# Poster: Understanding Tagged Text

Michael A. Correll* University of Wisconsin

Michael Gleicher† University of Wisconsin

## ABSTRACT

This poster describes visualization tools to support statistical literary analysis. Our observation of scholars' work shows the importance of connecting the large-scale statistical analysis of texts with close reading of the text. To support this work, we have designed a system that integrates a corpus-wide overview of text-properties with a viewer for examining specific passages. The corpus viewer is designed to help identify "important" features in a text; the text viewer provides a focus+context view for examining the passages containing those features. These tools, in concert, allow one to observe a corpus wide pattern of tags and then propose a causal explanation for this pattern using the techniques of passage analysis.

**Index Terms:** J.5 [Computer Applications]: Arts and Humanities—Literature

## 1 INTRODUCTION

Statistical literary analysis seeks to identify rhetorical patterns and devices by looking for patterns of words and using those patterns to identify rhetorical features. Tagging-based analysis uses rules to label words or phrases with "tags," and then looks at the statistical properties of these tags over an entire corpus of text. Such analysis can couple low-level "signatures" with higher-level concepts such as authorship and genre. However, while these techniques can display information about *what* is going on in a text, they provide little insight into *why*. Statistical methods might be *necessary* to observe a pattern in a corpus (especially for corpora which are beyond the scale of a human reader's ability to analyze in a reasonable time frame), but they may not be *sufficient* to explain the phenomena. Literary scholars follow through on the statistical analysis with close readings, analyzing the identified passages to develop causal explanations. A convincing argument from the standpoint of literary analysis is one that revisits the text with the techniques of close analysis.

One group taking advantage of text tagging for statistical literary analysis purposes is the Digital Humanities interdisciplinary group at the University of Wisconsin. This group uses automatic tagging software (specifically Docuscope [2]) to perform tasks such as suggesting rhetorical signatures that distinguish Shakespeare's comedies from his tragedies, or dating plays based on rhetorical forms of contemporaneous plays [6]. Identifying resulting rhetorical patterns and connecting them to their occurrence in the text has been labor intensive, making the scholars' work tedious and difficult to scale to larger text corpora. Our goal is to support literary scholars' work in studying large text corpora. We have developed a tool that integrates a corpus-wide overview of tagged text with specific important passages in order to facilitate the creation of sophisticated causal stories at the level of text.

Our efforts in understanding scholars' work has lead to a design that aims to support the connection between corpus-scale pattern identification and passage-scale close reading. Our prototype consists of two components: a "Corpus Viewer," designed to help identify patterns of tags of interest; and a "Text Viewer," designed to help identify and examine the specific passages fitting these patterns. The text viewer uses focus+context techniques to avoid the need for manual searches of documents that may be tens of thousands of lines long.

*e-mail: mcorrell@cs.wisc.edu

†e-mail:gleicher@cs.wisc.edu

## 2 TASK ANALYSIS

While some corpora are too large for a single human to read in detail, parsing these texts for certain rhetorical forms yields interesting results. Automated tagging software places words and phrases into certain categories which can then be used to summarize the structure of large text corpora. Unfortunately such automated tagging tools are not designed for passage selection and so are currently of limited utility for aiding close readings. Any tools we designed needed to assist seeing not just the statistical "big picture" results of a tagging scheme across an entire corpus, but also finding a method of ascribing causation for these large scale statistical phenomena using passage analysis.

The workflow of analysis involves a scholar looking at corpus wide summary of tag statistics to identify patterns of interest, followed by examining specific examples of these patterns. Since line breaks, stage directions, and dialogue labels can vary across editions and can frequently skew results, the tagging software currently strips out this information. While standard statistical views can be sufficient for examining patterns across a corpus, a tool needs not only to integrate well with the tagging system, but also provide integration with tools for examining specific passages in the text.

Once the patterns of interest have been identified, visualizing these phenomena within a specific tagged text has a number of needs. These tools must:

1. Allow the user to see global information on the distribution of tags across an entire document, in order to identify relevant passages and patterns.
2. Allow the user to easily consult passages corresponding to these patterns of interest.
3. Allow the passages to be read in context.
4. Work together with existing tools, including being consistent with the coloring and naming schemes used.

Fulfilling the first three requirements requires both local focus as well as global context. Similar tools such as SequenceJuxtaposer use the metaphor of the rubber sheet with tacked down edges to allow this sort of navigation[4][5]. The problem of passage selection has some additional requirements when compared to traditional focus+context techniques, viz.

1. Passages in focus must have a minimum per line size in order for text to be legible.
2. There must be some way of automating the selection of foci (in order to prevent our visualization tool from simply being another type of manual search, especially since lines not in focus may or may not be legible).

