WordSeer: Exploring Language Use in Literary Text

Aditi Muralidharan
Department of Computer Science
UC Berkeley
aditi@cs.berkeley.edu

Marti Hearst
School of Information
UC Berkeley
hearst@ischool.berkeley.edu

ABSTRACT
Increasing numbers of primary and secondary source texts in the humanities have been digitized in recent years. Humanities scholars who want to study these new collections in depth need computational assistance because of their large scale. We have built WordSeer, a text analysis tool that includes visualizations and works on the grammatical structure of text extracted using highly accurate off-the shelf natural language processing tools. We have focused on the task of exploring language use patterns in a collection of North American slave narratives, but the technique is applicable to any text collection. Our preliminary 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 keyword-based search interface. In this paper, we present the system currently under development and describe text analysis features we plan to include in the next iteration.

Keywords
Sensemaking, exploratory search, search user interfaces, digital humanities, text analysis, visualization, visual analytics.

1. 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. What balance will scholars strike between in-depth examination of a few sources, and a more "distant reading" [20] of a large number of them? 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.

In collaboration with English scholars, we have built WordSeer (http://bebop.berkeley.edu/wordseer), a visual analytics [14] system that allows scholars to compare documents' grammatical features, and analyze the distribution of textual patterns throughout an entire collection. Our goal is for humanities scholars to be able to use our system to gather accurate information about language use patterns in a way that is intuitive and natural to them.

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. Scholars agree about the slave narrative's most basic conventions but it is likely that these narratives, with their extreme repetitiveness, may also manifest other regular features that have yet to be detected by scholars. This project aims to assist literary scholars in uncovering these patterns with computational techniques.

In the course of our collaboration with humanities scholars, we have learned that humanistic analysis of text collections tends to take the form of a scholar coming up with a number of vague hypotheses or questions, and then looking through a text collection for evidence to support, disprove, or characterize them. This is a type of sensemaking process, as described by Russell [24] because the scholar is not always sure what kinds of information exist in the collection, what evidence will ultimately be needed, what form it will take, or where to find it.

In this paper, we present the system currently under development, give results of our preliminary evaluations, and describe the next iteration of the system, which will include features for finding similar texts, and for characterizing a given collection of text snippets along grammatical, narrative, and entity-based dimensions

2. RELATED WORK
The closest work to our project comes from other text-focused visual analytics efforts in the field of digital humanities. Tools developed in this field usually have two parts. First, they apply some form of natural language processing to extract aggregate statistics about word usage, named entities, and parts of speech. Second, they display the extracted information with visualizations such as word clouds, node-and-link diagrams, and concordances.

One of the first visual analytics interfaces for humanities text was Compus [5], which allowed users to visually search and explore a collection of XML-encoded 16th-centry legal documents. At present, there are three well-known analytics efforts focused around text collections in the digital humanities. The first is the MONK project (http://monkproject.org) incorporating the SEASR analysis toolkit [18]. These projects offers two computational linguistics tools in addition to word distribution and frequency statistics: 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. This research led to visual text-mining analyses of Emily Dickinson's correspondence [21], and of Gertrude Stein's "The Making of Americans" [2] and an interface for exploring the parts of speech used near query words of interest [29]. The second is Voyeur [23], which operates entirely at the word level. It allows users to plot word frequencies, see concordances (contexts in which words occur) and create tag clouds. Third, at a massive scale, though still at the word level, is the analysis by Michel et. al. [19] of millions of books in the Google books collection. This work also resulted in the Google books n-gram viewer (http://ngrams.googlelabs.com/).

Other projects have used more advanced language processing, but have not developed them into user interfaces or combined them with visualizations. Topic modeling is being applied to 19th Century British and American novels [12]. These novels were also

HCIR '11, October 20th, 2011, Mountain View, CA USA.

the subjects of recent research that showed how to automatically extract social networks from free text [4].

Outside the digital humanities, we are informed by the field of visual analytics for text collections. Sandbox system [31] and its companion information retrieval system, TRIST [13], addressed hypothesis generation and evidence gathering. It let users see how a certain concept had been reasoned about in the past and provided templates that the users could fill in with evidence relevant their particular case. Next, the problem of tracking, highlighting, and comparing relevant entities in a collection was tackled by Jigsaw [8, 26],which Goerg. et. al. extended to include text similarity, clustering, summarization and sentiment analysis [9]. For a more complete history of efforts to create text analysis tools, see Hearst 2009, Ch. 11 [10].

