a particular edition are called *omitted* or *deleted* variants).

When a column has many pixels of the same color, it is an indication that many witnesses may agree on a particular variant. Conversely, a column of pixels

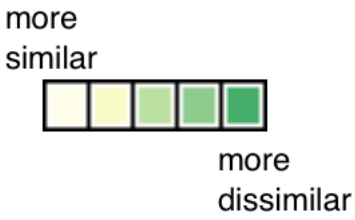


Figure 4.2: Mapping from Levenshtein's string similarity metric to pixel color.

with varying colors indicate the presence of multiple variants in different copies of the work, suggesting that the text may warrant scholarly attention. In both cases the distribution of colors conveys uncertainty about the editor's chosen lemma relative to the variants used in earlier versions of the text.

Interaction

The Overview has a set of interaction features that aid users in their analysis. Mousing over a pixel in the Overview (as indicated by the red arrow in Figure 4.1) immediately highlights the corresponding word in the base text by changing the font color from black to red. The line in which the word occurs is highlighted by filling it in yellow. All other occurrences of the word in the text are also drawn in red, but without the yellow background highlight.

Mousing over a pixel also highlights its entire row and column in the Overview (see Figure 4.3). Pixels in the same row are colored red. This feature can be useful when a scholar needs a list of lemmata for which a particular witness has provided a variant. Pixels in the same column are colored blue. This feature can be helpful when a scholar needs to check which witnesses provide a variant for a particular lemma of interest.

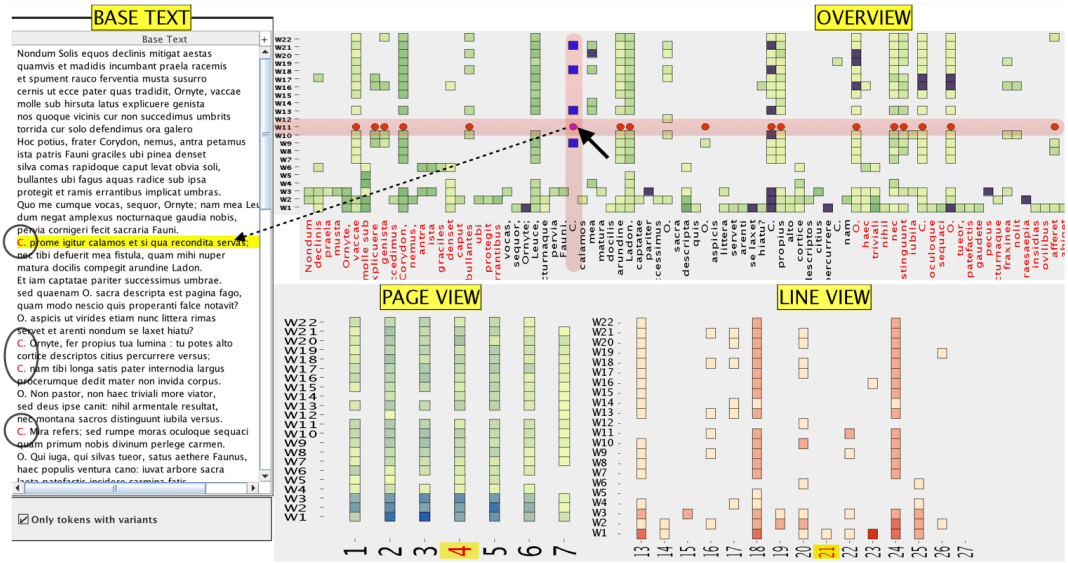


Figure 4.3: An example of vertical and horizontal changes to pixel color, as a result of hovering over a pixel in the Overview. Mousing over a pixel for a lemma 'C.' immediately highlights the pixels in its row and column. The base text view is coordinated with mouseover interactions in the Overview. The line in which 'C.' occurs is indicated by a yellow background and all occurrences of 'C.' are drawn in red.

4.1.2 Hierarchical Views

The detail view consists of subviews that summarize levels of text at different scales. The three Hierarchical Views — Page, Line, and Word — help users to focus more on tokens being analyzed in the Overview. This feature allows users to focus on particular scale in text. A draggable box in the Overview connects the Overview with these three hierarchical views. Using this box, the user can select a range of tokens in the Overview. The subviews are filtered to indicate or show only the pages, lines, and words for that range.

In the Page view, the horizontal axis has the page numbers listed from the entire text. The vertical axis displays the list of witnesses that occur in the critical apparatus. The Line view's horizontal axis displays the line numbers

filtered based on the page selection (see Figure 4.1e and 4.1f). The vertical axis has the same witness list as in the page view. The only difference from the Overview is in the metric used to encode the pixel color in these three views. The total number of variants present on each page and line level is mapped to the color scheme in the three detail views. The count of variants is normalized before mapping to the color scheme. Finally, the Word view is used to focus on a particular lemma locally, surrounded only by lemmata that occur on the same line.

4.1.3 Interaction and Query Capabilities

The tool includes interactive querying features. The representation in Figure 4.1, allows basic examination of lemmata and variants across multiple scales of text. A closer examination is often needed to characterize patterns and find outliers. Interaction lets scholars navigate over text and focus closely on particular parts of the text. Many words with no variant reading results in a sparse distribution of pixels. A denser representation of data, can help to reveal patterns and similarities present in data. The tool provides a check box (see Figure 4.1a) to filter and display only tokens with at least one variant reading in the Overview.

To explain the interactions and querying that a scholar might want to perform on the data using the pixel-based visualization, an example task is provided in Figure 4.4. Let us consider a scholar analyzing the text with the help of our visualization to answer a question: *What are the two common variants for words "fraxinea" and "nolit"? on page 3, line 39 of the poem.* This is an example that illustrates a task to characterize the relationship among variants in two lemmata. The sequence of interactions involved are:

- Navigate to page 3, by moving the draggable window in the Overview to cover page 3. The lemma color in the horizontal axis of the Overview and the page number shown in the Page view act as guidance for this navigation.

- In the Page view, selected page numbers are highlighted in yellow in the horizontal axis. Upon clicking page number 3, the data is filtered to display the lines on that page in line view (see Figure 4.4b). Lines 28 to 45 are filtered and displayed.

- In the Line view, clicking on line 39 displays the words on this line in the Word view (see Figure 4.4c). Now a scholar can compare the variants present for lemmata “fraxinea” and “nolit”.

- By viewing the words locally, they can identify that the two common variants are W10 and W16, which stands for sources δ and λ in the set of witnesses, respectively.

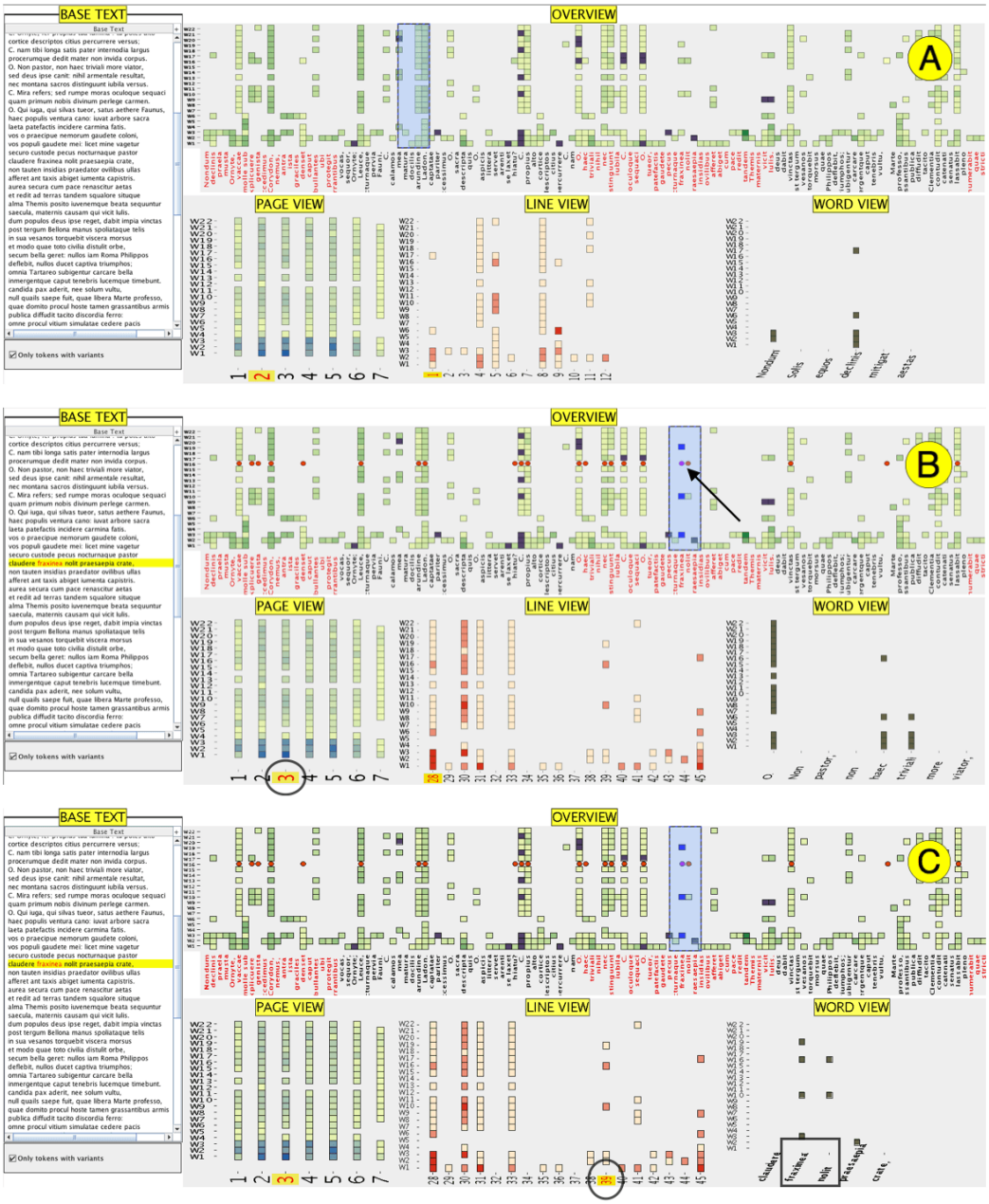


Figure 4.4: Sequence of interactions involved in performing a pattern recognition task, *"What are the two common variants for words "fraxinea" and "nolit" in poem 1, page 3?* using Giarratano's 1910 critical edition of Calpurnius Siculus: (a) initial visualization state; (b) navigating to page 3, (highlighted in the Page view); (c) selecting line 39 then comparing the patterns of variants of words "fraxinea" and "nolit". By viewing the words locally, we can see that the two common variants are W10 and W16.

On the other hand, let us consider a scenario in which the query is to identify the degree of similarity between two variants of the word "fraxinea" for witnesses W10 and W16. In this case one can navigate to page 3 in the Overview and look for the word "fraxinea" (see Figure 4.4a). It can be observed that the variant used in W10 has a brighter pixel color than the one used in W16. This indicates that W10 has a higher degree of similarity with the lemma than W16.

Below is a general list of tasks that a scholar can perform using this tool. By examining pixel placement and coloring, patterns and outliers in the variant-witness combination can be identified and characterized.

- **Examine the pattern of variants of different words.** Hovering over a pixel (Figure 4.1, red arrow) highlights all the pixels in the row in red to mark the words for a particular witness.

- **Examine the pattern of variants over different witnesses.** Hovering over a pixel also highlights all the pixels in the same column in dark blue to mark a lemma's witnesses.

- **Brush a word to see it in context.** Clicking a pixel highlights the corresponding word in the Base Text view. This exploits the familiarity of the layout in the Base Text view, helping the user navigate the less familiar arrangement of text in the pixel views (see Figure 4.3).

- **Drill down using the filter lens to an area of interest.** Filtering in the Overview highlights numbers and words in the hierarchical views. This feature allows the user to view a lemma following two approaches. First, by using the Overview, in which the lemma and its variants are surrounded by other variant occurrences. This approach is preferable when looking

for patterns in the entire text. The second approach looks at a lemma as part of the aggregated sets of lemmata in its parent line and page, which is preferable when exploring line and page level statistics.

