What Shakespeare Taught Us About Text Visualization

Michael Correll
University of Wisconsin-Madison
mcorrell@cs.wisc.edu

Michael Gleicher
University of Wisconsin-Madison
gleicher@cs.wisc.edu

ABSTRACT
In our work we have developed text visualization tools to meet the needs of literary scholars. While our work in this domain may on the surface seem quite different from other text visualization applications, we have encountered principles that generalize and will be useful for text data as distant from literature as social media. The way that digital humanities scholars use and argue about texts are not idiosyncratic to their field: their requirements and rhetoric offer implications for design generally, including renewed focus on outliers, constant connection from visualizations to underlying texts, and the ability to generate explanations for higher level patterns by moving back and forth between these patterns and low-level data.

Author Keywords
Text Analytics; Digital Humanities; Text Visualization

ACM Classification Keywords
H.5.0 Information Systems: Information Interfaces And Presentation

General Terms
Design.

INTRODUCTION

Several years ago we began a collaboration with an interdisciplinary group of researchers in the Digital Humanities, with a special interest in visualizing the history of early modern English print [5, 6]. The group was underserved by existing text visualization techniques; our original thinking was that this group was idiosyncratic enough that the research exercise would be how to adopt known techniques to new environments. Existing ethnographic information about how these literary scholars argued, used tools, and found information suggested we would need to adopt new strategies of design and deployment [2, 8].

Our conclusions from this ongoing line of research is that the digital humanities environment is neither as niche nor as idiosyncratic as we thought: rather, the design principles we developed for their text analytics applications are, in general, principles that apply to text visualization, and text analytics, as a whole. In particular:

- Text is often an end in itself and central to analysis; they thus allow users to ground the insights they've made at higher levels by turning to details in the underlying text.
- Insights are often about things that stick out: outliers can be more interesting than general trends.
- Those who work with text corpora need ways of creating explanations, both to help communicate insights but also to ground insights in more semantically useful language, including explanations of mathematical terms for those who do not have backgrounds in those fields.
- Questions can arise at all levels, so designers should afford multiple scales, multiple views, and multiple perspectives.

In this paper we will briefly explain each of these principles, how they arose from our work with humanities scholars, and why they are relevant to the broader text analytics problem.

THE LITERARY ANALYST

Literary scholars operate in an epistemic framework where there is often no ground truth to consult, existing algorithms are noisy, there are many answers to most questions, and arguments ultimately must be supported by consulting individual passages of text placed in larger contexts, contexts that are usually provided by other texts. Information visualization has recently begun to intervene in this space, with a special focus on visualizing patterns of word usage [3, 7]. These efforts (including our own) have emphasized the differences between this domain and the "general" text analytics problem, but our thesis is that these superficial differences are underlying many key similarities in structure, leading to a number of shared design principles.

Include Links To Text

Literary scholars ultimately argue *about* text, *using* text. Passages are taken as exempla of particular themes, quotations are placed in context of the text from which they came as well as related texts, and word choice is examined in detail. To make visualization tools that support this style of argumentation, our tools had to combine the ability of traditional visual displays to aggregate and display large amounts of data at different scales with the necessity of translating high level visual insights into the realm of particular sections or subsections of text. Figures 1 and 2 show one attempt at creating a tool suite with these abilities: while we still rely on standard abstractions in use in the wider field of text analytics (texts represented as high dimensional vectors, using dimensionality reduction to create useful spaces), our tool allows users to create their own definitions of importance (based on directions in the high level abstract space) and filter the entire corpus for the purpose of rapidly finding *passages* (rather than relationships, entities, or topics) of interest.

A focus on creating links to text still allows a great deal of flexibility in the design space. In our research we've investigated focus+context displays for viewing large-scale patterns in texts, and we've also looked at ways to aggregate

and disaggregate texts at semantically meaningful boundaries for more traditional visualization tools. Part of this effort will also be to create visual metaphors that work at multiple scales, and see when summarization must step in, and when the text "speaks for itself" and when it should be presented in its original form.

Sentiment analysis is an example where the original text as such (tweets, yelp reviews, &c.) is frequently given a short shrift: since the punchline of such sentiment analysis visualization is often an aggregate judgment: Do people like our product at time *t* or not? The actual artifacts of individual texts are often excluded from the central visual metaphors of the design. However, we would argue that the ability to return to the text is valuable even in these situations. As sentiment analysis incorporates more complexity from ML and NLP domains, users will need to be able to step into the loop to look at individual texts and make their own judgments without layers of algorithmically introduced structure and noise.

