WordSeer: Exploring Language Use in Literary Text

Aditi Muralidharan
UC Berkeley
420 Soda Hall.
Berkeley, CA 94701 USA
aditi@cs.berkeley.edu

Abstract
Increasing numbers of primary and secondary source texts in the humanities have been digitized in recent years. Scholars who want to study these new collections in depth need computational assistance because of their large scale. The non-programmer tools for text analysis currently available operate at the word level, and they show tables of counts and lists of occurrences, but rarely interactive visualizations. We have built WordSeer, a text analysis tool that includes visualizations and works on the grammatical structure of text extracted using highly accurate technologies from computational linguistics. We have focused on the task of exploring language use patterns in a collection of American slave narratives, but the technique is applicable to any text collection. Our user studies with humanities scholars show that WordSeer makes it easier for them to translate their questions into queries and find answers to their questions compared to a standard word-based interface.

Keywords
Digital humanities, search user interfaces
Introduction
The increasing prevalence of digitized source material in the humanities has led to uncertainty about how this suddenly available information will change scholars’ research methods. Our focus is specifically on text collections: comparing texts, and identifying and tracing patterns of language use. These tasks are not widely supported by any current software, but if humanities researchers want to use digitized text collections on a larger scale, they will need to do exactly such things.

We restrict ourselves to a particular collection: the North American antebellum slave narratives, written by fugitive slaves in the decades before the Civil War with the support of abolitionist sponsors. This project aims to assist literary scholars in exploring patterns of language use in these narratives with computational techniques. Our goal is for our English scholars to be able to use our system to gather accurate information about language use patterns in manner satisfying to them.

Our project is innovative because of the advanced computational language processing that brings to the service of humanistic analysis. It is the first system to allow interactive visual comparison of texts along grammatical dimensions.

Related Work
In the digital humanities, there are two close text analytics efforts. The first is the MONK project [1] incorporating the SEASR analysis toolkits [2]. These projects offer two computational linguistics tools: tagging words with their parts of speech and extracting named entities. Users can visualize occurrence patterns of word sequences within a chosen text, and plot networks of how often named entities occur near each other. These tools were used to produce several visual text-mining analyses of literature [3][4][5]. The second is Voyeur [6]. Voyeur operates entirely at the word level. It allows users to plot word frequencies, see concordances (contexts in which words occur) and create tag clouds.

Other digital humanities projects have used advanced language processing, but have not developed user interfaces or visualizations. These projects include topic modeling applied to 19th Century British and American novels [7], and to the Scandinavian folklore [8]. In addition, social networks were automatically extracted from 19th century novels [9].

From human-computer interaction, we are informed by general principles of search user interface design described by Hearst [10], and of visual exploration of large data sets described by Shneiderman [11]. Visually representing text as a block of tiles goes as far back as 1995 [12]. Interfaces for exploring the grammatical relationships between words are a newer area of research, we are only aware of Phrase Nets, by van Ham, Wattenberg and Viegas [13].
System Description

Computation
The goal of the computation is to extract grammatical structure from text. For example, the sentence, “Marsupial mammals have pouches.” contains the meaningful unit “marsupial mammals” that functions as a noun and is therefore a noun-phrase. It further contains the word “marsupial” which is an adjective modifier of the word “mammals”, which is a noun. Algorithms called parsers in computational linguistics extract such information accurately. We use the current standard, the Stanford Parser [14], which processes 4-5 sentences of length 30 a second. The complete list of inter-word dependency relationships we use is described in the Stanford Parser’s dependencies manual [15]. Some intuitively familiar relationships include adjective- and adverb-modifier, verb-subject, verb-object, conjunction, and various prepositions.

Interface
We walk through our interface with an example use case.

Suppose a scholar is interested in how the words “true” and “authentic” are used in the narratives. As a first step, she can search for sentences in which “true” or “authentic” are used to describe something. Figure 1 shows the main search boxes. There is an ordinary keyword search form on the left, but the grammatical search form allows the scholar to exactly express her query. She chooses the is-described-as relation, enters the terms “true, authentic” into the right-hand-side box, and leaves the other empty.

Figure 1 Searching for all things described as true or authentic.

Figure 2 Search results along with a visualization of the frequencies of different words described as “true” or “authentic”. The scholar has filtered on “authentic”. Her query returns a result page Figure 2. In addition to a list of results, the interface visualizes of the frequencies of all the words that match the search, giving her an instant picture of the entire result set. From the graphs, the scholar notices that “authentic” is much less frequent than “true”, and clicks the bar to filter the results, leaving only sentences in which “authentic” appears. Scanning this list, the scholar quickly realizes that “authentic” is used to describe sources of information, and decides to focus on the word “true”.
Figure 3 Reading narratives. The narratives are automatically scrolled to the sentences shown in the search results earlier, which are also automatically highlighted for convenience. She filters the results by clicking on “true”, and selects a few narratives from the search results to read in detail, by clicking the View button.

Clicking the View button takes the scholar to the reading screen, shown in Figure 3, where two narratives are opened up for closer inspection. She sees that “colored citizens” are described as being inspired by “the good and true”, and wants to know what else is said about them. She highlights the words in the text and clicks the magnifying glass icon. This takes her to a pattern view screen, Figure 4, which shows her all the patterns in the section of text that she highlighted. She selects the patterns corresponding to “colored citizens” and clicks “See in text”.

Figure 4 The textual and grammatical patterns that were automatically detected in the section of highlighted text.

Figure 5 Sidebars showing where “colored citizens” occurs. A comparison-visualization appears, shown in Figure 5. We represent each document as a stack of small bars, each corresponding to a section of text. A colored
bar indicates that the feature of interest occurs in that section, and clicking on any bar takes the scholar to the corresponding section. The scholar can immediately spot repetition and see patterns of distribution. For example, the second narrative does not contain this phrase at all, but that the first contains it several times.

**Evaluation**

*Pilot*

Our participants were a professor and graduate students in the department of English and we discussed a functional prototype. Early prototypes included Phrase Nets [13], but feedback indicated that the scholars preferred simple graphs. During these pilot sessions, we decided on representative tasks for our user study.

*Study*

We recruited 5 graduate students from the departments of English and History at UC Berkeley. In a 45-minute study, the participants were first shown a walk-through of the interface, and a standard keyword-based interface. Then, they were given three tasks to be done on each interface while thinking aloud, one easy, one medium, and one hard. Together, the tasks built towards *typifying* a pre-selected type of event in the narratives. The easy task was to collect a set of narratives that included the event, the medium task was to identify and select patterns of interest in the texts relevant to the question, and the hard task was to answer a summary question. The questions were, "give a list of 5 adjectives that characterize slave punishment", and "describe 5 places to which slaves escaped". We gave each user both questions, but alternated the interfaces on which users did the tasks.

![Figure 6 Answers to Likert-scale questions.](image)

**Figure 6** Answers to Likert-scale questions.

We compared the users’ *effectiveness, efficiency, and satisfaction* [10] on the two interfaces. We asked them Likert-scale questions about how easy it was for them to formulate their query, how easy it was to determine the relevance of results, and how satisfied they were with the accuracy and completeness of their results. The results are shown in **Figure 6**: WordSeer made it *significantly easier* for users to formulate their queries (p = 0.0104), and to answer the question (p = 0.0097). It was also easier to determine the relevance of search results, with p = 0.067. Perhaps a larger user study would be able to establish significance. The remaining dimensions were also positive for WordSeer, but not significant.

**Acknowledgements**

Many thanks to Marti Hearst, Bjoern Hartmann, Bryan Wagner, and Rachelle Annechino.
Citations


