



HAL
open science

What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections

Romain Vuillemot, Tanya Clement, Catherine Plaisant, Amit Kumar

► **To cite this version:**

Romain Vuillemot, Tanya Clement, Catherine Plaisant, Amit Kumar. What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections. IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST), Oct 2009, Atlantic City, New Jersey, United States. pp.107-114, 10.1109/VAST.2009.5333248 . hal-01437767

HAL Id: hal-01437767

<https://hal.science/hal-01437767v1>

Submitted on 29 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

What’s Being Said Near “Martha”?

Exploring Name Entities in Literary Text Collections

Romain Vuillemot¹

Tanya Clement²

Catherine Plaisant²

Amit Kumar³

¹Université de Lyon

²University of Maryland

³University of Illinois, Urbana Champaign

ABSTRACT

A common task in literary analysis is to study characters in a novel or collection. Automatic entity extraction, text analysis and effective user interfaces facilitate character analysis. Using our interface, called POSvis, the scholar uses word clouds and self-organizing graphs to review vocabulary, to filter by part of speech, and to explore the network of characters located near characters under review. Further, visualizations show word usages within an analysis window (i.e. a book chapter), which can be compared with a reference window (i.e. the whole book). We describe the interface and report on an early case study with a humanities scholar.

KEYWORDS: Visual Analytics, Design, Experimentation, Human Factors.

INDEX TERMS: H.5.2 Graphical user interfaces (GUI)

1 INTRODUCTION

The development of digital libraries now gives scholars access to large bodies of literature. MONK (<http://www.monkproject.org>) is an example of a digital environment designed to help humanities scholars discover and analyze patterns in the texts they study. It aims to support both micro analyses of the verbal texture of an individual text and macro analyses that let you locate and analyze texts in the context of a large document space consisting of hundreds or thousands of other texts. These explorations allow scholars to practice forms of what Franco Moretti has provocatively called “distant reading” [6].

Getting salience out of data and formulating new hypotheses for making further explorations are among the many goals of visual analytics tools. While humans are good at quickly identifying shapes from diagrams or faces in images, it is very difficult to get an overview of a text collection at a glance, much less make comparisons. For example the sentence “*Peter is greater than John, and both are smaller than Adam*” takes more time for humans to understand than a simple picture depicting the same height relationships.

A common task for literary scholars is to study characters in a book or collection. They may try to characterize the relationship between family members in a novel, or study the evolution of the mentions of an historical figure in a collection of biographies. Text analysis and effective user interfaces might facilitate the

exploration of the topics discussed or the vocabulary used in the neighborhood of the characters. Using our interface, called POSvis, scholars may use word clouds and self-organizing graphs to review the vocabulary in the vicinity of one or more entities, filter by part of speech, explore the network of other characters in that vicinity, and compare different text segments.

Before going further we define some of the terms we use in the paper. The term *name entity* is used loosely to refer to names that can be extracted automatically (typically proper names). The *part of speech* classification is a grammar classification of words, based on eight categories: verb, noun, pronoun, adjective, adverb, preposition, conjunction, and interjection. We say that words or name entities *co-occur* if they both appear at least once within a fixed text window, typically set by the user (e.g. a 20 word window, or a paragraph). The *document structure* is a hierarchy based on document abstraction levels found in the documents (e.g. using XML tags). For a book it could be: book > chapter > section > paragraph etc.; these are used to choose regions to be compared.

We start by describing the problem that motivated our work, then describe POSvis’ interface, the query workflow, and finally, results exploration. In section 4, we describe POSvis architecture and text analysis techniques. Finally in section 5 we describe our early study case results; these are discussed in section 6.

2 MOTIVATION

We worked with Tanya Clement who was a senior PhD student in the English department at the University of Maryland. As part of her research Clement has been studying *The Making of Americans* by Gertrude Stein. The book is 9 chapters and 517,027 words long. According to Clement, this postmodern writing is almost impossible to read and digital tools bring a new perspective into the nature of the text and the seemingly nonsensical, non-narrative structures. Data mining and text analysis methods have been used to facilitate a new reading of this text [2] [3]. For example, data mining and visualization have been combined to analyze patterns of repetitions in the text [4]. The task addressed in this paper is an attempt to understand how the identity and relationships of family members changed over time by examining the words that co-occur with these characters. Because of the chaotic structure of the text, even an expert reader, may become lost or confused. Manually keeping track of name entities and their relationships is also difficult (we found 190 entities in the book).