4.2 Data Set

Our analysis of the tool uses Giarratano's 1910 critical edition of *Calpurnius Siculus* [53]. The base text and critical apparatus are stored in a conventional edition format. Figure 3.1 shows the first page of the edition. Latin scholars familiar with the text, converted the printed edition and its accompanying critical apparatus into XML following Text Encoding Initiative (TEI) conventions [5]. Figure 4.5 shows a sample. The format contains information about variant sources, where the variants appear in a text, the variants themselves, and general editor's comments on how the text was reconstructed from the available variants. This encoding process helps provide an organization to the data. The data until this point is either unstructured (as in conventional edition format) or semi-structured (as in the TEI encoding).

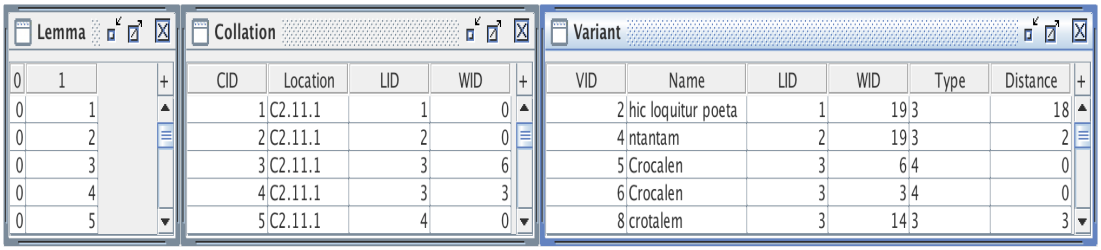
The scholars generally use a simple text editor for encoding critical editions. This is a manual process and undergoes multiple iterations. The reconstruction of the first poem in Giarratano's *Calpurnius Siculus* involves approximately 1200 variants across 7 pages, with an average of 6 lemmata for each of the 94 lines. There are totally seven poems in total. Once the scholars finished the encoding, they shared the XML with us. We used a set of XPath queries to convert the XML into tabular data for visualization purposes. XPath queries in general, can be defined as path expressions to select nodes or sets in an XML file. For example, we call a populateVariantTable() function that uses XPath queries to

```
<body>
    <sp who="#Corydon">
        <speaker>
            <app>
                <lem wit="#G #P #A #Φ" source="#Ulit. #Wernsd. #GlaeserSqq">C.</lem>
                <rdg wit="#N #π #χ" source="#p"/>
                <rdg wit="#ε #β #γ #μ #ρ" source="#eddAnteGlaeser"> O.</rdg>
            </app>
        </speaker>
        <l n="1">
            <app>
                <lem xml:id="lem1.1nondum">Nondum</lem>
                <rdgGrp type="anteCorr">
                    <rdg wit="#G1" varSeq="1" xml:id="rdg1.1nundum">nundum</rdg>
                    <rdg wit="#G1" varSeq="2" copyOf="#lem1.1nondum"/>
                </rdgGrp>
                <rdg wit="#P" copyOf="#rdg1.1nundum"/>
            </app>
            <app>
                <lem wit="#N" source="#Heins. #Schenkl">declinis</lem>
                <rdg wit="#G #V" source="#edd.">declivis</rdg>
                <rdg wit="#P" source="#p">declivus</rdg>
                <rdg wit="#μ">declives</rdg>
            </app> equos declinis mitigat aestas,</l>
        <l n="2">
            <app>
                <lem>quamvis</lem>
                <rdg wit="#φ #π #η #θ #r">quatinus</rdg>
            </app> et madidis incumbant <app>
                <lem>praela</lem>
                <rdg wit="#P">praeda</rdg>
            </app> racemis</l>
```

Figure 4.5: A fragment of the TEI – encoded file of the first poem in Giarratano's 1910 critical edition of *Calpurnius Siculus*. Tags such as "wit" and "source" attributes are used to encode witness names. The <lem> tag contains lemmata from the base text. The <rdg> tag is used to encode the text variant of a particular lemma. The tag <l> indicates the line number of a lemma.

extract information from <rdg> tags in the XML and populates the variant table with it.

The first step in this process is to define the schemas of the tables. The schemas are designed to support the specific visual queries in our visualization tools. The next step is to populate the tables with data by mapping XML information into table columns. The three tables generated using XPath queries represent lemma, variant, and collation information. As shown in Figure 4.6, the collation table provides details about the location (page and line numbers) of



Lemma

0	1
0	1
0	2
0	3
0	4
0	5

Collation

CID	Location	LID	WID
1	C2.11.1	1	0
2	C2.11.1	2	0
3	C2.11.1	3	6
4	C2.11.1	3	3
5	C2.11.1	4	0

Variant

VID	Name	LID	WID	Type	Distance
2	hic loquitur poeta	1	19	3	18
4	ntantam	2	19	3	2
5	Crocalen	3	6	4	0
6	Crocalen	3	3	4	0
8	crotalem	3	14	3	3

Figure 4.6: The three relational tables — Lemma, Collation, and Variant — with primary and foreign key identifiers that support generation of and querying in the visualizations.

each lemma. Several identifiers act as primary and foreign keys (LID for Lemma ID, WID for Witness ID, VID for Variant ID) to uniquely identify each record and also represent attribute relationships across tables. The key relationships support the multiscale queries used to perform visual queries and populate the visualization views.

4.3 Prototype Tool Implementation

The prototype tool is implemented in Improvise [56], a visualization environment for designing and implementing highly interactive visualizations. The tools is an Improvise document (a .viz file) that scholars can download from the project website. The data flow is shown in Figure 4.7. A set of XPath queries are applied to the encoded TEI XML to extract the values of specific tags and attribute values. The extracted values are stored in three relational tables as discussed in the previous section. Once the data tables are loaded in the Improvise system, they are visually mapped through instructions for drawing each data item (*visually encoded*). The pixel views are generated and rendered on the screen. One can then interact with the views to navigate and query the data. While querying, two types of filters can be applied. When a *data filter* is applied, the original

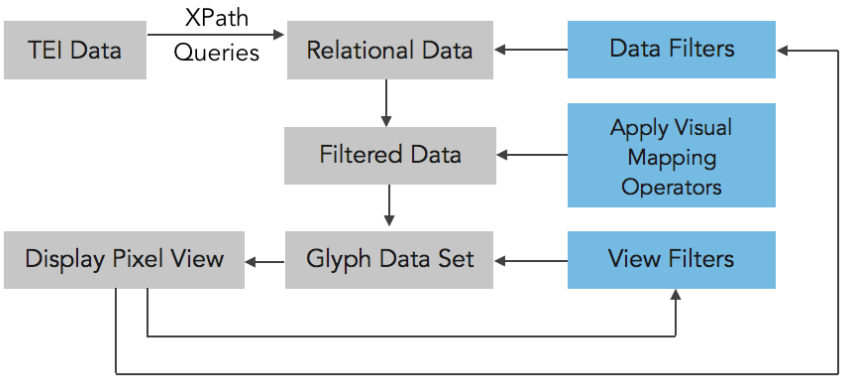


Figure 4.7: Flow of data in the pixel-based visualization.

data is queried and results in a subset of data that satisfies a specific condition expressed by the filter. When a *view filter* is applied it dynamically adjusts the visual encoding such as for pixel color changes and label highlighting.

4.4 Evaluation

To gain understanding of the usability of the prototype tool, we studied and analyzed the differences in task-based user performance. We evaluated the prototype tool as it was designed to reveal patterns of variation across *witnesses*—manuscript copies, earlier editions, and other sources—for a particular *lemmata*—words or segments—in a text.

The pixel-based representation allows focusing on text variation at the level of an individual lemma, or an entire line or page. Apart from assessing the overall usability of the tool, the three major goals for this user study were to verify:

1. the effectiveness of representing variation count in multiple pixel views, aggregated at the levels of lines and pages;

2. the effectiveness of pixel coloring in identifying patterns; and

3. the ease of use in performing queries, using focus+context, across multiple pixel views.

We chose three performance metrics to assess user performance: (1) accuracy in answers, (2) confidence level in answers, and (3) time taken to perform a task.

We calculate these metrics for three different task groups as explained in the following section. The scenarios are designed to reveal, how well people perform groups of tasks, and thereby inform our understanding of the usability and learnability of the visualization tool overall.

4.4.1 User Study

Following the guidelines described by Carpendale [14], we designed a user study to assess the effectiveness and usability of our text variant analysis tool. To define representative user tasks, we interviewed scholars to identify typical questions that they raise when examining printed critical editions by hand. To formulate an experimental hypothesis, we identified three major visual analysis tasks, each associated with one of our three user study goals.

We grouped these tasks into *synoptic tasks* as explained by Andrienko et al. [9] and Etemadpour et al. [19]. The three task groups are the following:

1. The **interaction-related (I)** tasks were for investigating ease in using selection, panning, and zooming features to perform a query on the data.

2. The **perception-related (P)** tasks were for investigating the effectiveness of pixel color mapping for performing grouping and pattern identification related queries on a given data set.

3. The **navigation-related (N)** tasks were aimed at verifying the usability

of the multiple coordinated view visualization features, particularly the distributed focus+context organization.

The complete questionnaire used in this user study is provided in Appendix A. Our null hypothesis is that *"the pixel-based text analysis tool performs equally well in all three task groups."* The alternate hypothesis is the opposite. By performing a comparison between task groups, we intend to identify the least performing task group and analyze further to understand the reason for the lower performance. On the other hand, the best performing groups will be an indication of features that are useful for analysis.

4.4.2 Experimental Settings

We conducted a pilot study with the participation of four graduate students and the classics faculty member on the DLL project. Together with assessment of preliminary results, we aimed to test the duration of tasks, the amount of training required to perform various tasks for a particular user group, and the robustness of the study user interface. The pilot study suggested some interesting trends in support of the alternate hypothesis. The actual study was conducted over several days with 14 participants. All were undergraduate or graduate students with normal or corrected to normal vision. Sessions were individual, included a brief training phase, and engaged five questions in each of the three task groups. Participants used typical interaction techniques with mouse, keyboard, and a 21.5" display. A think-aloud protocol was used to capture the approach that participants used to perform a query, and general observations were transcribed. Once a participant indicated understanding of the task and began interaction, time taken to answer was recorded. Since recording time starts only after the

Task group	Representative Tasks
Interaction	In page 2, click on any pixel for word "Fauni". From the base text view, find out how many occurrences (count) of "Fauni" is present in this poem?
	In page 7, click on the pixel for W3 and word "aures". What are the other witnesses that are highlighted in blue?
	On line view, line 45, what are the two words displayed on the word view that have a similar variant reading pattern?
Navigation	Which two words on page 3 have omissions for the same word by W16 and W17?
	On page 4, click on the pixel that represents variant (W2) for word "carcare". Which is the first and last word on page 4 that has a variant reading by W2?
	On line view for W1, between lines 13 to 27, which line has more number of variants?
Perception	Which page out of pages 1, 2 and 3 has the most number of omissions?
	On line 28, how many different variants are present?
	In page 1, which word has a varying pixel color pattern on its Y-axis?

Table 4.2: Representative tasks used to check usability of the prototype tool, grouped into three kinds of synoptic activity (Interaction, Navigation and Perception).

think-aloud of each task, it does not effect the timing of user behavior in performing a task. We collected qualitative feedback about the design and usability of the tool at the end of each session. Some of the representative tasks used in the study are provided in Table 4.2.

4.4.3 Study Metrics

This study uses three metrics—accuracy, confidence in answer, and time taken to perform a task—to assess the usability of the pixel-based text analysis tool.

Accuracy in answers explains the degree of correctness achieved in perform-

ing a task. A successful task is the one that is performed correctly and completely. Correctness is assessed based on ground truth. For each task we established ground truth answers at different levels of accuracy and precision. Each answer carried a specific weight and this enabled us to calculate the degree of correctness in answers provided by participants. This metric helps us to assess the overall level of correctness in observations that can be made using this visualization tool.

Confidence in answers, explains how sure participants are with the answers they provided for a task. After each task, participants were asked to mention their confidence in the given answer on a five-step Likert scale. This helps us to perform correlation between the responses given for a task and the corresponding confidence levels. This also helps us to identify tasks that users stated were difficult or complicated to perform. This again helps us to assess overall usability and identify specific features that require improvement.