## 3 DESIGN

We preliminarily created two interconnected tools, a Corpus Viewer for selecting important tags based on corpus-wide patterns and a Text Viewer for selecting salient passages in a single text based

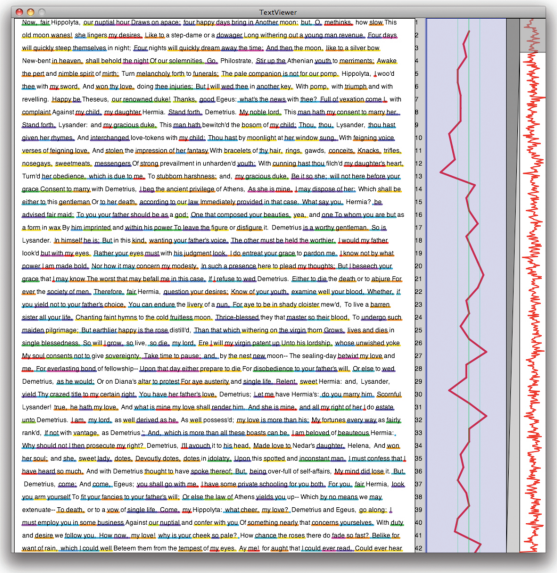

Figure 1: The initial view of a text. The graph to the far right is the global graph, which measures tag density of the entire document, with a small window showing current text position. To the left of the global graph is the local graph of tag density in the current window of viewable text. Tagged words are underlined with the color specified in the earlier subset selection task.

Figure 2: The text, after a range of relevance has been set. Notice that the window now covers a much larger portion of the text, and that portions of the text below the threshold fade into the background both in the actual text window as well as in the local graph.

on preselected important tags. The process of selecting important tag categories is amenable to any number of statistical or visual selection methods. Our Corpus Viewer tool displays aggregated, normalized counts of tags across documents, with the ability to filter out both categories of tags and groups of texts, allowing the user to generate a subset of tags that are considered relevant for some task. Once the user has selected a relevant subset of tag categories the user then selects a text to investigate using Text Viewer.

Figure 1 illustrates the initial view of the document with three main elements: a window view of the tagged text currently visible, a local graph of tag counts in the visible text, and a global graph of tag counts throughout the entire document. If the users wishes to select salient passages he or she sets threshold lines on the local graph. All lines that have tag counts within these lines are placed under a virtual lens with some variations of previous document focus+context techniques [3][1]. The lens is bridge shaped, i.e. the focus and several lines above and below it are rendered at the maximum font size, but beyond that window font size decreases linearly to some minimum. Originally this minimally sized font was greeked and rendered at a small size, but since greeked text contains no useful information for close analysis other than relative length between important lines this unimportant text is now simply not displayed, represented by a discontinuity in the local graph and a widening of the position box on the global graph. Only lines which are within a focus are labelled by line number, which adds to the ability to determine context without having to waste time and space rendering large numbers of semantically unimportant lines of text.

Figure 2 shows the document after a tag relevance range has been set. The text is now divided into a number of "bubbles" of tagged passages. If a line is part of a bubble it is a drawn in bright red on the global graph as well as the thickest possible red line on the local graph, providing context information about the relative spacing of passages in the document as a whole. If a line is not part of a bubble it is represented by thinner lines leading to a discontinuity on the local graph, and by thiner, dark red lines on the global graph.

## 4 Conclusion

The initial reaction of our domain collaborators has been positive. They appreciated the visual metaphor of irrelevant text receding into the distance and they liked the ability to select passages without having to leave the level of actual text.

Natural extensions to the tools are to develop better overviews of distributions over large corpora, allow more complicated aggregation techniques, allowing more sophisticated filtering and aggregation in the Corpus Viewer, display metadata such as line, act, and chapter breaks, and lastly to generalize from Docuscope tags to any sort of tagged document. Ideally an investigator using this tool would be able to perform passage selection based on differences determined by arbitrarily complicated statistical methods or tagging procedures and yet have most of this actual selection work hidden from the user.

## Acknowledgements

The authors wish to thank Profs. Michael Witmore and Robin Valenza of the UW-Madison English department and the rest of the Digital Humanities group for their assistance and feedback during the design process. This research was funded in part by NSF award IIS-0946598.

## References

[1] 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*, pages 133–140. ACM, 2004.

[2] J. Collins and Kaufer. Description of DocuScope, 2001. [Online]. http://betterwriting.net/projects/fed01/dsc_fed01.html [Accessed: Aug. 9,2010].

[3] G. G. Robertson and J. D. Mackinlay. The document lens. *Proceedings of the 6th annual ACM symposium on User interface software and technology - UIST '93*, pages 101–108, 1993.

[4] M. Sarkar, S. S. Snibbe, O. J. Tversky, and S. P. Reiss. *Stretching the rubber sheet*. ACM Press, New York, New York, USA, June 1993.

[5] J. Slack, K. Hildebrand, T. Munzner, and K. John. SequenceJuxtaposer: Fluid navigation for large-scale sequence comparison in context. In *German Conference on Bioinformatics*, pages 37–42. Citeseer, 2004.

[6] M. Witmore. The Funniest Thing Shakespeare Wrote? 767 Pieces of the Plays, 2010. [Online]. http://winedarksea.com/?p=600 [Accessed: Aug., 9,2010].