More and more source text in the humanities gets digitized every day, making it accessible to large scale computational analysis. By contrast, traditional methods of humanistic analysis are based on detailed arguments built upon close readings of individual texts. How will the field adapt? How do we use statistics and text mining to answer humanistic questions?

**Text Analysis in the Humanities**

To date, text analysis systems for humanities scholars 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, literary study is a form of sensemaking: a cycle of reading, interpretation, explanation and understanding. As useful as they are, current digital humanities text analysis systems leave the exploration and understanding part of the cycle unsupported.

**A Sensemaking Challenge**

Past studies of sensemaking have focused on decision-making or intelligence analysis tasks. In these domains, the objects of analysis are explicit and clearly defined: people, events, facts, relationships, and numbers. In literature study, by contrast, the objects of study are style, themes, imagery, form and stereotypes. How can text mining and visualization help literature scholars analyze these nebulous kinds of information?

**Related Projects**

The MONK project at CMU, and the Voyeur and TaPor project at McMaster University share the same cause as WordSeer. They use basic visualization and language processing to help literature scholars explore language use. The WordHoard program uses detailed metadata to give users tables of word frequencies in different subsets of text. The Prism project at the University of Virginia’s Scholar’s lab is a tool for “crowdsourcing interpretation”. It visualizes the results of many users’ interpretations of a text.

---

**Digital Humanities**

“How does the portrayal of men and women in Shakespeare change in different circumstances?”

We demonstrate WordSeer’s capabilities by using it to explore this open-ended question. We find that when love is a major plot point, the language around women becomes more physical, for men, more sentimental.

**Grammatical Search**

What are some things that are portrayed as ‘his’ and some things that are ‘hers’?

Ordinary keyword search would return a list of sentence matches. 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 WordSeer, we make headway on this problem with grammatical search.

**Visualisation**

We use WordSeer’s newspaper-strip visualization to examine the prevalence of “her body parts” and “her male relatives” in the comedies, tragedies, and histories.

**Exploration**

Womens’ body parts are mostly mentioned in romantic contexts.

We created a final pair of categories, “love plays”, in which love was a major plot point, and “not love plays”.

---

WORDSEER.BERKELEY.EDU

A Sensemaking Environment for Literary Text