Time taken is how long it takes a user to complete a visualization query task. To perform a task, users need to know which view to use and set of interactions to perform to achieve a desired result. If a user has insufficient understanding of a tool's functionality, they spend more time to understand the tool, particularly to make sense of their actions and the corresponding results. The time taken to perform each task was recorded (in minutes). This measure helps us to identify tasks that are time-consuming, look for reasons why, and consider how to improving the features that cause this behavior.

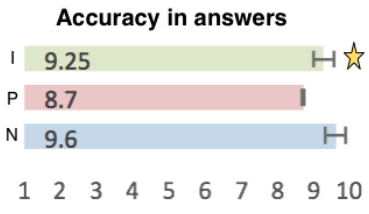
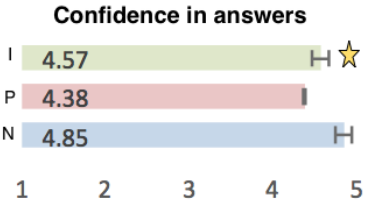
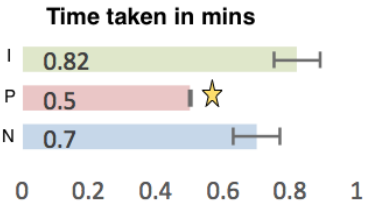
Means with Standard error bars	Friedman test results	Post-hoc test results
	$\chi^2 = 390.31$ $p << 0.01$	Tasks I and P has a significant difference with $p << 0.05$. Perception related tasks have less accurate results.
	$\chi^2 = 326.35$ $p << 0.01$	Tasks I and P has a significant difference with $p << 0.01$. Perception related tasks display lower confidence.
	$\chi^2 = 410.24$ $p << 0.01$	Tasks N and P has a significant difference with $p \leq 0.01$. Navigation related tasks take more time to accomplish.

Figure 4.8: Accuracy, confidence, and time taken results for tasks in the three task groups (I, P, N), with mean, error bars, and p-value. A star indicates the best performing task group.

4.4.4 Study Results and Discussion

In figure 4.8, we summarize the performance measures and analysis results for the I, P, and N task groups. We used three performance metrics: **accuracy** (in %), specifically the correctness of answers based on ground truth, normalized; **confidence level** in answers, on an increasing scale from 1 to 5; and **time taken** (in minutes) to perform a task. A check for data normality using the Shapiro-Wilk test revealed that all three measurements had a non-normal distribution. To further check the statistical significance of data, we applied a non-parametric

approach using the Friedman rank sum test to compare group means. All three metrics had $p \ll 0.01$, which indicates a significant difference in group means.

A post-hoc analysis using a pairwise Wilcoxon signed rank test helped us to identify the best/least performing task group for each metric. With respect to accuracy and confidence level, I and P tasks had a significant difference in means (accuracy: $Z = -1.35b$, $p \ll 0.05$, confidence: $Z = -3.82$, $p \ll 0.01$). Overall the study indicates that there is a significant difference in performance between task groups and hence we reject our null hypothesis.

We look further to identify the least and best performing task groups. Task I results were the most accurate and involve the highest confidence in answers. Task P results displayed less accuracy and confidence in answers. The primary reason behind this is the effort involved in identifying patterns in colored pixels and the rotated orientation of text along the horizontal axis of the overview.

To understand further the lower performance in P tasks, we verified the qualitative feedback elicited from participants received at the end of each session. Eight out of 14 users mentioned that, although they were able to identify patterns in colors easily, they were not entirely certain of the correctness of answers they provided. The orientation of text is a common design trade-off in visualization tools, and remains an open problem. When we consider time taken as a measure of efficiency, N and P tasks both displayed significant difference in means ($Z = -3.09$, $p \leq 0.01$). P tasks were less time consuming; patterns in color can be readily identified due to the compact representation of data points. N tasks needed more time to accomplish. They require traversal of multiple different views and a variety of drill-down steps.

4.5 Summary

There were several notable highlights observed from user task performance and qualitative user feedback. Participants found interaction, especially panning and zooming in the Overview, particularly useful when there is a close-call in differentiating pixel colors. Brushing a pixel to highlight the corresponding pixels in the same column (lemma-wise) and row (witness-wise) proved helpful for interpreting the distributions of variations. We relate the study results with the main goals listed in Section 4.4 to identify features that need improvement.

- **Displaying variation in text using multiple pixel-based views.** From the results of task group N, we observe that multiple discrete pixel views used to display texts at different scales is time consuming. This suggests a navigation mechanism that requires fewer steps to traverse different text scales. The results of task group I, which includes applying filter, panning, and zooming within views, indicate good performance with respect to all three measures.

- **Mapping string edit distance to pixel color.** From the results of task group P, we observe that differentiating similarity between variants is less time consuming, but displays a lower confidence level in answers provided. This is an indication that users face difficulty in readily identifying differences in colors under the univariate color scheme. An analogous color scheme with high contrasting colors, for instance the example provided by O'Donovan et al. [41], can help us differentiate the degrees of similarity. Analogous colors are any three pairs of side-by-side colors chosen from a color wheel, such as yellow-green, yellow-orange, and orange-red.

- **Performing queries using focus+context across multiple views.** The results of task group I represent querying the data set using lemma filter and drill-down in the multiple detail views. Task group I has good performance results with respect to all three metrics. We observed that users preferred to use scrolling, zooming, and panning of the detail views themselves while performing queries on the data set.

Based on the user study results and feedback, the following chapter describes an improved version of the pixel-based variant analysis tool. The tool has an integrated, multi-tier, focus+context design inspired by the Perspective Wall [35].

The work discussed in this chapter was presented as a poster at the IEEE Conference on Information Visualization 2015 [10] and published as a short paper at the Eurographics/IEEE Conference on Visualization Workshop on Visual Analytics 2016 [11].

Chapter 5

TexTile

From the insights gained in our initial user study, conducted to evaluate the prototype version of the pixel-based variant analysis tool, we identified a set of improvements. We developed a new version of the tool, called TexTile. TexTile differs from its predecessor in two key ways.

First, the improved version implements a layered hierarchy of views to provide an integrated focus+context display. A similar approach was used by Mackinlay, et al. in their Perspective Wall [35] to visualize a collection of documents in a filesystem. The wall has a view in the center that displays the details of a filesystem, such as the document files in a folder. On either side of the central view are perspective walls that display context to the left and right, shrinking into the distance. These walls display files aggregated in their corresponding folders. The files are structured by modification date along the horizontal axis and by file type along the vertical axis.

Consider an example of two folders in the file system hierarchy that were modified on the same date. Scrolling left, the left perspective wall expands these folders and displays their files in the central view. One can readily see a pattern

in the file modification date, and compare the two folders based on their files. Here the comparison of files happens in "visual parallel", along the horizontal axis. When a file is dragged to the right perspective wall, it is displayed as an aggregated part of folders progressively higher in the filesystem. The transition among views can be compared to a long sheet of paper brought into focus, one section at a time, with a perspective effect that fades as files move farther from the center of focus. In TexTile, aggregation at successive levels of lines and pages are displayed in tiered side views with individual lemmata in the central focus view.

Second, the pixel color is mapped to a uniformly applied, higher contrast yellow-orange-red color scheme. By implementing these two modifications in the design of TexTile, we aim to overcome the navigation and perception difficulties that we observed in the previous version of the tool.

5.1 Design Specification

A screenshot of TexTile is shown in Figure 5.1. Its design integrates two fundamental visualization techniques: pixel-based visual representation and stepped focus+context interaction. Together, the techniques support continuous multi-scale navigation over text. Overall, navigation is similar to SmoothScroll [59].

Navigation happens in five tiered views arranged in a shallow hierarchy horizontally: a central Lemma view that acts as a focus, two Line views on either side, and two Page views on the ends. These surrounding views collectively provide context around the lemma in focus. The following sections explain the design specification of the tiered view hierarchy.

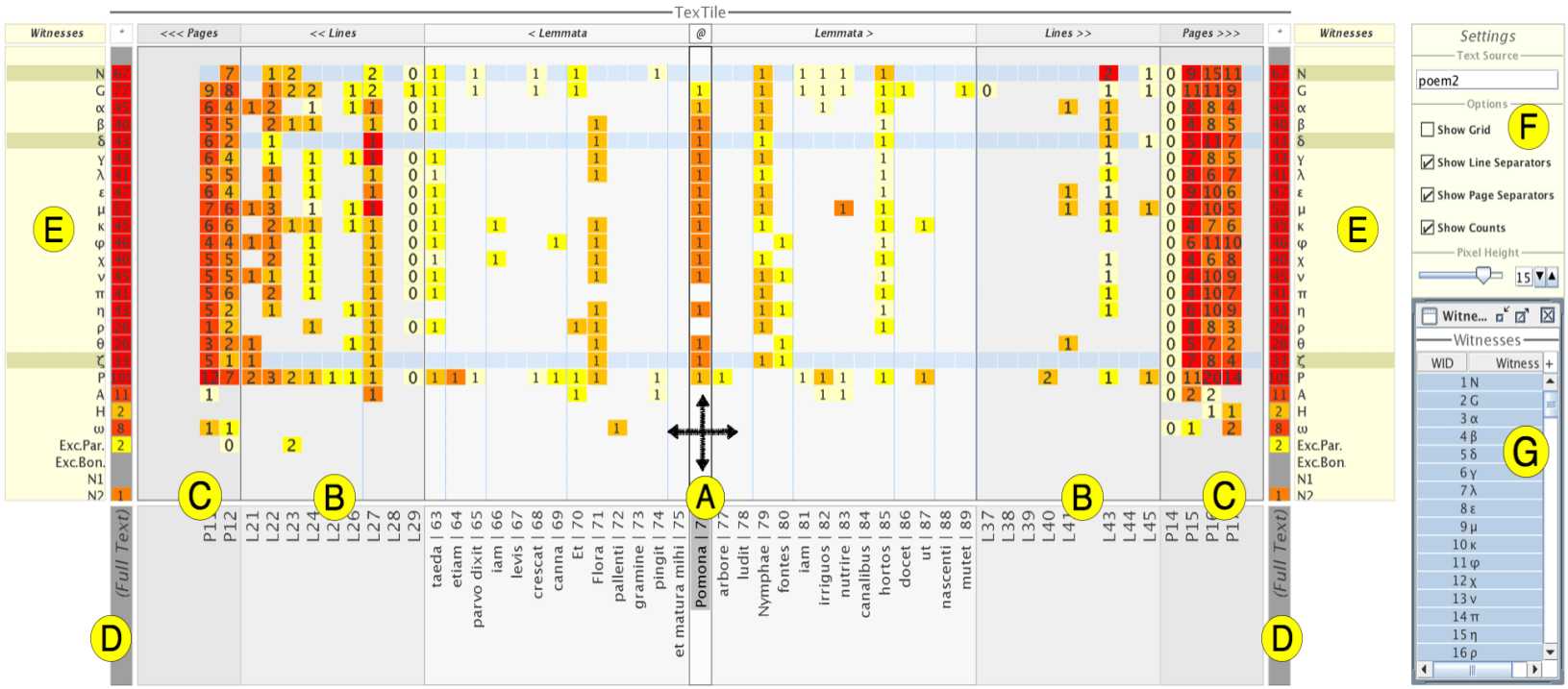


Figure 5.1: TexTile - A Pixel-based Text Analysis tool with an integrated focus+context technique. The pixels are at the intersection of witnesses and lemmata. The views are aggregated at the levels of pages, lines, and lemmata. (A) The central Lemmata view with focus on a single lemma that can be dragged horizontally and vertically to pan across all four directions. (B) Line views on either side of the central view, with pixel color and variant count aggregated at the level of lines (C) Page views on either side of the central view, with pixel color and variant count aggregated at the level of pages. (D) Full text view, with a summary grand total of variant counts. (E) The list of witnesses that are considered for analysis. (F) user features that can be enabled on demand - text box to select a poem file; show grid; show line/page separators; show variant counts; and pixel height slider. (G) A multiselect witness list box, to select subsets of witnesses for analysis.

