A Sensemaking Environment for Literature Study

Aditi Muralidharan
Dept. of Computer Science
UC Berkeley
Berkeley, CA 94720 USA
aditi@cs.berkeley.edu

Marti A. Hearst
School of Information
UC Berkeley
Berkeley, CA 94720 USA
hearst@ischool.berkeley.edu

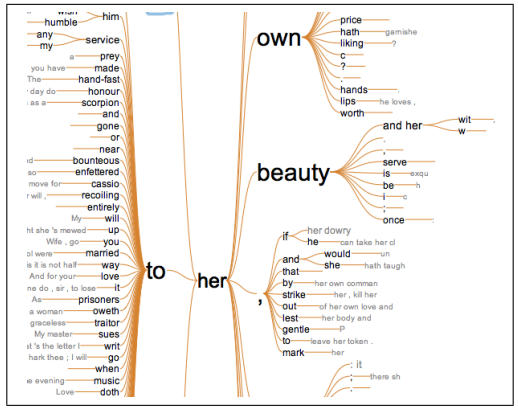


Figure 1: Word tree [17] for the word "her" generated by WordSeer on Shakespeare's complete works

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract

We present a sensemaking environment for literary text analysis. Literature study is a cycle of reading, interpretation, exploration, and understanding. While there is now abundant technological support for reading and interpreting literary text in new ways through text-processing algorithms, the other parts of the cycle – exploration and understanding – have been relatively neglected. Motivated by the literature on sensemaking, we are developing a software system that integrates tools for algorithmic processing of text with interaction techniques that support the interpretive, exploratory, and note-taking aspects of scholarship. At present, our project supports grammatical search and contextual similarity determination, visualization of patterns of word context, and examination and organization of the source material for comparison and hypothesis-building. This article illustrates its capabilities by analyzing language-use differences between male and female characters in Shakespeare's plays. We find that when love is a major plot point, the language Shakespeare uses to refer to women becomes more physical, and the language referring to men becomes more sentimental. Future work will incorporate additional sensemaking tools to aid comparison, exploration, grouping, and pattern recognition.

Keywords
Sensemaking; User interfaces; Visualization; Text Analysis; Information Retrieval

ACM Classification Keywords
H.5.2 [**Information interfaces and presentation**]: User Interfaces.

General Terms
Human-Computer Interaction, Information Retrieval

Introduction
To date, text analysis systems for humanities scholars [5, 10, 3, 11, 16, 12] have focused on aiding interpretation. First, they apply some form of natural language processing to extract aggregate statistics about word usage, topics, named entities, and parts of speech. Second, they display the extracted information with visualizations like word clouds, node-and-link diagrams, and lists of word contexts. Such systems make patterns of style, form, and theme visible, and interpretable by people.

However, literature study is a form of sensemaking [13]: a cycle of reading, interpretation, exploration and understanding. As useful as they are, current digital humanities text analysis systems leave the exploration and understanding part of the cycle unsupported.

According to Hearst [7], a tool that supports sensemaking would give an overview of the contents of the collection, help the user keep track of what they had already seen, suggest what to look for next, encourage the user to try new queries, find documents similar to those already found, and allow for aliasing of terms and concepts.

The WordSeer project (http://wordseer.berkeley.edu), is our effort to create a sensemaking environment for literature and language study. Like other systems for the humanities, it has search and visualization capabilities, but it also supports sensemaking activities like collecting and reorganizing information, exploring related words, and annotating and tagging items. At present, it is being used by 3 groups of literature scholars to analyze the North American slave narratives, the works of Stephen Crane, and the complete works of William Shakespeare.

In this paper, we demonstrate the software's current capabilities by using it to explore the following open-ended question:

> "How does the portrayal of men and women in Shakespeare's plays change under different circumstances?"

As one answer, we show how the tool suggests that when love is a major plot point, the language referring to women changes to become more physical, and the language referring to men becomes more sentimental.

System Description
The tool is run on a collection of documents. The input is a set of XML files in a directory, each representing a document in the collection, and the output is a web application with search, visualization, and annotation capabilities. We chose XML because TEI [15], an XML specification for encoding documents, is a widely-adopted digitization standard in the humanities. Many documents of interest to literature scholars are encoded as TEI-XML files.