3 DESCRIPTION OF THE INTERFACE

Figure 1 shows the graphical interface of POSvis [14], loaded with *The Making of Americans*, a novel from Gertrude Stein. The *Document Overview* panel (top strip) shows an overview of the document’s structure. The X-axis shows the different chapters while the height of the vertical bars is proportional to the number of words in each section. Below the overview two range sliders allow users to set scopes for comparisons: A (*analysis text*, in red) and B (*reference text*, in black).

¹E-mail: romain.vuillemot@insa-lyon.fr

²E-mail: {[tclement](mailto:tclement@umd.edu), [plaisant](mailto:plaisant@umd.edu)}@umd.edu

³E-mail : amitku@uiuc.edu

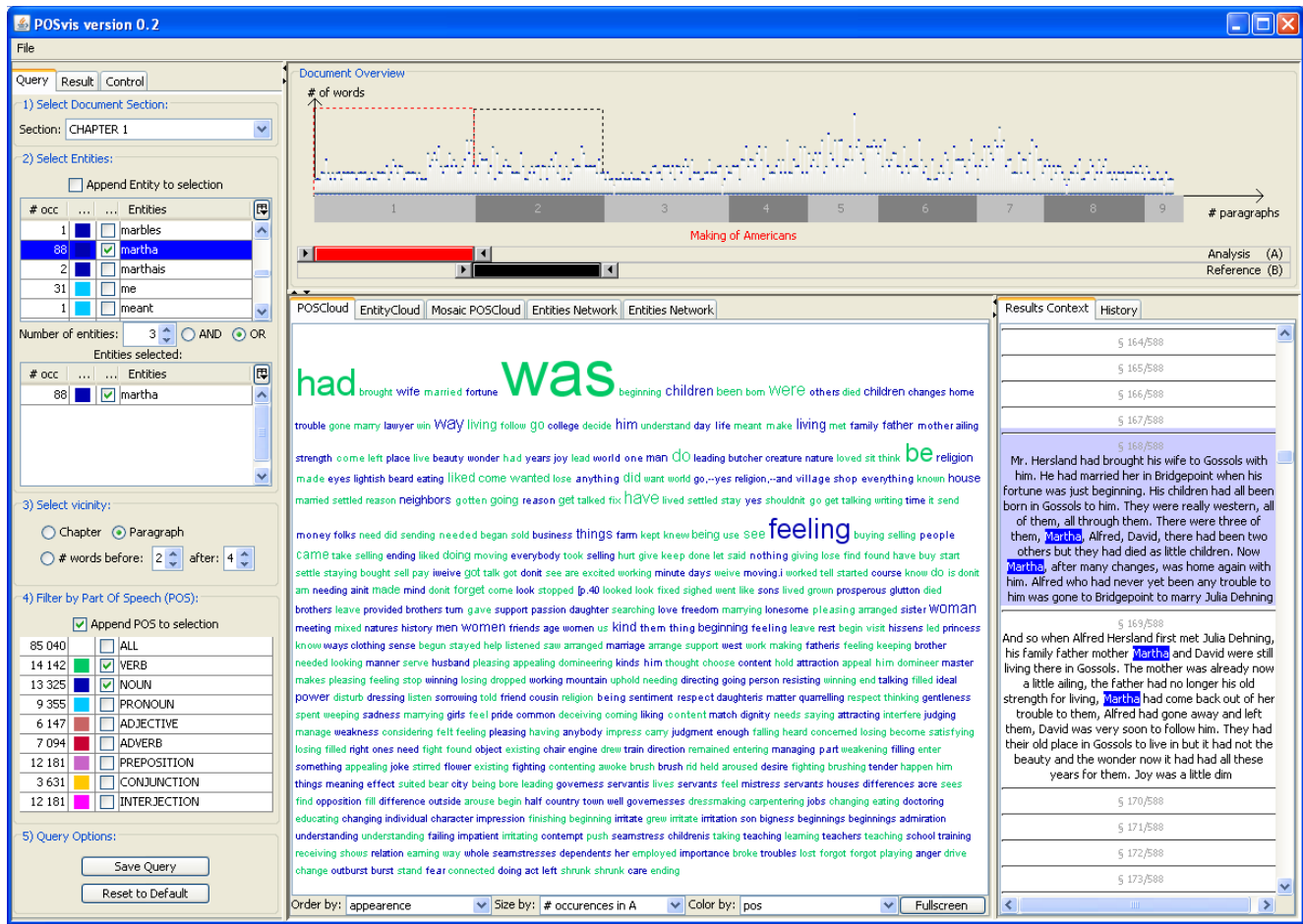


Figure 1. POSvis loaded with Gertrude Stein's book *The Making of Americans*. The top strip shows a document structure overview (here chapters) which allow users to select regions for analysis (red slider) and reference (black slider). In the query panel on the left, users can set the size of the vicinity window (here a paragraph) used for determining co-occurrence. In the control panel (tab not visible on the screenshot), entities were defined as NNP and NNPS (singular and plural proper names). One entity (Martha) was selected in the list of name entities, and verb and nouns were selected in the Part of Speech (POS) menu. Verbs and nouns found in the vicinity of Martha are summarized in a word cloud in the middle, and details on-demand are available in the right panel.