3. SYSTEM DESCRIPTION

WordSeer is built around the text collection of interest to our literary scholar collaborators: the North American antebellum slave narratives. This is a collection of 3,000 autobiographical accounts written by fugitive slaves in the decades before the Civil War with the support of abolitionist sponsors. Our current experimental version only includes about 120 narratives, but not only will the next iteration be collection-independent, we will also use it the full set of 3,000.

WordSeer has four components: search, analysis, reading, and annotation. Browsing is notably absent in the present iteration, but future iterations will automatically generate a faceted search and browsing experience using the Castanet algorithm, developed by Stoica et. al. 2007 [27].

3.1 Search

The text is broken up and indexed by document, sentence, paragraph, and even on the word level. This design allows us to compute frequencies and co-occurrences information at every level of granularity. We also have full-text search ability.

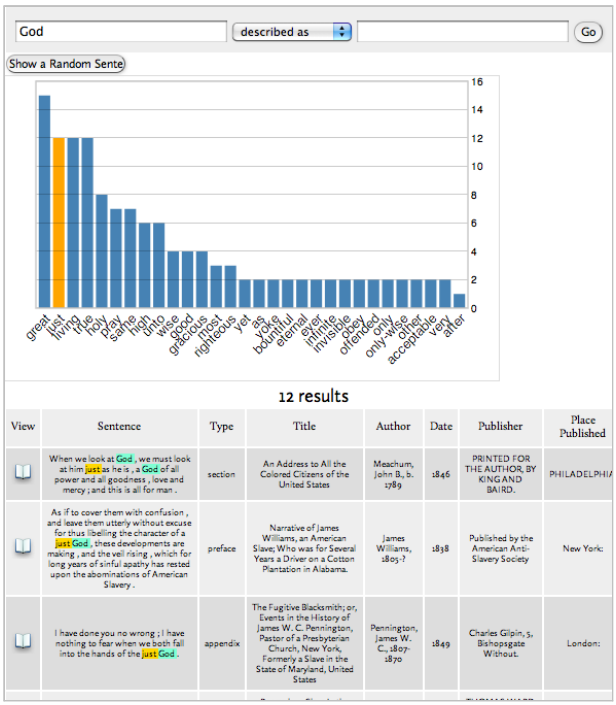

Figure 1: The adjectives describing 'God' filtered on the word 'just'.

3.1.1 *Grammatical search*

We have applied natural language processing in the form of syntactic and dependency parsing [15] to allow users to search over grammatical relationships between words, with a bar graph summarizing the matched grammatical relationships.

Grammatical search allows users to be precise about the relationships between query words. For example, in (Figure 1), instead of just typing in "God kind" to retrieve sentences in which "God" is described as "kind", a user directly specifies that the *amod* (adjective modifier) relationship should exist between "God" and "kind". The query "God *described as* ___", would give the results shown. All the adjectives that ever describe "God" are displayed in an interactive graph. Clicking on an adjective filters the result set. In the figure, the results have been filtered on the word "just".

By allowing users to search through the grammatical "neighborhoods" of words, this feature allows them to see other words that are used in meaningful relationships with query words. The information we extract could previously only be learned from reading. Now, users can make a quick assessment of the contexts in which a word is used, and decide whether or not a hypothesis will bear further investigation.

3.1.2 *Evaluation*

We conducted a pilot study of the grammatical search interface in which we recruited 5 graduate students from the departments of English and History at UC Berkeley. After a walk-through, they were given 3 tasks to be done on WordSeer, and 3 tasks on a non-grammatical keyword-search 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 results were encouraging: participants reported that WordSeer made it easier to formulate queries ($p = 0.01$), and to answer the question ($p = 0.01$).

3.2 Analysis

Humanities researchers are often interested in understanding the prevalence, emotions, and connotations associated with word usage [2, 21, 29].

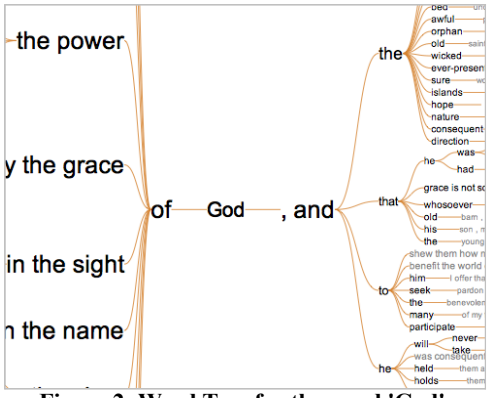

Figure 2: Word Tree for the word 'God'.