Grammatical Search
We began our analysis with the question, "what are some things that are portrayed as 'his' and some things that are

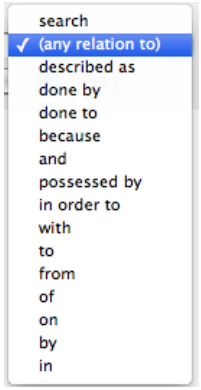


Figure 2: Searchable grammatical relationships between words.

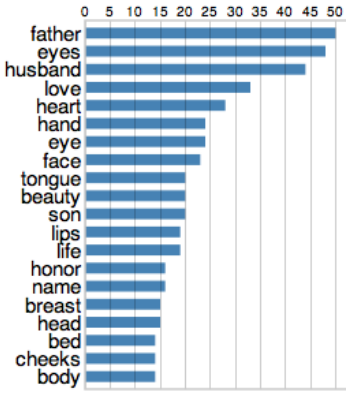


Figure 3: Results for the grammatical search *possessed-by* `her`. The bar graph shows the 20 most-frequent matches along with their frequencies.

'hers'?. A typical keyword search returns an unstructured lists of results. A standard approach in literature study is to view search results in a *concordance*: a list of all the sentences in which a word occurs, with the target word aligned in the center of the view, exposing the contexts to its left and right, sorted in some manner. Our tool uses the word tree visualization [17] which makes common contexts in a concordance easier to view by grouping them in an arced tree-like structure. The word tree for `her` is shown in Figure 1. Some words like `beauty` stand out, but constructions like `her own` muddy the picture.

The word `his` is always a possessive pronoun, so word sequences containing `his` would nearly always be relevant. But `her` can also be a 3rd-person pronoun, and will yield constructions like "I told her that X" and "I gave her the Y". With the WordSeer project, we make headway on this problem by providing an easy interface to view the results of *grammatical search* (Figure 4). The system uses natural language processing (NLP) to extract relationships between words (see [8] for an overview), and allows users to specify *both* keywords *and* relationships between them. This mode of searching is related to information extraction systems such as TextRunner [1] and Bindings Engine [2]. These system also use NLP to allow queries for relationships between words. In the tool's search interface, pairs of words are specified using input boxes, and the relationship between them is selected from a drop-down menu (Figure 2). Leaving a word-input box blank returns all matches.

With this feature, we can take advantage of the fact that possessive relationships between words can be automatically detected, to express our question precisely: "what are all the words with which `her` has a possessive relationship".

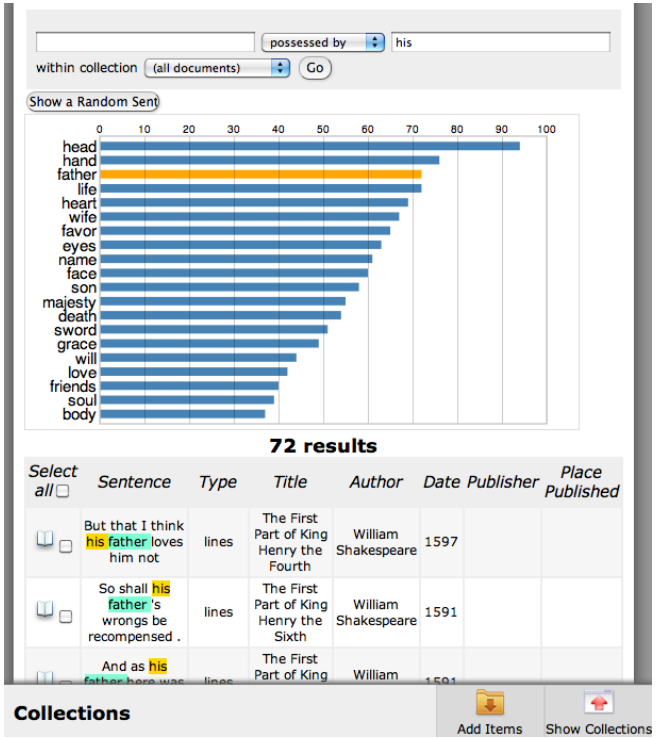


Figure 4: Grammatical search results for *possessed-by* `his`, filtered on the word `father`. Individual search results, corresponding to matching sentences, are highlighted to show the words in the relationship

Figure 4 shows search results for all words for which `his` has a possessive relationship. Comparing these words with those for `her` (Figure 3) reveals immediate differences. The word `father` is most common for `her`, with `husband`, and `son` close behind. Several body parts enter the picture: `eyes`, `hand`, `face`, `tongue`, `lips`, `cheek`. A picture emerges: women's most commonly-mentioned possessions are their male relatives and their bodies.

