Towards an Electronic Variorum Edition of Cervantes' *Don Quixote:* Visualizations that support preparation

Rajiv Kochumman, Carlos Monroy, Richard Furuta, Arpita Goenka, Eduardo Urbina, and Erendira Melgoza
TEES Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA
{rajiv, cmonroy, furuta, arpita, e-urbina, ere}@csdl.tamu.edu

ABSTRACT
The Cervantes Project is creating an Electronic Variorum Edition (EVE) of Cervantes' well-known *Don Quixote de la Mancha*, published beginning in 1605. In this paper, we report on visualizations of features of a text collection that help us validate our text transcriptions and understand the relationships among the different printings of an edition.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]; J.5 [**Arts and Humanities**].

General Terms
Design, Experimentation, Human Factors.

Keywords
Cervantes Project, EVE, MVED, variance visualization.

1. INTRODUCTION
The Cervantes Project is engaged in creating an Electronic Variorum Edition (EVE) of Miguel de Cervantes Saavedra's seminal *Don Quxote de la Mancha*. The EVE will contain and interlink all significant early editions of the work in textual and facsimile form. In this paper, we report on observations from our work with copies of the first printing (the *princeps*, dating from 1605) of the first book (of two) of the *Quixote*. Only about 18 copies of this printing are known to have survived, and of these we estimate that only about 12 are accessible, the others being, for example, in restricted private collections. We are well on the way to acquiring microfilm copies of these 12 editions. We believe that our current collection of 8 copies is the largest obtained for a study of this text. We have manually transcribed the text of each copy into computer-readable form, manually verified the accuracy of the transcription, and have linked text to image representation.

Elsewhere we describe the MVED [4], developed for use by an EVE's editor. The MVED's processing begins with the linked transcriptions. The editor selects one of the texts to serve as *base*. The MVED then *collates* the texts against the base, producing a list of *variances* (i.e., differences). The editor can annotate, correct and emend the texts, and select among the variances, to produce a unified text. The EVE's reader, will have access to both the unified text and the original texts, allowing the reader to independently evaluate and understand the editor's decisions.

The variances and the editing actions are stored in a database repository. Variances and editing actions are distributed across the length of the text. A visualization of the variances is useful, since the distribution of variances across the texts, the number of variances between two texts, and the length of the variances, provide information about the nature of the texts themselves, and the relationship between them.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

2. VISUALIZATION OF VARIANCES
2.1 Variance Visualizer
We decided to visualize variances so that we could see their distribution along the length of the text. We expected that this would help in many ways, for example, in identifying patterns or clusters [2, 5]. We have implemented a number of different representations, and discuss two here. The first shows variances as points distributed along the length of chapter one of six copies of the *princeps* text (Figure 1). Vertical lines indicate the end of a page. The topmost display shows all the variance points and the remaining show variances between the base text and a particular copy (numbered successively as texts 1 through 5). We note that the topmost display is often useful, but loses clarity when the

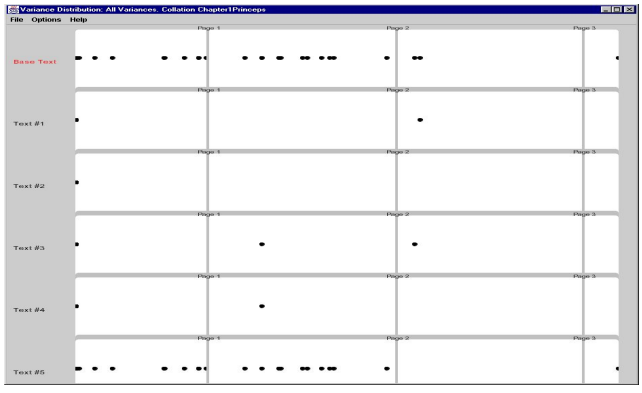

Figure 1: Variance distribution shown as points.

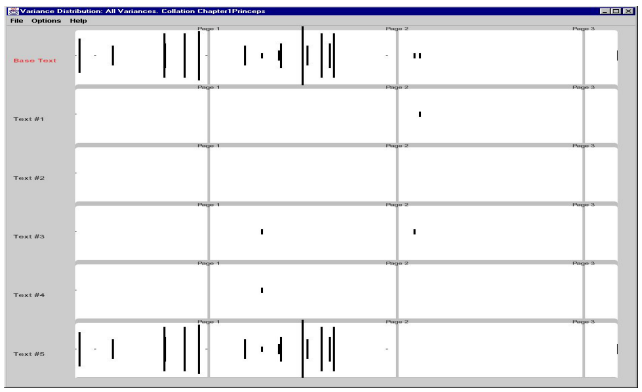

Figure 2: Variance distribution with variance length.

number of variance points is very large, since the points then tend to overlap in the visualization because several successive text locations must be mapped to the same display pixel.