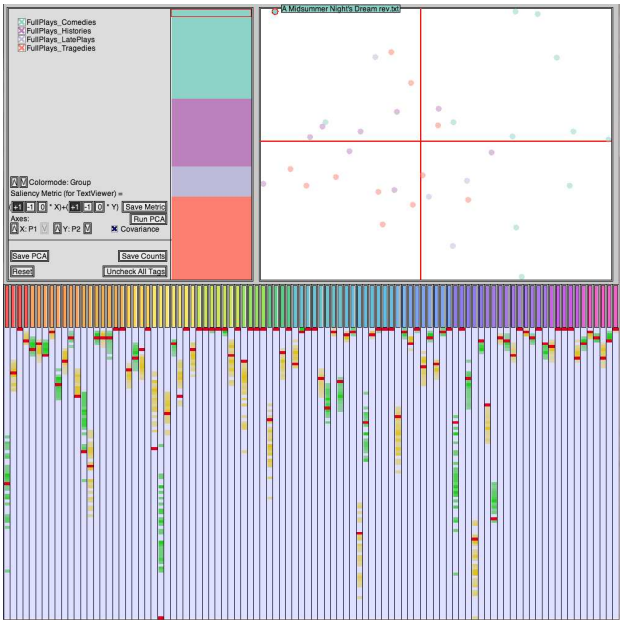

Figure 1: Overview of the Shakespeare corpus in our tool suite showing comedies (teal data points) being pushed to the right [6]. A high level pattern is visible, but there is no direct route to what this means in terms of low-level text.

Deal With Interesting Outliers

In many text analytics applications outliers represent either noisy data to be excluded or isolated (and so likely irrelevant) sections of a corpus. We expected our humanities scholars to take a similar view, especially given how many of the corpora they used were noisy or drawn from many divergent sources with variable standards of curation. Yet, examining both humanities scholars' language of proof and their language of discovery, outliers constituted the bulk of their targets of analysis (see Fig. 3). Outliers, especially in corpora large enough to make even a cursory reading by human beings of many texts impractical if not impossible, were targets of interest,

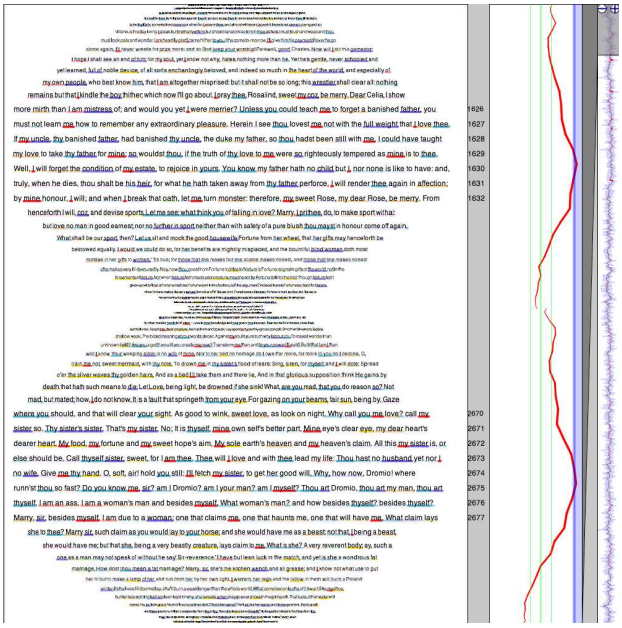

Figure 2: To create an explanation for this clustering seen in Fig. 1, the user dives into the actual text to important passages. Annotation of relative word importance lets users create plausible explanations grounded in their own expertise [6].

exemplars of particular patterns, or items that required explanation.

In our tools we are investigating ways to highlight noise in the data, particularly exploring encodings that afford the quick estimation of overall trend while allowing users to pick out outliers; we've found that non-traditional encodings can, for tasks like these, be effective and intuitive [4].

This lesson can be especially useful in the wider field of text analytics: in intelligence and law it is the outliers and exceptions that should be brought to the attention of the analyst. Analysis that relies on real-time streaming data (for the goal of de-noising and filtering what can be very noisy and very large corpora) will frequently rely on clusters as the unit of analysis [11], which can result in outliers being unduly removed from consideration. While noise should still be dealt with in text corpora, care should be taken not to eliminate all items in a corpora that break the trend: preserving them, even if not in the central visual display, makes it possible both to *report* on streaming data and to *interrogate* this data.