5.1.1 Tiered Views

In TexTile, each tier corresponds to a specific level of text aggregation such as lemmata, lines, and pages. As a result of user interaction in the central focus, the data to display in the adjoining views needs to be derived and filtered. The following paragraphs and Figure 5.2 describe in detail the modulus arithmetic used to aggregate data for display in each tier.

Lemmata View

The central view, as shown in Figure 5.1A, represents the Latin text at its lowest level of granularity, the lemma. The horizontal axis displays lemmata in their natural order of occurrence in the text. The vertical axis in all views displays the list of witnesses present in the critical apparatus portion of the text. The central view always has one lemma ('w') as its focus (see Figure 5.2). The focus column

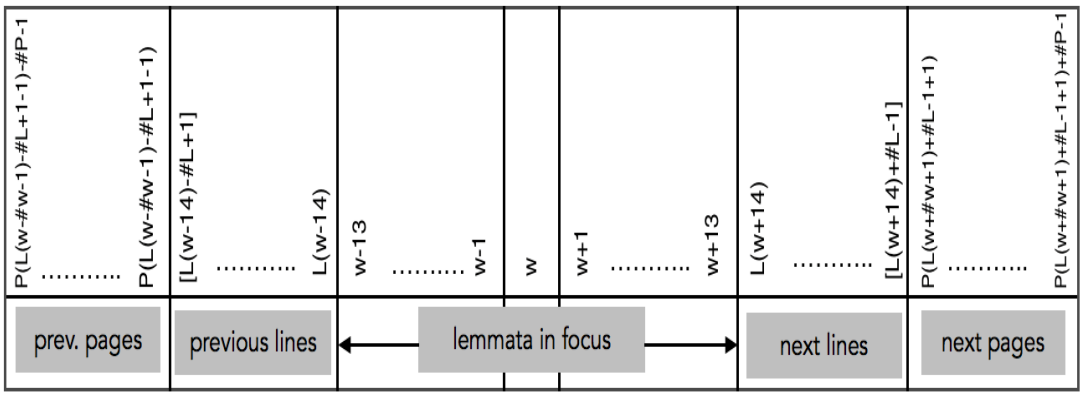


Figure 5.2: The figure depicts the modulus arithmetic for tiers on each side of the TexTile visualization. 'w' indicates the lemma in central focus. L(w) indicates the line number of lemma 'w'. P(L) represents the page number of line 'L'. #w, #L, and #P represent the fixed number of columns allocated for display in each of the tiered views (Lemmata, Lines, and Pages, respectively). These values are currently set as 13, 9, and 5 in TexTile, and were chosen based on typical lemmata per line and lines per page in our target Latin texts.

is labelled (@) in the TexTile interface.

The presence of a variant for a lemma-witness combination is indicated by a colored pixel at the intersection of columns and rows. To the left ("< Lemmata" view) and right ("Lemmata >" view) side of the central focus are 13 columns, showing immediately earlier and succeeding lemmata. The range of lemmata to display in the left Lemmata view is [w - 13, w - 1], and in the right Lemmata view [w + 1, w + 13]. As described in Section 4.2, the lemmata can be fetched from the collation table using the primary key lemma identifier (LID). If there are fewer than 13 lemmata prior to or following the lemma in focus, only the lemmata present are displayed and some columns appear blank.

Line Views

On both sides of the central Lemmata view are the line views (see Figure 5.1B). Columns in the line views show witness-variant co-occurrence for entire lines. The left ("<< Lines") and right ("Lines >>") line views display 9 columns (lines) each. Let us consider L as a function of the focus lemma position in the text, w, relative to all other lemmata in sequence. The function used to calculate the range of line numbers (and the corresponding variants) to display in the left Line view are [L(w - 14) - 8, L(w - 14)]. Similarly, the function for the right Line view are [L(w + 14), L(w + 14) + 8]. The modulus arithmetic is defined such that the innermost column of each line view shows a partially aggregated line, when that line still has lemmata visible in the Lemma view.

Page Views

On both sides of the Line views are the Page views (see Figure 5.1C). Columns in the page views show witness-variant counts aggregated at the level of entire

pages. The left ("<<< Pages") and right ("Pages >>>") Page views display 5 columns (pages) each. Let us consider P as a function of line number. The function used to calculate the range of page numbers in the left Page view are [P[L(w - 14) - 8] - 4, P[L(w - 14)]]. Similarly, the function for the right Page view are [P[L(w + 14) + 4], P[L(w + 14) + 8] + 4]. The innermost column of each Page view aggregates only the portion of a page not visible as lines in the adjacent Line view.

Full Text View

On the outside of each Page view, one additional column aggregates witness-variant counts for the entire text (see Figure 5.1D). Both sides are identical to show information for the full text and to anchor navigation. Both views effectively show average edit distance for each witness, using the same pixel color encoding as the other views.

5.1.2 Navigation

Panning in all four directions (left-right and top-bottom) is enabled in the central focus column (@). When a user navigates from one lemma to another in the central Lemmata view, pages and lines shift smoothly in the Line and Page views. Navigation supports both mouse drags and stepping with the keyboard. For example, consider the usage scenario depicted in Figure 5.1. A user interacts with TexTile to set the point of interest at lemma "Pomona". The 13 lemmata prior to and following "Pomona" are displayed in the two Lemmata views.

Using the modular arithmetic discussed in the previous section, the left Line view displays lines 21 to 29 and the right Line view displays lines 37 to 45. The



Figure 5.3: The five-level color scheme used to map Levenshtein edit distance between two strings into pixel color.

left Page view, displays pages 11 and 12. The right Page view displays pages 14 to 17 (which are the final pages in the text).

In response to interaction, including navigation, the full data set is queried through the primary/foreign key relationships that exist between the Collation, Variant and Lemma tables as discussed in Section 4.2. Each view receives a filtered subset of lemmata corresponding to its lemma, line, or page range. The subsets are aggregated (for views except the lemma views) then the witness-variant pairs are counted for each pixel.

5.1.3 Pixel Color

The presence of a variant for a lemma-witness combination is indicated by a rectangle "pixel". The color of a pixel is mapped from a string edit distance derived using a case-sensitive version of the Levenshtein edit distance [33]. The edit distance in our case indicates a degree of similarity between the lemma, and a variant (as strings of characters).

The calculated edit distance value is mapped into a five-level color scheme: pale yellow, yellow, pale orange, orange, and red. The color scheme is displayed in Figure 5.3. The range of edit distance mapped into the five colors is in increasing level of dissimilarity: 0, 1, 2 – 3, 4 – 9, and >27. For the Line and Page views, the individual edit distances are summed over the entire line or page, then mapped into the same five-scale color scheme. As a result, color accumulates in a sensible manner in the innermost columns of the Line and Page views during navigation.

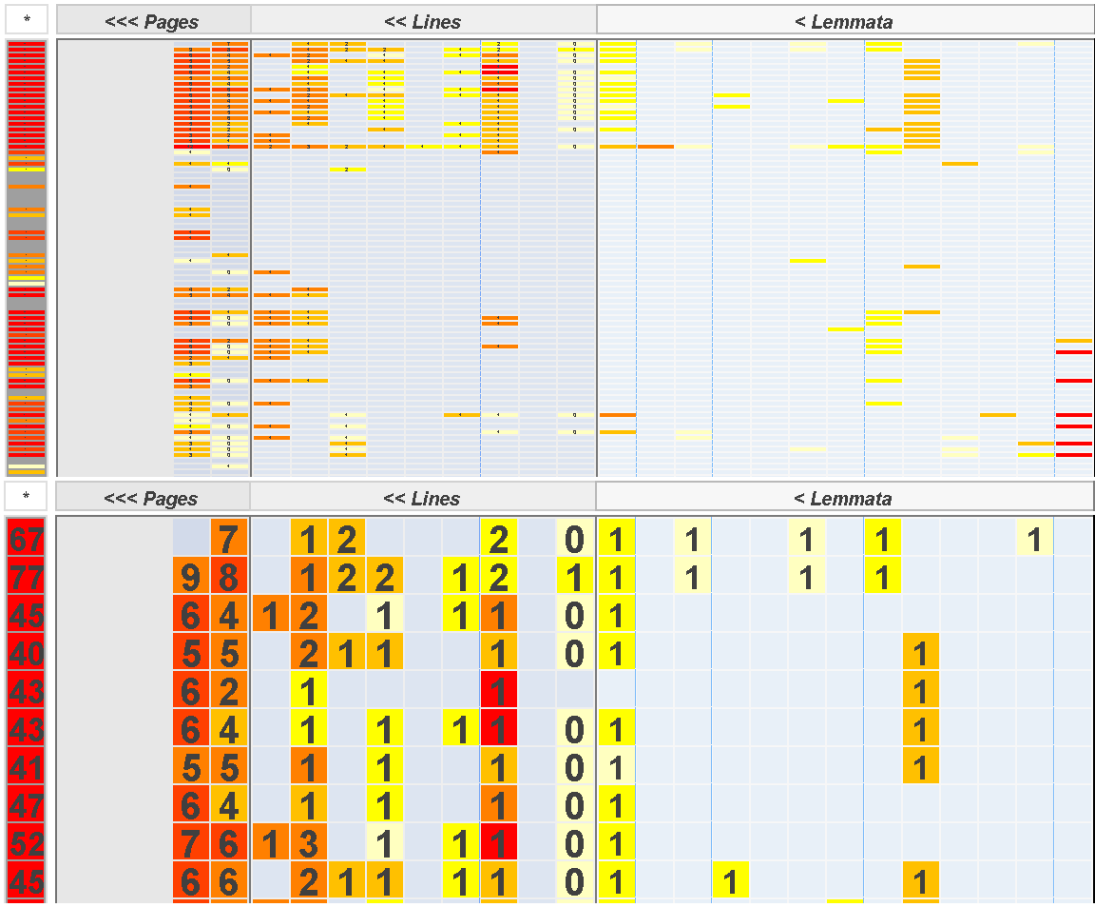


Figure 5.4: TexTile views at pixel height zoom levels of 3 (top) and 20 (bottom).

5.1.4 Interaction and Query Features

To indicate the start and end of a page or line in each of the context views, two checkboxes (refer Figure 5.1F), "Show Line Separators" and "Show Page Separators" are provided. Users can enable these features on demand. This helps to keep track of one's reading location in the entire text. Enabling the "Show Counts" checkbox displays the actual total number of variants on top of each pixel, aggregated at the corresponding level of text. As a user traverses from one lemma to another in the central Lemma view, the variant count is summed up at the level of lines and displayed as the total count in the rightmost column of the



Figure 5.5: TexTile filtered to show a subset of witnesses selected in the multiselect Witness list box at bottom right.

left Line view and the leftmost column of the right Line view. The corresponding calculation is applied in the Page views.

By enabling scrolling along the vertical axes, users can interact with the text data and view a long list of witnesses and their variants. Displaying more records than the available display space allows is a well-known issue (discussed for instance for Chimera's Value Bars [15]). Enabling scroll bars would help to overcome this issue to a certain extent, but would also create the overhead of users scrolling until they reach a desired point. To provide an alternative for displaying more data in the available space, we provide a "Pixel Height" slider. This feature allows the user to adjust the pixel height to display more or fewer witness rows vertically. Figure 5.4 displays the left half of TexTile with row heights 20 (fewer than 10 witnesses visible) and 3 (more than 30 witnesses visible).