In a second representation, we made the height of the rectangle, and not its width, correspond to the length of the variance (Figure 2). This representation was clearer in showing the distribution as well as the lengths of the variances, even for large data sets.

Some interesting points come out when we look at the visualizations. Consider Figure 2. The display shows that the base text and texts 1, 2, 3, and 4 are similar to each other, since the number of variance points is not too large. The base text and copy 5 are different from each other. Only a single variance separates the base text and copy 1, and that variance is not present in the other copies (the cumulative display makes it clear that the variances between the base and copy 1 and that with copy 3 are distinct).

Since all six copies are from the same printing, we would expect that variances would be small, so the large-scale differences with copy 5 stood out. Inspection revealed that the microfilm images of this copy were unclear, either because of damage to the original copy (e.g., by water staining) or because of poor camera settings. Also of interest was the single point difference with copy 1, which could signify a transcription error. In this case, the difference was found to result from a differing punctuation mark.

Of particular interest are differences that involve only one copy. We saw another instance, the inverse of the one just discussed, in Chapter 2. Here, a short difference was found in all copies except for the base copy. Inspection showed that this corresponded to an "o" that was accented in all copies except for the base; see Figure 3. This could correspond to a stop press correction, or it could correspond to wear to the type during printing. Clearly, the visualization tool does help us in identifying patterns, but we need to go back to the original image for an accurate interpretation.

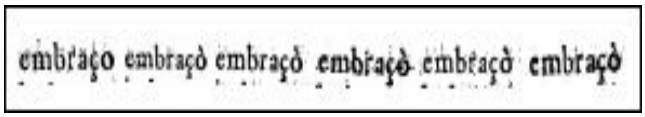

Figure 3: A variance across the 6 texts.

The single copy variance patterns have lead us to reexamine our base copy choice. Our original choice was made primarily for administrative reasons—the copy that currently serves as the base is relatively unencumbered by restrictions imposed by the copy's owner. However the visualization results suggest that perhaps another copy may be more representative. We will continue to evaluate the other copies, along with the recently-acquired copies 7 and 8, not yet fully incorporated into the set, for this role.

Our observations about the copies are echoed by those previously identified in the literature. Casasayas [1] believes that of the approximately 18 known copies of the *princeps*, that there are no two alike. This is for a variety of reasons, ranging from mundane to apparent deception. Knowles [6] says about one copy: "an amazing sample of clever repairing and binding…certainly made up of two copies…the title page is probably faked…repaired (pages) with amazing skill."

Flores [3] identifies two "family groups" of the *princeps* and others suggest that there are up to four different varieties. We expect that our visualizations will be an important tool in confirming family groups and resettings, as well as patterns of correction or deterioration of types.

From a project standpoint, visualizations provide a means of verifying our techniques, ranging from transcription accuracy to image quality of the source materials. They also serve to help us validate the algorithms used in processing the text collection—an unexpected behavior of our collation algorithm was signaled by an unusually long variance in one of the later chapters.

3. ACKNOWLEDGMENTS

Urbina and Melgoza also are affiliated with the Department of Modern and Classical Languages. The remaining authors also are affiliated with the Deparment of Computer Science. See http://www.csdl.tamu.edu/cervantes/ for project information.

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

4. REFERENCES

1. Casasayas, José M., *Ensayo de una guía de bibliografía Cervantina*, Tomo V. Mallorca, 1995.
2. Eick, S., Steffen, J.L., Sumner, E.E., "Seesoft—A Tool For Visualizing Line Oriented Software Statistics". *IEEE Trans. Software Engineering*, 1992.
3. Flores, R. M., *The Compositors of the First and Second Madrid Editions of Don Quixote, Part I*. London: The Modern Humanities Research Association, 1975.
4. Furuta, R., Kalasapur, S.S., Kochumman, R., Urbina, E., Vivancoz-Pérez, R., "The Cervantes Project: Steps to a Customizable and Interlinked On-line Electronic Variorum Edition Supporting Scholarship". *ECDL* 2001.
5. Havre, S., Hetzler, B., Nowell, L., "ThemeRiver™: In Search of Trends, Patterns, and Relationships", *IEEE Symp. Information Visualization, InfoVis* 1999.
6. Knowles, Jr., E.B., "Notes on the Madrid, 1605, editions of *Don Quijote*." *Hispanic Review* 14 (1946): 47-58.