Provide Explanatory Tools

The feedback on our initial tools pointed out features our users needed: while our tools provided ways to cluster and aggregate large text corpora, it was difficult to see what cluster membership or non-membership meant at the level of text. To complicate matters further, the typical insights one would gain from a clustering of a text corpus (topics, shared entities, &c.) were not useful for our collaborators: they knew the standard groupings (genre, authorship, date) but wanted to see what effect these groupings had on the underlying text. Tools we created for them had to be adapted to provide visual ways

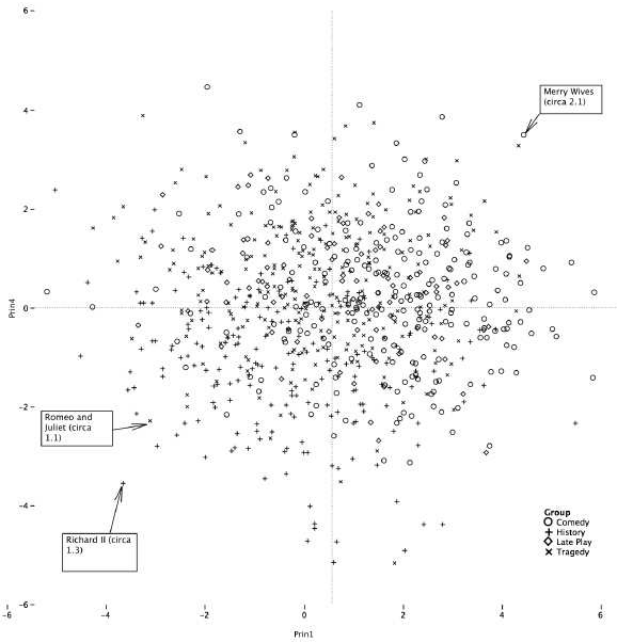

Figure 3: An example use of a visualization by our collaborators [9], embedding 1000 word "chunks" of Shakespeare's plays in a two dimensional space. Sections of plays that fall outside of their expected regions are more interesting than the large numbers of plays that follow the general trend, and so the outliers are the main target of later analysis.

of interrogating differences, and explaining surface-level features in terms of word-, sentence-, and paragraph-level details. Our collaborators were used to looking for interesting insights across texts, authors, and corpora, and "interesting" for them frequently meant atypical and unexpected (see Fig. 4).

To support this explanatory need, we are investigating ways to visualize not only what the patterns in the overall corpus are, but also how these patterns are viewed on the micro scale, and how to leverage the domain expertise of the user to steer large-scale visualizations qualitatively and quantitatively. Our research has begun to tackle this problem for the common text analytics workflow of embedding corpora in low dimensional spaces or small numbers of clusters. We have been tackling the problem from both ends: how can we make spaces that are semantically meaningful (*a priori* explanations) or, how can we create ways to visualize algorithmically (but not semantically) meaningful spaces in a way that affords semantic insights (*post hoc* explanations)?

Provide Multiple Perspectives

The Google n-grams dataset contains information about word occurrence and co-occurrence for hundreds of years of English print [10]. There are many ways to visualize these data but these ways are dependent on both the type of research and type of argumentation required of our humanities collaborators. What was surprising was that our collaborators were able to use multiple tools in concert (even repurposing tools originally intended for other purposes) to quickly learn to navigate the dataset from multiple perspectives (see Figs. 5 and 6). Insights gained from one perspective were explained with another, and patterns that could be intuited to exist in one tool were confirmed by exploration of the other. More than just the creation of linked views or tool suites, users were letting mental models guide their choice of tool rather than vice versa.

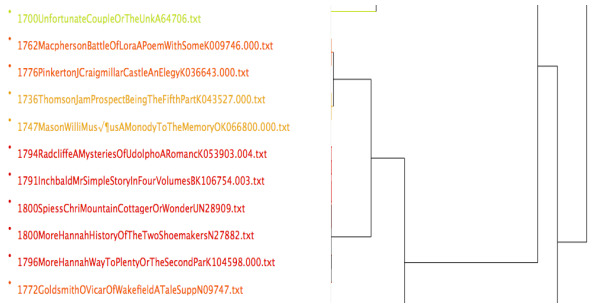

Figure 4: A section of a large dendrogram of of English renaissance plays, clustered by rhetorical similarity. Our collaborators, when finding areas where (for instance) an author's works were spread across multiple clusters, had to develop methods (and use new visualization tools) to explain the structure of the dendrogram: the dendrogram was good at showing the structure of the data per se, but did not afford ways of providing explanations.

We opened the problem up to our graduate level course in visualization: each group produced a visualization that was different from the others (the design space included, among others, radial pie charts, multiple linked views, and traditional scatterplots) yet this plurality of designs did not add harmful complexity but offered the potential for new insights.