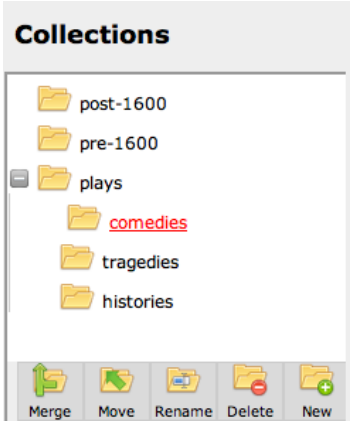


Figure 5: We initially created the above collections of documents. Collections can also contain words, sentences, and "snippets" of text .

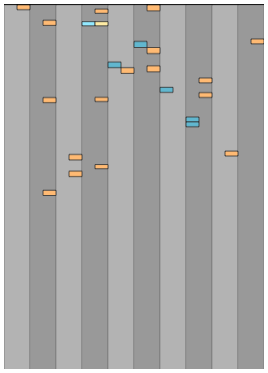


Figure 6: Visualization of the *histories* collection comparing the prevalence of body parts *possessed-by* `her` (blue) and relatives `possessed-by her` (orange).

Visualization, Reading, and Hypothesis-Generation
Our next question was whether this physical, male-dominated picture of women was consistent, or whether it changed in different types of plays. We used the tool's collections feature (Figure 5) to divide the plays into *comedies*, *tragedies*, and *histories* – the three most commonly-accepted categorizations of Shakespeare's plays. We also created *pre-1600*, and *post-1600* categories to check whether there were temporal differences. Collections were created using the "collections" bay, a collapsible window at the bottom of the screen. We added the appropriate plays through the document listing (sortable and filterable by date, title, full-text search, grammatical search, and length).

We used the tool's *newspaper-strip* visualization [4, 6] (Figure 7) to compare the prevalence of the two categories of words in different types of plays. Each play is represented as a long column. Within each column, small, colored horizontal blocks (corresponding to 10 sentences each) highlight the presence of a match.

The results for the *tragedies* collection were similar to the results for *comedies* (Figure 7) but in *histories* (Figure 6), an interesting pattern emerged. It seemed that body parts were somewhat less prevalent in these plays, but family remained unchanged.

Hovering over a few body-part results quickly led to a new hypothesis. In our rough sample, many of the mentions sounded romantic. We used the reading and annotating interface to follow up on this by clicking on the highlighted blocks in the newspaper-column visualization. We selected the speeches referring to body parts and tagged them by the topics they seemed to contain (Figure 8). It soon became apparent that many of the mentions were speeches by a lover.

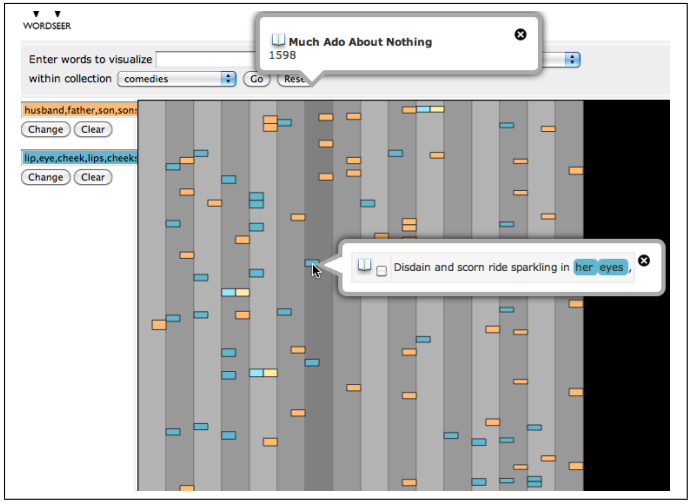


Figure 7: Comparing the prevalence of body parts *possessed-by* `her` (eyes, lips, cheeks, and face)(blue) and relatives *possessed-by* `her` (husband, father, sons, daughters, children) (orange) in the comedies. Each column is a comedy, represented in alternating shades of grey. Hovering over a column (e.g. "Much Ado About Nothing" above) darkens it and displays the title. Hovering over a highlighted block displays the matching sentence.