The *Query* panel located on the left side of the screen shows in a first list extracted entities with their occurrence count in the analysis section. Entities can be sorted by alphabetical or cumulative count order, to get a quick access to respectively a specific or most/less frequent item. Entities can be selected or unselected and can be (or not) appended to each other, and appear in a second list. The third list gives Part of Speech (POS) categories with cumulative occurrences counts for each category. Embedded checkboxes in strips permit multiple POS selections. At the bottom of the panel, two buttons: one for saving the query for further usage, and another one for resetting the query to start with default values.

By default the *Results* panel shows a word cloud using usage frequency information (size) and POS information (color). An export feature enables data resulting from the query to be as raw data (XML) or as a picture (PNG) to be included in presentations or blogs.

A *Control* panel (as a tab behind the *Query* panel) allows users to adjust parameters of the result display e.g. the frequency threshold for a word to be displayed in the tag cloud. Font selection, minimum and maximum font sizes and graph layout options are also available.

3.1 Query specification

First users select an analysis section (and optionally a reference section) in the *Overview* panel. Users then set the size of the text window -or the range of area around a chosen entity- to be used to determine co-occurrences, e.g., 20 words before and after, or a paragraph. The selection of name entities is iterative, using items such as checkboxes and dynamically updated menus. There is no submitting query button: every click or focus results in an action, allowing non expert users to perform complex queries that would have required advanced knowledge of Structured Query Language (SQL). Results are progressively revealed in the *results* panels. When users select items, details appear in the context panel and we emphasize focused entities / POS by respectively highlighting them in blue or in yellow. This bi-color code is used again in further selections/focus for visualizing selected items in lists, tag clouds or results in context. Another color coding is introduced for the POS categories: verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, interjections. The *properties* tab allows users to change these colors and to remove or add POS categories (e.g. to distinguish plural from singular nouns).