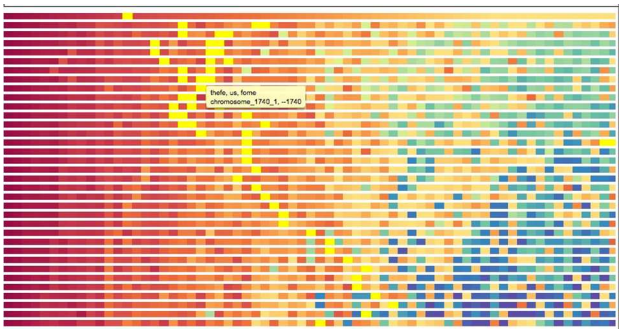

Figure 5: Our collaborators use a tool originally designed for genomics data to analyze patterns of word usage through time, and also find dirty data [1]. Here where OCR errors are rendering the early English "long s" as an "f" are shown to occur very frequently in the Google n-grams data set before disappearing entirely in the modern era.

CONCLUSION

We contend that the design principles we've learned from our work with the humanities have the form of general maxims for text analytics as a whole. We have already seen confluence in our work with applications made for humanities scholars providing insights to domains like genomics, virology,

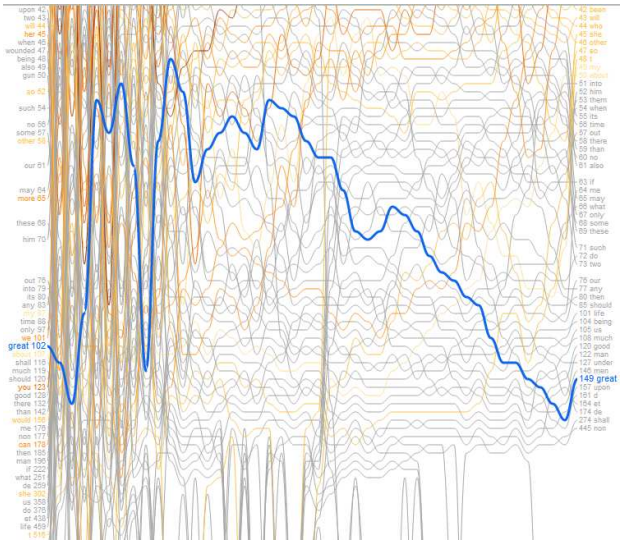

Figure 6: Another view of the Google n-grams dataset from a web viewer. Here the patterns are less clear, but the exact rank orders and positions are easier to see for individual words.

and proteomics and vice versa. That we have been able to broadly apply our tools lends credence to the position that we have stumbled on general practices rather than disciplinary idiosyncracies. Given that research on textual data is desirable in many domains, the ability to generalize from success in a domain that has experts who have devoted their working lives to studying texts and thus can use our tools and make use of the explanations generated from these tools is evidence that our methods can apply to a broad base and have a broad impact in text visualization and analytics.

ACKNOWLEDGMENTS

Our visualization work was supported by NSF awards IIS-0946598 and IIS-1162037. Our domain work, and the work of our humanities collaborators, was supported by a Mellon Foundation grant.

REFERENCES

1. Albers, D., Dewey, C., and Gleicher, M. Sequence surveyor: Scalable multiple sequence alignment overview visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (dec 2011), 2392 – 2401.
2. Chu, C. Literary critics at work and their information needs: A research-phases model. *Library & Information Science Research 21*, 2 (1999), 247–273.
3. Clement, T., Plaisant, C., and Vuillemot, R. The Story of One: Humanity scholarship with visualization and text analysis. *Relation 10*, 1.43 (2009), 8485.
4. Correll, M., Albers, D., Franconeri, S., and Gleicher, M. Comparing averages in time series data. In *Proceedings of ACM CHI* (2012).
5. Correll, M., and Gleicher, M. Poster: Understanding tagged text. In *IEEE Information Visualization Conference Poster Proceedings* (2010).
6. Correll, M., Witmore, M., and Gleicher, M. Exploring collections of tagged text for literary scholarship. *Computer Graphics Forum 30*, 3 (jun 2011), 731–740.
7. Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM (2007), 213–222.
8. Ellis, D. The English literature researcher in the age of the Internet. *Journal of Information Science 31*, 1 (Feb. 2005), 29–36.
9. Hope, J., and Witmore, M. The Hundredth Psalm to the Tune of" Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. *Shakespeare Quarterly 61*, 3 (2010), 357–390.
10. Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. Quantitative analysis of culture using millions of digitized books. *science 331*, 6014 (2011), 176.
11. Risch, J., Kao, A., Poteet, S., and Wu, Y. Text visualization for visual text analytics. *Visual Data Mining* (2008), 154–171.