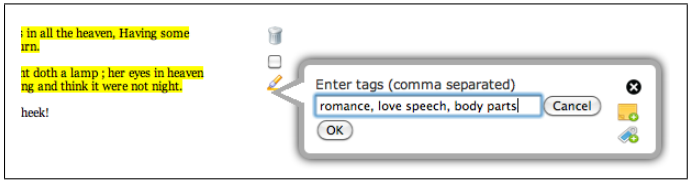


Figure 8: Highlighting text creates a "snippet", to which tags and notes can be attached.'

Related words (commonly used in similar contexts [9], or commonly used within a 10-sentence window) can be

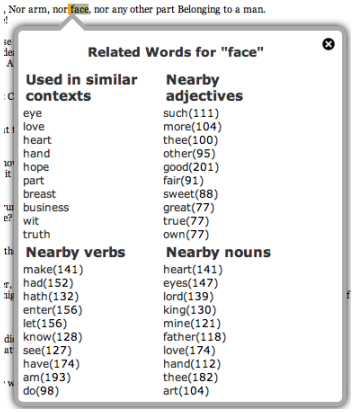


Figure 9: Related words for `face`

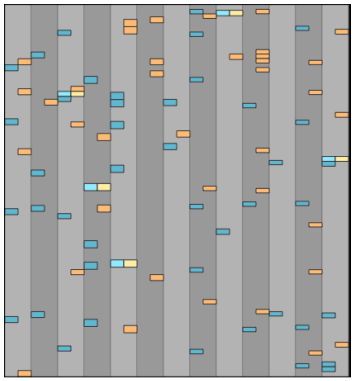


Figure 10: Visualization of the *love-stories* collection comparing the prevalence of body parts possessed-by `her` (blue) and relatives *possessed-by* `her` (orange).

viewed by right-clicking words in the reading interface. Our hypothesis was strengthened when we examined the related words for body-parts (Figure 9). Other body parts were frequently mentioned, along with `love`, `fair`, and `sweet`.

We created a final pair of categories: *not-love-stories* for plays in which love is not a major plot point, and *love-stories* for plays in which it is. When we reorganize the plays along these lines, the results are immediate.

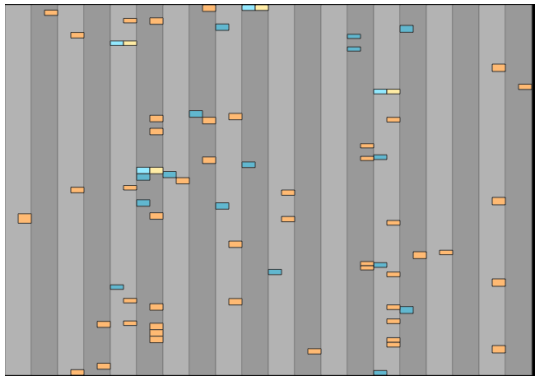
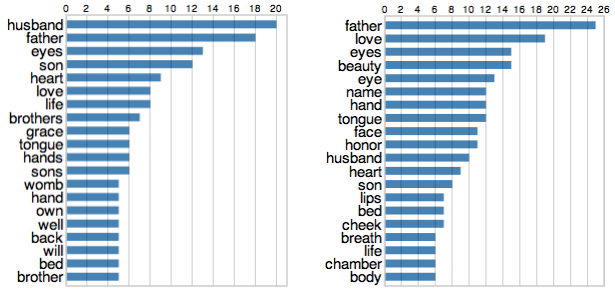


Figure 11: Visualization of the *not-love-stories* collection comparing the prevalence of body parts possessed-by `her` (blue) and relatives *possessed-by* `her` (orange).

In the *love-stories* (Figure 10) collection, we see *both* body parts *and* male relatives. By contrast, the *not-love-stories* visualization (Figure 11) shows *predominantly* male relatives, and hovering over the occurrences of body parts reveals a gloomy picture of `her` tear-stained cheeks and `her` sorrowful eyes.

The grammatical search results agree with the newspaper-strip visualizations and related words. We see more physical attributes *possessed-by* `her` in the in the *love* collection than in the *not-love* collection (Figure 12).