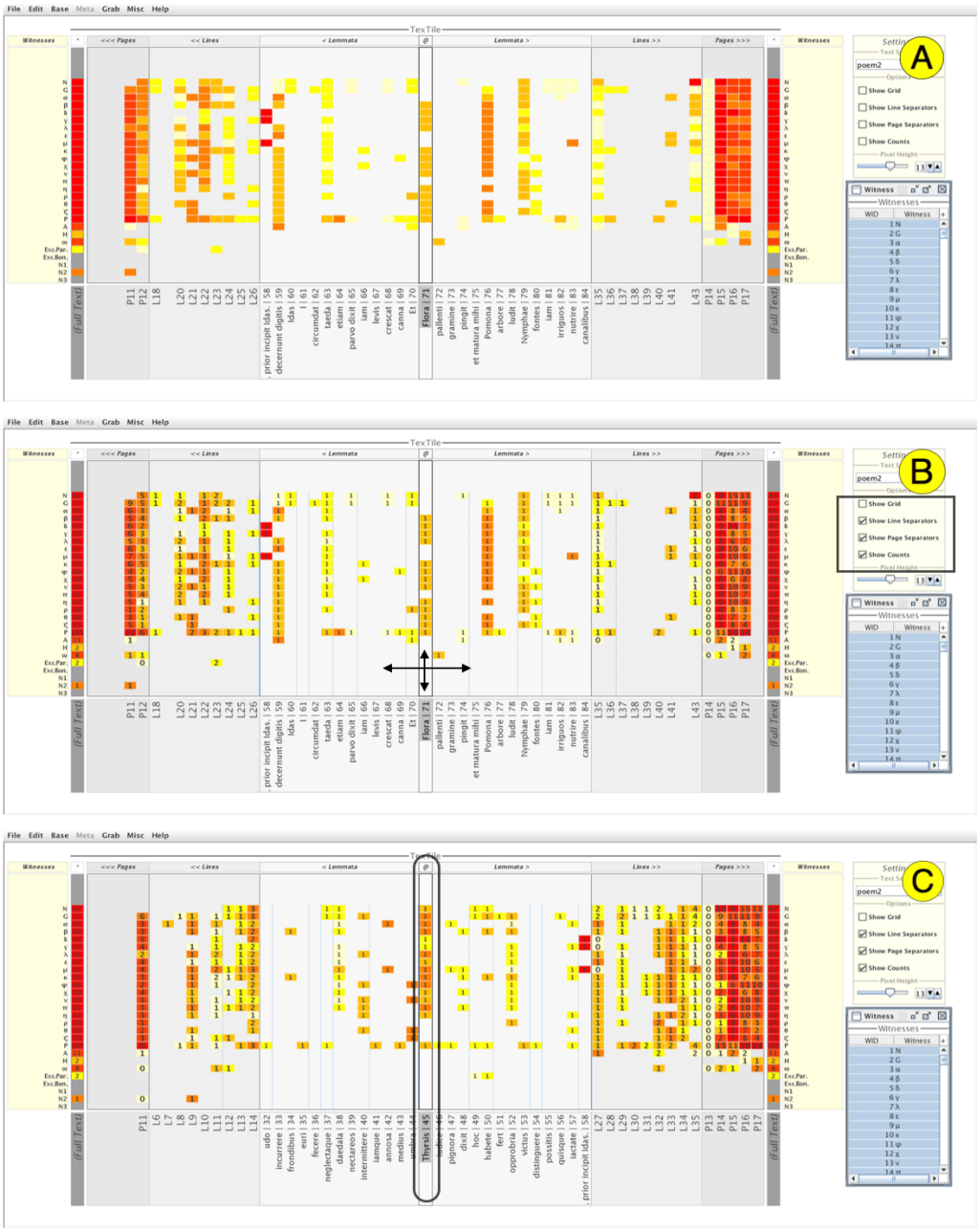


Figure 5.6: Sequence of interactions involved in performing a pattern recognition task using TexTile. *"How many groups of colored variants do you see for lemma 'Thyrsis' on page 12, line 21?"* (using poem 2 of Giarratano's 1910 critical edition of Calpurnius Siculus). A) Initial visualization state. B) Enabling the page and line separators for guided navigation. C) Identifying the lemma in the central focus view and analyzing degree of similarity between variants based on pixel colors. Interaction and querying identifies three groups of variants.

The data set can be filtered to view a subset of witnesses and their variants using a multiselect witnesses list box (see Figure 5.1G). This querying feature allows users to view only those witnesses required for their analytic needs. An example scenario is shown in Figure 5.5. In addition, outermost "Witnesses" views (Figure 5.1E) bracket the full text ('$*$') tiers. These views can be used to highlight one witness or a set of witnesses of interest. For each selected witness, the entire row is highlighted with a darker blue background. This feature helps the user to scan the rows of witnesses of current interest. An example analysis task, with the sequence of interactions involved in identifying an interesting pattern of similarity among variants, is shown in Figure 5.6.

5.2 Evaluation

The first user study conducted to assess the usability of the prototype pixel-based tool, helped us identify the highlights and shortcomings of the tool. Similar to the previous user study, we conducted one to evaluate the usability of TexTile. The study goals are to assess the overall usability, the learnability of an integrated focus+context based visualization, and the effectiveness of the pixel color encoding. Finally, we compare the TexTile study results with the evaluation results for the prototype tool.

5.2.1 Experimental Settings

The user study was carried out with 15 participants from both student and academic communities. All participants had normal or corrected to normal vision. Each session began with an overview of the Latin data set and the design features of TexTile. This was followed by a brief training phase, to explain to participants

the various tasks that can be performed using the tool. Questions that arose during the training session were addressed and tasks clarified. Participants also received a copy of the training material for reference, during the study if desired.

After the training phase, the task phase with 13 quantitative tasks and 10 qualitative questions began. The task phase lasted for approximately an hour per participant. At the end of the study, user feedback was elicited. The purpose of qualitative questions and debriefing was to obtain comments and recommendations concerning the experimental procedure, tool design, and overall usability of visualization. A complete list of the tasks used in this study is provided in Appendix B. Tasking, grouping, data collection, and analysis methods were all similar to the earlier study as explained in Sections 4.4.1 and 4.4.3.

5.2.2 Study Results And Discussion

Quantitative Tasks

The results of quantitative tasks are provided in Figure 5.7. The three metrics used for assessing TexTile are accuracy, confidence level in answers, and time taken to perform the task.

For instance in Figure 5.7, it is observed for task group I, that task 3 takes longer time (mean = 2.2 minutes) to perform. This task corresponds to the question: *Select witness ids 1, 7, 9, 11, 13, 16, 20, 22. On page 12, line 32, for lemma "Flora", what are the witness ids that agree to a variant of same degree of similarity?*. This task involves multiple different interaction states such as selection of witnesses from the witness list, navigating to the lemma "Flora", and then comparing the pixel colors of all witnesses. These multiple interaction steps are all necessary to complete the task. On the other hand, we observed that the

Task Number	Task Group	Task to perform *(Multiple choice answers provided in the study)*
1	N,P	Select witness ids 1 to 15 in the "Witnesses" list box. Navigate to poem 2, page 12, and line 21, lemma "Thyrsis". How many groups of colored variants do you see?
2	I,N	For the same lemma "Thyrsis" as in the previous question, what are the first line and page numbers that you see in the right side "Lines" and "Pages" views?
3	P	For the same lemma "Thyrsis" as in the previous question, what are the variants with the highest dissimilarity?
4	I	Enable "Show Counts" checkbox. On page 11, which witness has the most number of variants?
5	I, P	Select witness ids 1, 7, 9, 11, 13, 16, 20, 22. On page 12, line 32, lemma "Flora" what are the witness ids to agree to the same variant?
6	I, N	Enable "Show Line Separators" checkbox. Select witness ids 1,2 and 3. In lines 34 and 35, what are the lemmata with variants for all the 3 witnesses?
7	I	Enable "Show Page Separators" checkbox. Find the start and end line numbers on page 13.
8	P	Type "poem3" in Text Source textbox and press "Enter". Select witness ids 19 to 25. In the "Full Text" view, which witness id has the highest and lowest variant count?
9	N,P	Select witness ids 1 to 20. Compare lines 63 and 66. Do you see a common pattern in the witness list?
10	P	Select witness ids 1 to 20. On line 8, lemma "Lycidan" which witness id has the most dissimilar variant ?
11	I	Select witness ids 1 to 20. On line 4, lemma "ruscis", what are the witness ids with a variant reading?
12	P	Poem 2, page 12, line 20-lemma "intermittere". Select all witness ids. Are all variants displaying the same level of dissimilarity?
13	N	Select witness id 1. In poem 2, page 12, and line 35 How many lemmata with a variant are present?

Table 5.1: Quantitative tasks used to assess the usability of TexTile. The task group column indicates the classification of each task based on the three synoptic task groups: Interaction-related(I), Perception-related (P), and Navigation-related (N).

time taken does not significantly effect the accuracy and confidence level in the responses provided by participants.

In task groups N and P, one task corresponds to the question: *Select witness ids 1 to 20. Compare lines 63 and 66. Do you see a common pattern in the witness list?*. This is a pattern recognition task, that involves identifying similarity in variants between witnesses for lemmata, aggregated at the level of lines. There is a notable decrease in accuracy (mean = 4), confidence level (mean = 3.7) and an increased time taken (mean = 2.1 minutes) to perform this task.

Analyzing the qualitative feedback, participants expressed that it was difficult to decide if the two lines displayed same pattern in similarity between variant groups, since there were multiple witnesses that had to be compared. This suggests, that there is a limit to the number of variants that can be compared manually by a user. This task might benefit from an interactively dynamic automated facility to reorder witnesses based on the similarity of variants.

Qualitative Tasks

At the end of each study, users were asked to answer a ten item questionnaire. The questions were designed to elicit general impressions about the usability of the tool. More attention was given to the qualitative results of this study than the previous one, in order to better capture participants' diverse experiences exploring and analyzing patterns using the tool. Wehrend et al. [58] discuss a set of domain-independent operations that a user might need to perform using a visualization tool. We follow their approach in this study using the following kinds of tasks and describe a method suitable to assess the usability of user-centered systems.

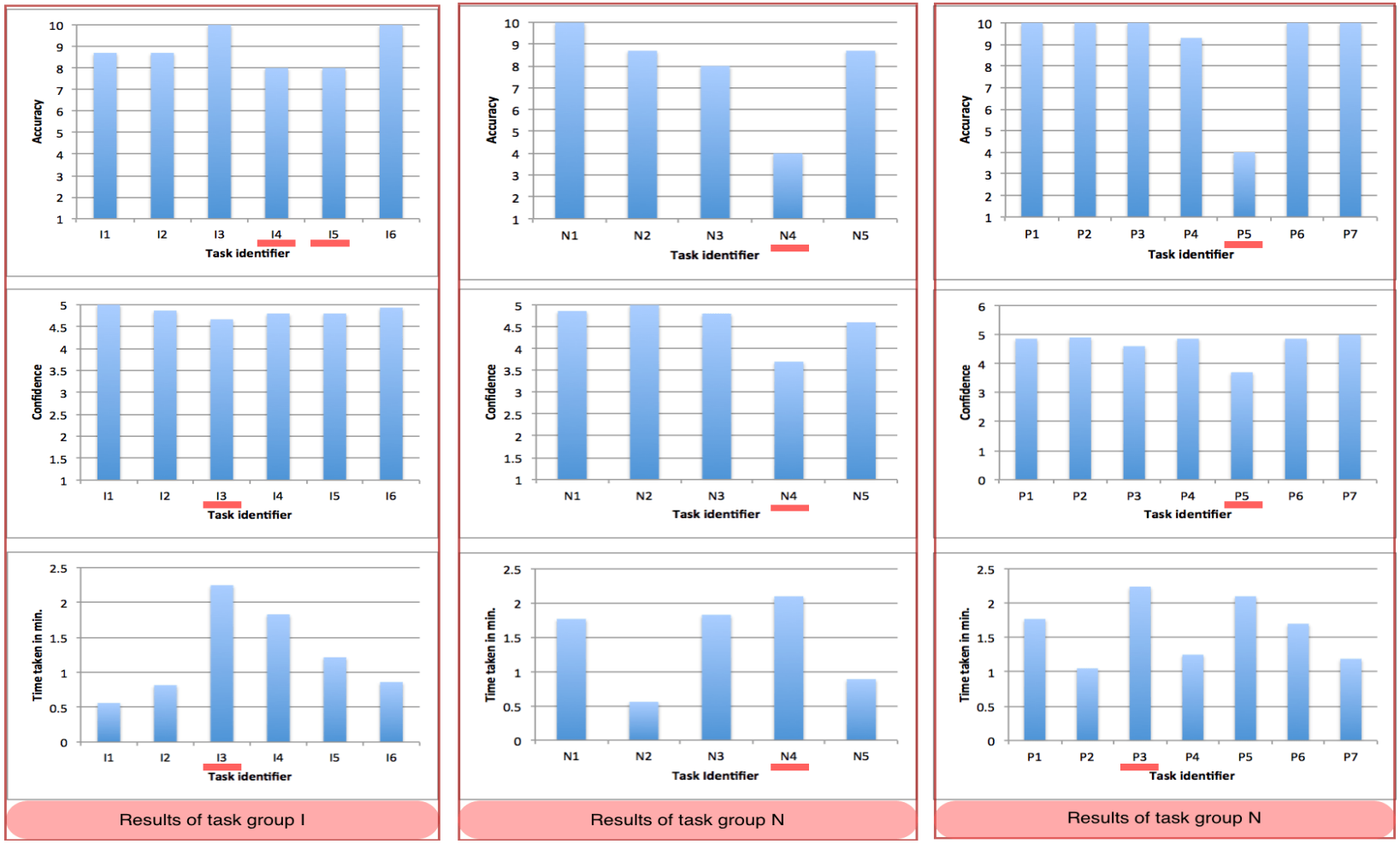


Figure 5.7: Accuracy, confidence, and time taken for tasks in the three task groups (I, P, N). The horizontal red line in the bar graph indicates a result of importance that is also discussed in the main text.