3.2.1 *Word Usage Patterns*

WordSeer allows exploration of the contexts surrounding word-usage on two levels. The first is at the sentence level, with the *word tree* visualization (Figure 2) introduced by Wattenberg and Viegas '08 [30]. Word trees show the same information as a traditional concordance, but contexts that begin with the same

word are grouped into a suffix tree. It is easy to scan, and can give an instant picture of the way a word is used in a collection.

The second is at the collection-wide level. A popular visual metaphor for this is one of text as a long newspaper column [2, 3, 5, 11]. WordSeer uses this visual metaphor to show how a query word or phrase is distributed through the collection (Figure 3). Each column is a narrative, and each bar is a group of 30 sentences. The bars are highlighted if the word occurs in that group of sentences. Clicking on a bar brings up the sentence, with an icon to visit the occurrence in the original text.

Figure 3: Collection-wide occurrence of 2 queries, 'blessing' (yellow), and 'faith' (blue).

3.2.2 Collection-specific Vocabulary

The vocabulary problem [7] refers to the great variety of words with which concepts are expressed in a collection. By suggesting other words that seem to behave similarly to words in a query (Error: Reference source not found), we aim to give the user a

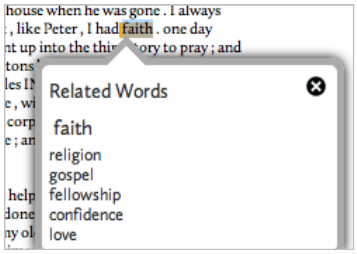

Figure 4: Related words for 'faith'.

sense of the vocabulary of the collection. This is query expansion adapted to our type of investigation, in which users are interested in the behaviors of specific words.

We have computed distributional similarity scores between words using dependency relationships as context [16]. These similarities calculate other words that are used in similar contexts, and might therefore have similar meanings.

3.3 Reading and annotation

Users can view a list of all the documents in the collection and select one to read in detail. While reading, we support note-making and tagging. Users can select and tag selections of text (with auto-suggest for consistency) and attach notes to their selections. Further, they can review all text tagged with a tag, view all their notes, and edit and delete notes and tags. In the future, we want to explore a canvas-like interface, where a user can simultaneously organize and manipulate all the collected snippets under all the tags.

4. FUTURE WORK

WordSeer lacks two much-needed text analysis features that humanities users require. The first is exploration based on interesting examples, and the second is entity characterization and exploration. More generally, Wordseer needs to support finding patterns around people, language use, and events and tracing similarities throughout a text collection.

4.1 Example-Based Exploration

There is currently no way to explore the collection based on a set of passages known to be interesting. Our scholars need to find similar snippets from other texts in the collection, characterize collected snippets along grammatical dimensions, find other types of events and topics that seem to accompany the snippets of interest. In addition, they need to be able to refine our automated suggestions by giving feedback. At present, we only have a very bare approximation to this, in word-to-word similarity.

We plan to borrow an approach from information retrieval to solve this problem. First, we will compute similar passages to any selected portion of text using a vector space model with bag-of-words features to characterize paragraphs. We will augment these features with the grammatical information we already compute, and others used in NLP applications like entailment detection, evaluation of machine translations, and summarization. These will be our initial suggestions, presented to the user for feedback as a thumbs-up/thumbs-down list or a choose-between-two-alternatives sequence. Using either active learning [28] or relevance feedback [22, 25], we will refine our suggestions and repeat the process.

4.2 Characterizing and Exploring Entities

Users say it would be useful to be able to track a particular person or set of people through a text, seeing the events in which they participate, what they say in various places, and they ways in which they are described. It would also be useful to have a way to find similar entities, and see their general characteristics and prevalence in the collection.

There have been efforts from the summarization literature to create timelines and story lines around particular queries or news topics [1, 17], but tracking a *person* through a single text would require accurate named entity recognition and pronoun resolution. Further, the literature on finding related entities seems to be sparse. Filippova & Strube '07 [6] clustered entities into related groups in the context of evaluating the coherence of automatically generated text, but measured "relatedness" using a Wikipedia-based metric, inapplicable to our situation.