The grammatical search results show that the language around men changes as well (Figure 13). In the *not-love case*, the only woman to appear is `mother`, at number 20, but in the love case, `wife` takes first place, followed by `favor`. Compared to the physical language for women, these words have a more sentimental quality.

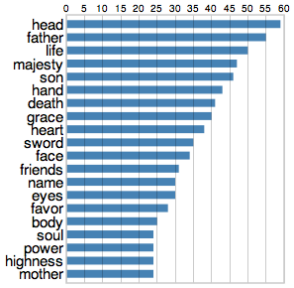


(a) The not-love plays (b) The love plays

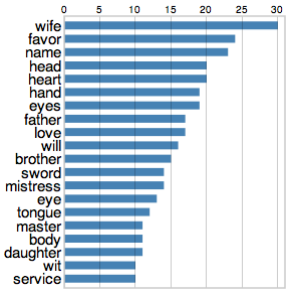
Figure 12: Comparison of grammatical search results for possessed-by `her`

Future Work

Our tool will be improved and evaluated through extensive case studies of real questions with our literature-scholar collaborators. These, along with the exploration, collection, and comparison features are the subjects of ongoing work. Our collaborators say that when studying patterns of imagery, theme, and rhetoric, it would be useful to see groups of related words, and other passages expressing similar concepts. A number of NLP algorithms for calculating text similarity and grouping words are available. We plan to incorporate these into the system with relevance feedback [14], so users can refine the matches. Second, in order to compare visualizations and



(a) The not-love plays



(b) The love plays

Figure 13: Comparison of grammatical search results for possessed-by `his`.

grammatical search results, users currently have to switch between collections using a drop-down menu. A side-by-side interface would be less cumbersome. It would also be useful to highlight differences in frequency and distribution between data represented with the various visualizations.

Acknowledgements

This work was supported by NEH Digital Humanities start-up grant number HD-51244-11.

References

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. *Proc. IJCAI*, 2007.

[2] M. J. Cafarella and O. Etzioni. A search engine for natural language applications. In *Proc. WWW*, WWW '05. ACM, 2005.

[3] T. E. Clement. A thing not beginning and not ending: using digital tools to distant-read gertrude stein's the making of americans. *Literary and Linguistic Computing*, 23(3):361, 2008.

[4] S. G. Eick. Graphically displaying text. *Computational and Graphical Statistics*, 3(2):127–142, 1994.

[5] J. Fekete and N. Dufournaud. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In *Proc. ACM Digital libraries*, DL '00. ACM, 2000. ACM ID: 336632.

[6] M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *Proc. ACM SIGCHI*, 1995.

[7] M. A. Hearst. Supporting the search process. In *Search User Interfaces*. Cambridge University Press, 1 edition, Sept. 2009.

[8] D. Jurafsky and J. H. Martin. Chapter 13 syntactic parsing. In *Speech and language processing*, pages 427 — 459. Pearson Prentice Hall, 2nd ed. edition, 2009.

[9] D. Lin. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proc. ACL*, 1997.

[10] X. Llora, B. Acs, L. S. Auvil, B. Capitanu, M. E. Welge, and D. E. Goldberg. Meandre: Semantic-driven data-intensive flows in the clouds. In *Proc. IEEE eScience*, pages 238–245, 2008.

[11] C. Plaisant, J. Rose, B. Yu, L. Auvil, M. G. Kirschenbaum, M. N. Smith, T. Clement, and G. Lord. Exploring erotics in emily dickinson's correspondence with text mining and visual interfaces. In *Proc. ACM Digital Libraries*, pages 141–150, Chapel Hill, NC, USA, 2006. ACM.

[12] G. Rockwell, S. G. Sinclair, S. Ruecker, and P. Organisciak. Ubiquitous text analysis. *Poetess Archive Journal*, 2(1), 2010.

[13] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proc. INTERACT and CHI*, CHI '93. ACM, 1993.

[14] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.

[15] e. TEI Consortium. Guidelines for electronic text encoding and interchange. http://www.tei-c.org/P5.

[16] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near Martha? exploring name entities in literary text collections. In *IEEE VAST*, pages 107–114, 2009.

[17] M. Wattenberg and F. B. Viegas. The word tree, an interactive visual concordance. *IEEE Visualization and Computer Graphics*, 14(6):1221–1228, 2008.