Task Number	Task
1	Can you **locate/identify** lemmata with interesting characteristics using this visualization?
2	Can you **distinguish** between similarities of variants of different witnesses using the color mapping?
3	Do you think the color mapping helped you to **categorize** variants to different groups of similarities?
4	Are you able to identify **groups/clusters** of witnesses based on variant similarity in this pixel-based visual representation?
5	Do you think this representation of data gives a clear overview of the **distribution** of variants over the whole data set?
6	Can you easily **rank** the witnesses based on the number of variants contributed using this visualization?
7	Do you think this visual representation helps you to make **comparison** between variants?
8	Do you think this visual representation helps you to form **association** between witnesses?
9	Given two witnesses, do you think relationship or **correlation** between them can be observed easily?
10	What would you suggest for improving the visualization tool?

Table 5.2: Qualitative tasks used to check usability of TexTile, inspired by Wehrend et al. [58]. Responses were entered on a 1 ("easy") to 5 ("difficult") Likert scale.

- **Locate/Identify.** Assess if the users are able to locate or identify a particular data point (lemma) from the entire data.
- **Distinguish.** Assess the tool with respect to differentiating multiple data points (similarity between two variants).
- **Categorize.** Assess the tool with respect to identifying divisions of item categories (witnesses) through visual representation (pixel color).
- **Cluster.** Assess if the users are able to group/cluster data points based on a particular feature (edit distance).
- **Distribution.** Assess if the users are able to get an overview of the distribution of data points (variants) across the entire text.
- **Rank.** Assess if the users are able to order the items (witnesses) based on a metric (variant count).
- **Compare.** Assess if the users are able to make comparisons between similar items (variant similarity with lemma).
- **Associate.** Assess if the users are able to identify any relationship or trend that exists between items (distribution of variants for subsets of witnesses).
- **Correlate.** Assess if the users are able to find relationship between two given data items (pairs of witnesses).

Based on the above tasks, the qualitative questions were formulated and used in the study. The results are provided in Figure 5.8. The answers were recorded on a Likert scale with increasing level of difficulty in performing a task from 1 to 5. The operations for which the tool had a significantly high rating are distinguish,

categorize, distribution, rank, comparison and association. The operations for which the users provided a mixed rating are locate/identify, group/cluster and correlation. The reasons why the users might have faced difficulty in performing these tasks, from our observations are listed below:

- To locate/identify a particular lemma in the entire poem, a user needs to first pan the view in the central focus multiple times, until one reaches the desired location in the text. An easy mechanism such as clicking on the page or line number directly, instead of scrolling through multiple lemmata in the central focus would be a possible solution. Users also mentioned that, if line number is mentioned in the lemmata view, it can help them easily locate a lemma.

- Many users (10 out of 15) mentioned that it was hard to differentiate between the page and line separators, since both were represented using the same color encoding (a vertical blue line).

- Orientation of text in the horizontal axis posed issues for some users and this was also highlighted in the earlier study. This is an unsolved visualization problem in general and has yet to be addressed. Some users mentioned that swapping the data displayed in the two axes could be helpful. They also stated that displaying the lemmata along the vertical axis would let them read that text in the normal horizontal orientation. However, doing so would go against the conventional left-to-right reading order of Latin text that the current design retains. By altering the design, the user would need to read and scroll text word-by-word vertically. This issue calls for further discussion with domain experts to understand their preference regarding

readable presentation of Latin text. This is identified as a part of our future work.

- Selecting subsets of data, to view a set of witnesses, can be performed using multiple selection in the witness list box. This feature helps the user analyze variation by putting witness rows in close proximity. Whereas, when the witnesses are located farther from one another, rows are harder to compare. A feature to reorder witnesses, to prioritize their proximities could be a helpful feature.

- Latin experts expressed their desire to have a summary of variant count and degree of similarity along the horizontal axis. This is similar to the current "Full text" view along the vertical axis. Scholars mentioned that this additional feature would help them more readily scan for lemmata having high degrees of uncertainties (due to substantial variation across witnesses, for instance).

- Custom visual encoding of particular variant correction types, such as omissions made them easier to spot.

- Many users expressed that they would like to see a popup of the actual variant text when they mouse-over a pixel.

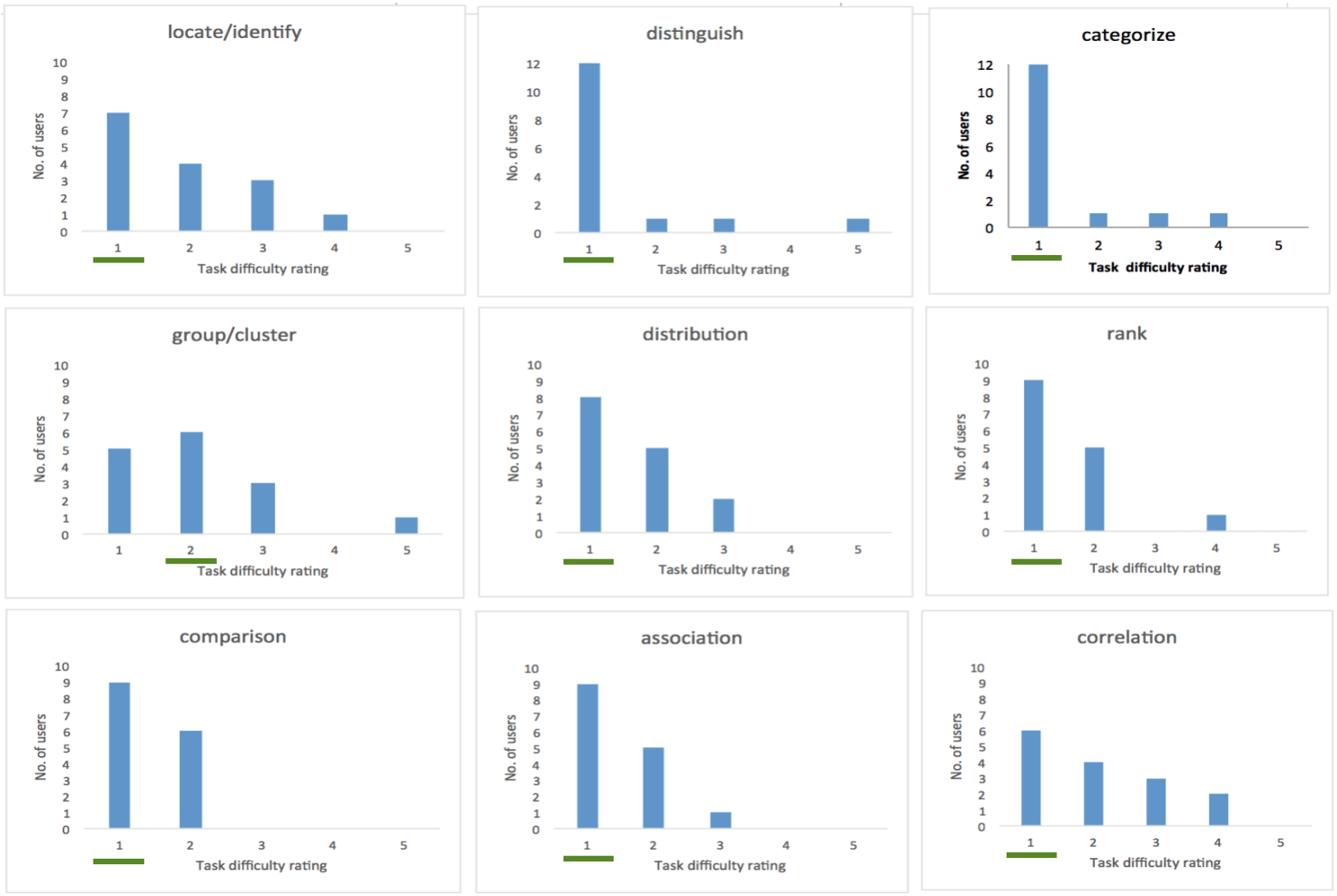


Figure 5.8: Qualitative results of nine different domain-independent operations a user might need to execute to analyze data, following Wehrend et al. [58]. The green bar under each bar graph indicates the best performing task.

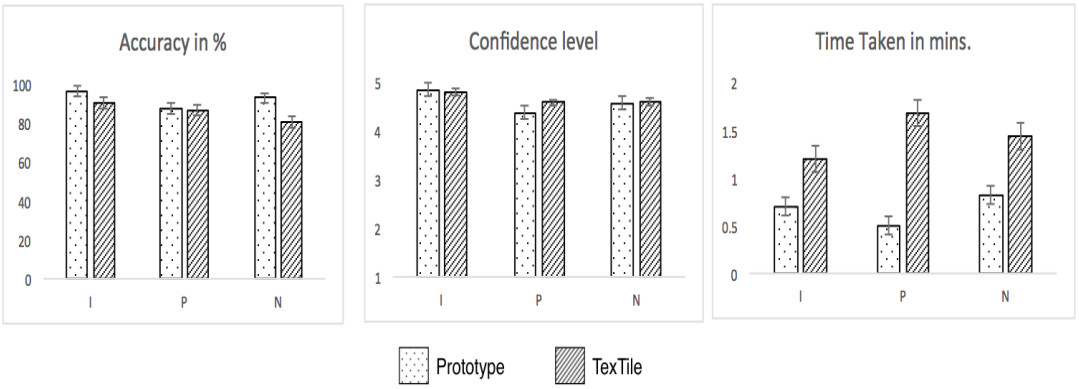


Figure 5.9: Comparison of the prototype and TexTile pixel-based text analysis tool with respect to the three performance metrics. The horizontal axis indicates the three task groups(I, P, and N).

5.3 Summary

We compared the three user study metrics results between the prototype and the TexTile visualization tool. The results are shown in Figure 5.9. With respect to accuracy in answers, the results are empirically the same for task groups I and P. For task group N, there is a 12% decrease in accuracy achieved using TexTile. The confidence level does not correlate with accuracy. The confidence level in answers is statistically the same between the two versions of tool for I and P tasks. For task group P, TexTile shows an improvement in confidence. This is an indication that the new analogous color scheme fares better and helps users perform perception-related tasks like similarity comparison.

Tasks take longer time to perform in TexTile. The prototype tool was straightforward in its design. Many participants were already familiar with separate drill-down views. In contrast, TexTile has an unfamiliar integrated focus+context design. This requires users to traverse a new arrangement of views having a hierarchy of focus+context relationships. We suspect that the limited training

period (10 – 20 minutes) was insufficient to gain familiarity. This is a learnability issue. It will be interesting to verify if there is an improvement in usability with respect to time, with a longer training session and hands-on experience.

Chapter 6

Conclusion

In this thesis, we have described novel pixel-based focus+context variant analysis tools for scholars to analyze Latin critical editions. The tools draw from and innovate on well-established visualization techniques used to display and interact with text at multiple scales. We anticipate that correlating witnesses and grouping them into categories will help scholars analyze prior versions of the text and propose improved reconstructions. Our user studies verify the usability and level of accuracy in interpretation that can be achieved using the tools. The pixel-based and focus+context aspects of design compliment each other well and promote efficient data exploration.