Another sensemaking system that has tackled this problem is Jigsaw [8], with excellent visual aids for tracking and connecting entities. Nevertheless, Jigsaw used named entity recognition, but no pronoun resolution, and their concept of "entity connection" was limited to co-occurrence in the same document.

5. REFERENCES

[1] Chieu, H.L. and Lee, Y.K. 2004. Query based event extraction along a timeline. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2004), 425–432.

[2] Don, A. et al. 2007. Discovering interesting usage patterns in text collections: integrating text mining with visualization.

Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (Lisbon, Portugal, 2007), 213-222.

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

[4] Elson, D.K. et al. 2010. Extracting social networks from literary fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010), 138–147.

[5] Fekete, J.-D. and Dufournaud, N. 2000. Compus: visualization and analysis of structured documents for understanding social life in the 16th century. *Proceedings of the fifth ACM conference on Digital libraries* (New York, NY, USA, 2000), 47–55.

[6] Filippova, K. and Strube, M. 2007. Extending the entity-grid coherence model to semantically related entities. *Proceedings of the Eleventh European Workshop on Natural Language Generation* (Stroudsburg, PA, USA, 2007), 139–142.

[7] Furnas, G.W. et al. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*. 30, 11 (1987), 971.

[8] Gorg, C. et al. 2007. Visual Analytics with Jigsaw. *IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007* (Nov. 2007), 201-202.

[9] Gorg, C. et al. 2010. Combining Computational Analyses and Interactive Visualization to Enhance Information Retrieval. (New Brunswick, NJ, 2010).

[10] Hearst, M.A. 2009. *Search user interfaces*. Cambridge Univ Pr.

[11] Hearst, M.A. 1995. TileBars: visualization of term distribution information in full text information access. *Proceedings of the SIGCHI conference on Human factors in computing systems* (1995), 59–66.

[12] Jockers, M. 2011. Detecting and Characterizing National Style in the 19th Century Novel. *Digital Humanities 2011* (Stanford, CA, 2011).

[13] Jonker, D. et al. 2005. Information triage with TRIST. *2005 International Conference on Intelligence Analysis* (2005), 2–4.

[14] Keim, D. et al. 2008. Visual Analytics: Definition, Process, and Challenges. *Information Visualization*. A. Kerren et al., eds. Springer Berlin Heidelberg. 154-175.

[15] Klein, D. and Manning, C.D. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1* (Stroudsburg, PA, USA, 2003), 423–430.

[16] Lin, D. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (1997), 64–71.

[17] Lin, F.-ren and Liang, C.-H. 2008. Storyline-based summarization for news topic retrospection. *Decision Support Systems*. 45, 3 (Jun. 2008), 473-490.

[18] Llorà, X. et al. 2008. Meandre: Semantic-driven data-intensive flows in the clouds. *Fourth IEEE International Conference on eScience* (2008), 238–245.

[19] Michel, J.-B. et al. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. (Dec. 2010).

[20] Moretti, F. 2005. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso Books.

[21] Plaisant, C. et al. 2006. Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (Chapel Hill, NC, USA, 2006), 141-150.

[22] Rocchio, J.J. 1971. Relevance Feedback in Information Retrieval. *SMART Retrieval System Experimens in Automatic Document Processing*. (1971).

[23] Rockwell, G. et al. 2010. Ubiquitous Text Analysis. *Poetess Archive Journal*. 2, 1 (2010).

[24] Russell, D.M. et al. 1993. The cost structure of sensemaking. *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (New York, NY, USA, 1993), 269–276.

[25] Salton, G. and Buckley, C. 1997. Improving Retrieval Performance By Relevance Feedback. *Readings in information retrieval*. Morgan Kaufmann.

[26] Stasko, J. et al. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*. 7, 2 (2008), 118-132.

[27] Stoica, E. et al. 2007. Automating creation of hierarchical faceted metadata structures. *Proceedings of NAACL HLT* (2007), 244–251.

[28] Tong, S. and Koller, D. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2, (Mar. 2002), 45–66.

[29] Vuillemot, R. et al. 2009. What's being said near "Martha"? Exploring name entities in literary text collections. *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (2009), 107–114.

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

[31] Wright, W. et al. 2006. The Sandbox for analysis: concepts and methods. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), 801–810.