The pixel-based visual text analysis technique allows classics scholars to perform textual criticism effectively and efficiently. The technique also acts as a medium to capture snapshots of one's analysis which may help scholars to analyze texts collaboratively. This thesis contributes:

- an implemented prototype tool for variant exploration and analysis, with a quantitative evaluation to assess the potential of combination of focus+context and pixel-based visualization techniques;

- an implementation of TexTile, a novel pixel-based visualization design, that allows continuous focus+context navigation of full text across scales;
- an evaluation with analysis of the TexTile design and implementation; and
- an understanding of how visualization tools can be applied to text analysis in study of classical Latin.

Some of the future directions for this work follow.

- Add dynamic filtering and sorting features to let users flexibly arrange sources (witnesses) along the vertical axis.
- Explore alternative ways to display text (lemmata) along the horizontal axis, to overcome the reading issues caused by orienting text vertically.
- Embed pixel views directly inside conventional text displays including full page layouts, text blocks, and table cells. This approach might help scholars readily explore variants in context as they read the base text.
- Apply TexTile to data sets for texts in other Classic languages, such as *The New Testament* in Greek, to explore textual discrepancies between early translations.
- Use the tool to explore modern prose texts, particularly those that contain varying amounts of hierarchical structure.

We anticipate that successful application of new interactive visualization techniques and creation of pedagogical tools for text analysis in a complex domain like classics will provide a clear direction for application to humanities scholarship more generally.

Bibliography

[1] CollateX. http://collatex.net/. Accessed: 2016-04-13.

[2] Digital Latin Library. http://digitallatin.org/. Accessed: 2016-03-23.

[3] Juxta Software. http://www.juxtasoftware.org. Accessed: 2016-04-13.

[4] Philology. https://en.wikipedia.org/wiki/Philology. Accessed: 2016-04-11.

[5] TEI Consortium. http://www.tei-c.org. Accessed: 2016-04-13.

[6] Textual Criticism. http://theodora.com/encyclopedia/t/textual_criticism.html. Accessed: 2016-04-12.

[7] Visualizing Variation. http://individual.utoronto.ca/alangaley/. Accessed: 2016-04-13.

[8] D. Albers, C. Dewey, and M. Gleicher. Sequence Surveyor: Leveraging Overview for Scalable Genomic Alignment Visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2392–2401, Dec 2011.

[9] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[10] B. Asokarajan, J. Abbas, S. Huskey, and C. Weaver. Pixel-oriented Visualization for Analyzing Classical Latin Texts. Poster presented at IEEE Conference on Information Visualization 2015.

[11] B. Asokarajan, R. Etemadpour, J. Abbas, S. Huskey, and C. Weaver. Visualization of Latin Textual Variants using a Pixel-Based Text Analysis Tool. In *EuroVis Workshop on Visual Analytics*. The Eurographics Association, 2016.

[12] P. Baudisch, B. Lee, and L. Hanna. Fishnet, a Fisheye Web Browser with Search Term Popouts: A Comparative Evaluation with Overview and Linear View. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pages 133–140, New York, NY, USA, 2004. ACM.

[13] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[14] S. Carpendale. *Information Visualization: Human-Centered Issues and Perspectives*, chapter Evaluating Information Visualizations, pages 19–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[15] R. Chimera. Value Bars: An Information Visualization and Navigation Tool for Multi-attribute Listings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pages 293–294, New York, NY, USA, 1992. ACM.

[16] A. Cockburn, A. Karlson, and B. B. Bederson. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces. *ACM Comput. Surv.*, 41(1):2:1–2:31, Jan. 2009.

[17] C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing Document Content Using Language Structure. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'09, pages 1039–1046, Chichester, UK, 2009.

[18] G. Das, R. Fleischer, L. Gasieniec, D. Gunopulos, and J. Karkkainen. Episode Matching, 1997.

[19] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, F. de Oliveira, M. Cristina, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 21(1):81–94, 2015.

[20] J.-D. Fekete and N. Dufournaud. Compus: Visualization and Analysis of Structured Documents for Understanding Social Life in the 16th Century. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 47–55, New York, NY, USA, 2000. ACM.

[21] T. M. J. Fruchterman and E. M. Reingold. Graph Drawing by Force-directed Placement. *Softw. Pract. Exper.*, 21(11):1129–1164, Nov. 1991.

[22] G. W. Furnas. Generalized Fisheye Views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[23] G. W. Furnas and X. Zhang. MuSE: A Multiscale Editor. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST '98, pages 107–116, New York, NY, USA, 1998. ACM.

[24] J. Heer, M. Bostock, and V. Ogievetsky. A Tour Through the Visualization Zoo. *Commun. ACM*, 53(6):59–67, June 2010.

[25] 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, 2015.

[26] D. Jerding and J. Stasko. The information mural: A technique for displaying and navigating large information spaces. In *Proceedings on Information Visualization, 1995*, pages 43–50, Oct 1995.

[27] D. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, Jan 2000.

[28] D. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *VAST07: IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122, Oct 2007.

[29] D. A. Keim, H. P. Kriegel, and M. Ankerst. Recursive pattern: a technique for visualizing very large amounts of data. In *Visualization, 1995. Visualization '95. Proceedings., IEEE Conference on*, pages 279–286, 463, Oct 1995.

[30] A. Khella and B. B. Bederson. Pocket PhotoMesa: A Zoomable Image Browser for PDAs. In *Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, MUM '04, pages 19–24, New York, NY, USA, 2004. ACM.

[31] S. Koch, M. John, M. Wrner, A. Mller, and T. Ertl. VarifocalReader - In-Depth Visual Analysis of Large Text Documents. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1723–1732, Dec 2014.

[32] R. Kosara, S. Miksch, and H. Hauser. Semantic Depth of Field. In *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, INFOVIS '01, pages 97–, Washington, DC, USA, 2001. IEEE Computer Society.

[33] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics Doklady*, volume 10, page 707, 1966.

[34] H. Levkowitz and G. T. Herman. Color scales for image data. *IEEE Computer Graphics and Applications*, 12(1):72–80, Jan 1992.

[35] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The Perspective Wall: Detail and Context Smoothly Integrated. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pages 173–176, 1991.

[36] C. Monroy, R. Kochumman, R. Furuta, and E. Urbina. Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants among Documents. 2539:39–49, 2002.

[37] F. Moretti. Distant Reading. *Comparative Critical Studies*, 10(3):409–412, 2013.

[38] R. Mullin. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. *Canadian Journal of Statistics*, 13(2):167–168, 1985.

[39] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context with Guaranteed Visibility. *ACM Trans. Graph.*, 22(3):453–462, July 2003.

[40] S. B. Needleman and C. D. 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, 1970.

[41] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color Compatibility from Large Datasets. *ACM Trans. Graph.*, 30(4):63:1–63:12, July 2011.

[42] D. Oelke, H. Janetzko, S. Simon, K. Neuhaus, and D. A. Keim. Visual Boosting in Pixel-based Visualizations. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'11, pages 871–880, 2011.

[43] P. Pirolli and S. Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. McLean, VA, 2005.

[44] R. Rao and S. K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pages 318–322, New York, NY, USA, 1994. ACM.

[45] J. C. Roberts. Multiple view and multiform visualization. *Proceedings of SPIE - The International Society for Optical Engineering*, 3960:176–185, 2000.

[46] J. C. Roberts. State of the Art: Coordinated Multiple Views in Exploratory Visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV '07. Fifth International Conference on*, pages 61–71, July 2007.

[47] R. Sadana, T. Major, A. Dove, and J. Stasko. OnSet: A Visualization Technique for Large-scale Binary Set Data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1993–2002, Dec 2014.

[48] M. Sarkar and M. H. Brown. Graphical Fisheye Views, 1993.

[49] B. Shneiderman. Tree Visualization with Tree-maps: 2-d Space-filling Approach. *ACM Trans. Graph.*, 11(1):92–99, Jan. 1992.

[50] J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 57–65, 2000.

[51] K. L. Summers, T. E. Goldsmith, S. Kuhica, and T. P. Caudell. An Experimental Evaluation of Continuous Semantic Zooming in Program Visualization. In *Proceedings of the Ninth Annual IEEE Conference on Information Visualization*, INFOVIS'03, pages 155–162, Washington, DC, USA, 2003. IEEE Computer Society.

[52] J. J. Thomas and K. A. Cook. A Visual Analytics Agenda. *IEEE Comput. Graph. Appl.*, 26(1):10–13, Jan. 2006.

[53] Titus Calpurnius Siculus and Caesar Giarratano. *Calpurnii Et Nemesiani Bucolica*. 1910.

[54] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near Martha? Exploring name entities in literary text collections. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 107–114, Oct 2009.

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

[56] C. Weaver. Building Highly-Coordinated Visualizations in Improvise. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 159–166, 2004.

[57] C. Weaver, D. Fyfe, A. Robinson, D. Holdsworth, D. Peuquet, and A. M. MacEachren. Visual Exploration and Analysis of Historic Hotel Visits. *Information Visualization*, 6(1):89–103, Mar. 2007.

[58] S. Wehrend and C. Lewis. A Problem-oriented Classification of Visualization Techniques. In *Proceedings of the 1st Conference on Visualization '90*, VIS '90, pages 139–143, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[59] M. Wörner and T. Ertl. *SmoothScroll: A Multi-scale, Multi-layer Slider*, chapter Computer Vision, Imaging and Computer Graphics. Theory and Applications: International Joint Conference, VISIGRAPP 2011, Vilamoura, Portugal, March 5-7, 2011. Revised Selected Papers, pages 142–154. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[60] J. Ziolkowski. "What Is Philology": Introduction. *Comparative Literature Studies*, 27(1):1–12, 1990.

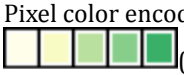
Appendix A

Evaluation Material - Study A

The training material and tasks used for the user study of prototype pixel-based focus+context text variant analysis tool is provided in the following pages.

Pixel-oriented text visualization of Latin text
Training session (15 minutes)

Briefly discuss about the different views and features in the tool:

- Base text view
  - X axis – List of words
  - Y axis – List of witnesses (W1, W2,…W22)
- Pixel Overview or Summary view
  - Selection of a pixel – changes that occur on vertical and horizontal axis of the selected point. (Red Oval pixels and Blue square pixels)
  - Pixel color encoding  (On a scale of 5, similar to more dissimilar)
  - Variant type "Omission" is represented by a purple square pixel
  - Pixel color on Page, Line and Word View represents the total number of variants present on each hierarchy. Darker pixel color represents more number of variants present for that witness.
  - Lens filter to select a range of words
- 3 other pixel views – page, line and word level (Each pixels represent no. of variants on each level)
- Check box "Only tokens with variants"

Training tasks:

1) On Overview, which axis represents the words with variants?
   a. X-axis
   b. Y-axis

2) On Overview, page 2 between words, "arundine" and "Ladon", which word has a more varied pixel color pattern?
   a. Arundine
   b. Ladon
   c. Both are same

3) How is the variant type "Omissions" represented in Overview?
   a. Red Oval pixels
   b. Blue square pixels
   c. Purple square pixels

4) On Overview page 1, click on any pixel for word "genista". Which line in Base Text View is highlighted in Yellow?
   a. 3
   b. 4
   c. 5

5) In Page View, between pages 1 and 2 which has the most number of variants for W1?
   a. Page 1
   b. Page 2

Pixel-based text analysis tool - User study

Confidence level in each question is in a scale of -
1(Low confidence) to 5(High confidence)

Quantitative tasks: (Time limit – 20 minutes)

1) How many words (count) with variants are present on Page 1?

   Confidence Level [1(Low) to 5(High)] –

2) In page 1, which word has a varying pixel color pattern on its Y-axis?
   a) Vaccae
   b) Corydon
   c) Bullantes

   Confidence Level [1(Low) to 5(High)] -

3) In page 2, how many omissions (count) are there for the word "C."
   (Hint:C. occurs multiple times on page 2)

   Confidence Level [1(Low) to 5(High)] -

4) In page 2, click on any pixel for word "Fauni". From the base text view, find out how many occurrences (count) of "Fauni" is present in this poem?

   Confidence Level [1(Low) to 5(High)] -

5) In page 7, click on the pixel for W3 and word "aures". What are the other witnesses that are highlighted in blue?
   a) W11, W14, W18, W19, W20, W21
   b) W11, W14

   Confidence Level [1(Low) to 5(High)] -

6) Which two words on page 3 have omissions for the same word by W16 and W17?
   a) C. and O.
   b) O. and raesaepia
   c) C. and Raesaepia

   Confidence Level [1(Low) to 5(High)] -

7) On page 4, click on the pixel that represents variant (W2) for word “carcare”. Which is the first and last word on page 4 that has a variant reading by W2?
    a) Vinctas and lassabit
    b) Deus and lassabit
    c) Quae and carcare

    Confidence Level [1(Low) to 5(High)] -

8) Which page out of pages 1, 2 and 3 has the most number of omissions?
    a) 1
    b) 2
    c) 3
    d) 2 and 3

    Confidence Level [1(Low) to 5(High)] -

9) On page view, which witness has entered the highest number of variants on Page 3?
    a) W1
    b) W2
    c) W3
    d) W4

    Confidence Level [1(Low) to 5(High)] -

10) On page 3, what is the first and last line number that you see on the line view?
    a) 13 and 27
    b) 28 and 45
    c) 46 and 61

    Confidence Level [1(Low) to 5(High)] -

11) On line 28, how many different variants are present?
    a) 5
    b) 2
    c) 3

    Confidence Level [1(Low) to 5(High)] -

12) On line 30, how many different variants are present?
    a) 1
    b) 4
    c) 3

Confidence Level [1(Low) to 5(High)] -

13) On line view for W1, between lines 13 to 27, which line has more number of variants?
    a) 13
    b) 18
    c) 23
    d) 24

Confidence Level [1(Low) to 5(High)] –

14) On line view, line 39 - What are the two common variants for these two words? “fraxinea” and “nolit” (i.e., two witnesses have variant reading for both “fraxinea” and “nolit”)
    a) W3 and W10
    b) W10 and W16
    c) W3 and W19

Confidence Level [1(Low) to 5(High)] -

15) On line view, line 45, what are the two words displayed on the word view that have a similar variant reading pattern?
    a) maternis and vicit
    b) maternis and lulis
    c) vicit and lulis

Confidence Level [1(Low) to 5(High)] -

Qualitative tasks: (10 minutes)

1) Rate the difficulty level in using this tool.
    a. 1 – Easy to use
    b. 2 – Little difficult, more training required.
    c. 3 – Managed to do it
    d. 4 – Hard
    e. 5 – Very hard

2) Which was the most challenging part. Explain briefly.

3) Where you able distinguish between different colored pixels easily?
    a. 1 – Easy to use
    b. 2 – Little difficult, more training required.
    c. 3 – Managed to do it
    d. 4 – Hard
    e. 5 – Very hard

4) Any other feedback?

Appendix B

Evaluation Material - Study B

The tasks used to evaluate TexTile, a pixel-based integrated focus+context text variant analysis tool is provided in the following pages.

User study - Evaluation of Pixel-based text analysis tool with an Integrated Focus + Context

Quantitative tasks (20-25 min.):

Confidence level in answers is in the scale of - 1(Low confidence) to 5(High confidence).
Please indicate your confidence in answer at the end of each question.

1) Select witness ids 1 to 15 in the "Witnesses" list box. Navigate to poem 2, page 12, and line 22, lemma "Thyrsis". How many groups of colored variants do you see?

   a. 1
   b. 2
   c. 3
   d. 4

   Your confidence level in answer - 1 2 3 4 5

2) For the same lemma "Thyrsis" as in the previous question, what are the first line and page numbers that you see in the right side "Lines>>" and "Pages>>>" views?

   a. L31, P14
   b. L32, P14
   c. L27, P13
   d. L40, P17

   Your confidence level in answer - 1 2 3 4 5

3) For the same lemma "Thyrsis" as in the previous question, what are the variants with the highest dissimilarity?

   a. Witness IDs – N, G, α, β
   b. Witness IDs – G, α, β, λ, μ, κ, φ, π
   c. Witness IDs – α, β, λ, λ, μ, κ
   d. Witness IDs – G, α, β, λ, μ, κ, φ, η

   Your confidence level in answer - 1 2 3 4 5

4) Enable “Show Counts” checkbox. On page 11, which witness has the most number of variants?

   a. G
   b. P
   c. δ
   d. λ

   Your confidence level in answer - 1 2 3 4 5

5) Select witness ids 1, 7, 9, 11, 13, 16, 20, 22. On page 12, line 32, lemma “Flora” what are the witness ids to agree to the same variant?

   a. λ, φ, ν, ρ
   b. λ, φ, A, ω
   c. λ, φ, ν, A
   d. λ, φ, ν

   Your confidence level in answer - 1 2 3 4 5

6) Enable “Show Line Separators” checkbox. Select witness ids N, G and α. In lines 34 and 35, what are the lemmas with variants for all the 3 witnesses?

   a. Pomona, Nymphae
   b. Nymphae, irriguos, hortos
   c. Nymphae, irriguos
   d. Iam, nutrine

   Your confidence level in answer - 1 2 3 4 5

7) Enable “Show Page Separators” checkbox. Find the start and end line numbers on page 13.

   a. 27 and 54
   b. 27 and 36
   c. 55 and 71
   d. 72 and 87

   Your confidence level in answer - 1 2 3 4 5

8) Type “poem3” in Text Source textbox and press “Enter”. Select witness ids 19 to 25. In the “Full Text” view, which witness id has the highest and lowest variant count?

a. A and N1
b. P and A
c. P and N1
d. H and N1

Your confidence level in answer - 1 2 3 4 5

9) Select witness ids 1 to 20. Compare lines 63 and 66. Do you see a common pattern in the witness list?

a. Yes
b. No
c. Not sure

Your confidence level in answer - 1 2 3 4 5

10) Select witness ids 1 to 20. On line 8, lemma “Lycidan” which witness id has the most dissimilar variant ▭?

a. μ, κ
b. κ, φ
c. φ
d. κ

Your confidence level in answer - 1 2 3 4 5

11) Select witness ids 1 to 20. Increase the Zoom level to 20. On line 4, lemma “ruscis”, what are the witness ids with a variant reading?

a. δ, γ, λ
b. δ, γ, ε
c. δ, γ
d. λ, ε

Your confidence level in answer - 1 2 3 4 5

12) Poem 2, page 12, line 20-lemma “intermittere”. Select all witness ids. Are all variants displaying the same level of dissimilarity?

a. Yes
b. No
c. Not sure

Your confidence level in answer - 1 2 3 4 5

13) Select witness id N. In poem 2, page 12, and line 35 – How many lemmas with a variant are present?

Your confidence level in answer - 1 2 3 4 5

Quantitative tasks: (10 – 15 min.)

1) Can you **locate/identify** lemmas with interesting characteristics using this visualization?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

2) Can you **distinguish** between similarities of variants of different witnesses using the color mapping?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

3) Do you think the color mapping helped you to **categorize** variants to different groups of similarities?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

4) Are you able to identify **groups/clusters** of witnesses based on variant similarity in this pixel-based visual representation?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

5) Do you think this representation of data gives a clear overview of the **distribution** of variants over the whole data set?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

6) Can you easily **rank** the witnesses based on the number of variants contributed using this visualization?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

7) Do you think this visual representation helps you to make **comparison** between variants?

   Circle any one - (Easy) 1 2 3 4 5 (Difficult)

8) Do you think this visual representation helps you to form **association** between witnesses?

Circle any one - (Easy) 1 2 3 4 5 (Difficult)

9) Given two witnesses, do you think relationship or **correlation** between them can be observed easily?

Circle any one - (Easy) 1 2 3 4 5 (Difficult)

10) What would you suggest for improving the visualization tool?

Thank you!!

Appendix C

Study Authorization Documents

Authorization to conduct user study to evaluate pixel-based text variant analysis tool is included in the following pages.

The UNIVERSITY of OKLAHOMA

Institutional Review Board for the Protection of Human Subjects

Notice of Deferral Approval for OU Collaborator to Conduct Research

Date: April 1, 2016

Principal Investigator: Bharathi Asokarajan

IRB#: 6697
Reference#: 650351

Study Title: Evaluation of Pixel-based text analysis tool with Integrated Focus + Context technique

This letter is to notify you that the University of Oklahoma (OU) Institutional Review Board (IRB) has approved your request for OU to defer all IRB responsibilities with regard to the above-referenced study to the IRB at the Oklahoma State University. This signed and IRB- approved OU Collaborator Assurance serves as the University of Oklahoma IRB's approval for you to conduct your research under the review and authorization of the Oklahoma State University.

On behalf of the OU IRB, I have reviewed the above-referenced study and determined that it meets the criteria for deferral. As a collaborating investigator on this study, you are responsible to:

- Comply with the IRB Authorization Agreement] signed by OU collaborating researcher(s) and the Oklahoma State University Principal Investigator, Ronak Etemadpour, PhD, referencing the Oklahoma State University IRB-approved study titled, "Evaluation of Pixel-based text analysis tool with Integrated Focus + Context technique ";
- Conduct the study in a manner consistent with the requirements of the IRB's of the Oklahoma State University and the University of Oklahoma and federal regulations 45 CFR 46;
- Request approval from the Oklahoma State University IRB prior to implementing any/all modifications, as changes could affect the exempt status determination;
- Notify the Oklahoma State University IRB of any protocol deviations or unanticipated problems;
- Maintain accurate and complete study records for evaluation by the Oklahoma State University and University of Oklahoma HRPP Quality Improvement Program and, if applicable, inspection by regulatory agencies and/or the study sponsor; and
- Notify the Oklahoma State University IRB at the completion of the project.

For circumstances involving the review of uses and disclosures of protected health information (PHI) under the Health Insurance Portability and Accountability Act (HIPAA), a determination will be made between the two institutions as to who will serve as the Privacy Board, if applicable.

If you have questions about this notification or using iRIS, contact the HRPP Office at 405-325-8110 or irb.ou.edu.

Cordially,

Lara Mayeux

Lara Mayeux, Ph.D.
Vice Chair, Institutional Review Board

Institutional Review Board
Authorization Agreement

Name of Institution or Organization Providing IRB Review (Institution/Organization A):
Oklahoma State University (OSU)
IRB Registration #: IRB00001305 **Federalwide Assurance (FWA) #, if any:** FWA00000493

Name of Institution Relying on the Designated IRB (Institution B):
University of Oklahoma, Norman Campus, through the Board of Regents of the University of Oklahoma (OU)
FWA #: FWA00003191

The Officials signing below agree that OU may rely on the designated IRB for review and continuing oversight of its human subjects research described below:

( ) This agreement applies to all human subjects research covered by Institution B's FWA.

(x) This agreement is limited to the following specific protocol(s):

Name of Research Project: Evaluation of Pixel-based text analysis tool with Integrated Focus + Context technique (BU-16-25)
Name of OSU Investigator: Ronak Etemadpour, PhD
Name of OU Investigators: Bharathi Asokarajan (graduate student); Christopher Weaver, PhD
Sponsor or Funding Agency: N/A
Award Number, if any:

( ) Other (*describe*):_____

The review performed by the designated IRB will meet the human subject protection requirements of Institution B's OHRP-approved FWA. The IRB at Institution/Organization A will promptly report its findings and actions to appropriate officials at Institution B. Relevant minutes of IRB meetings will be made available to Institution B upon request. Institution B remains responsible for ensuring compliance with the IRB's determinations and with the Terms of its OHRP-approved FWA. This document must be kept on file by both parties and provided to OHRP upon request.

Signature of Signatory Official (Institution/Organization A):

Date: 3/29/16

Print Full Name: Kenneth W. Sewell, Ph.D
Institutional Title: Vice President for Research

Signature of Signatory Official (Institution/Organization B):

Date: 3/26/16

Print Full Name: Glen Krutz, Ph.D.
Institutional Title: Vice Provost for Academic Initiatives