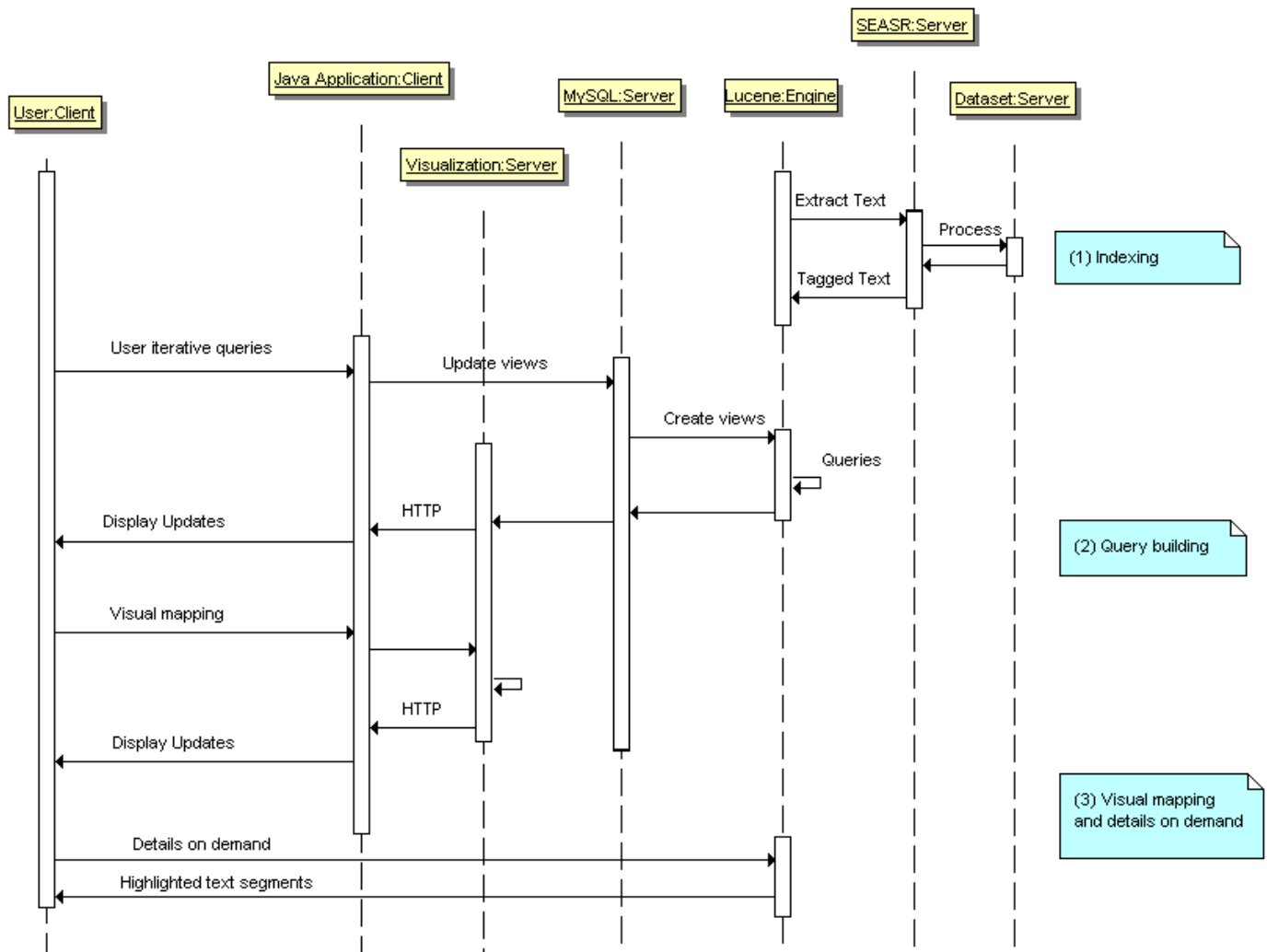


Figure 2. The system execution is threefold: 1) an offline indexing process extracts tagged chunks from text, keeping document structure 2) users select entities and POS of interest and gets tag clouds and social networks that can be customized during the last step 3) details on demand are available and retrieved by means of a query to the Lucene engine.

3.2 Word usage in vicinity of name entities

Results are presented in two tabbed panels showing either a word cloud or a social network of name entities. When the character *Martha* is selected in the list of entity, the overview shows where it appears and the word cloud shows a summary of neighboring words.

3.2.1 Word clouds by Part of Speech

Word clouds are a simple way to summarize content; it is widely appreciated by literary scholars who find it easy to use [3], and enjoy the often elegant resulting displays. The result can be displayed as a word cloud comprising the words found in the vicinity of the selected entities, and filtered to only show the entities that match the POS selected in the *query* panel. To increase the salience of the word cloud we give users the ability to define their own mapping with the controls at the bottom of the word clouds. Three independent visual variables are available: word order, size and color. They can encode text variables such as order of appearance in selection (analysis or reference) and

cumulative count of occurrence in analysis or reference selection. Other customizations to increase salience such as count thresholds or color assignments are possible using the *properties* panel on the left. Some POS can be excluded/included in the query panel, but specific entities can also be included or excluded by category or as individual entities. This is useful as some words may predominate and overwhelm the visualizations.

When multiple entities are selected the word clouds reflect the vocabulary used in the text windows where the entities co-occur.

3.2.2 Dunning log-likelihood Word clouds

To compare two text regions we used a Dunning's log-likelihood analysis [5] to highlight words that are underused or overused in the analysis region, compared to a reference region. The log-likelihood is a statistical measure that shows how significant the difference is between two sets (e.g. regions). Its application to words frequency in text comparison gives *relative* information about words usage difference in two regions.

views and holds the controller) and a remote server (where the model is hosted and views are generated).

The user client interface is a Java application. We used Prefuse [9] for interactive graph layout functions, internal data structures (tables) and the interaction widget. Otherwise we used Swing components, which are robust enough to allow users to formulate their needs.

The server becomes more complex, since we need to generate views from the model. We need 1) to build an index of the texts and answer queries using that index, 2) rank according to criteria (counts, relevance) and 3) integrate multi data sources. For that purpose we used Apache Lucene server (<http://lucene.apache.org/>), a high-performance, full-featured text search engine library written entirely in Java. The system execution process is threefold (Figure 2). First text collections are pre-processed (e.g. tagged) offline using a text analysis mashups on a server based on the Software Environment for the Advancement of Scholarly Research (SEASR) [12] and indexed by Lucene. The goal of the SEASR project is to create a flexible and scalable architecture that can be quickly deployed and reused for the humanities. This way, additional text processing such as ignoring stop words or porter stemming can easily be included in the data flow, even by a scholar using the SEASR interface drag and drop intuitive interface. Then, while users construct their queries, views are created in a cache MySQL database based on word attributes. Word clouds are generated from the MySQL database by a visualization server, available as a RESTful Web Service published over HTTP (without proxy restriction). This way, word clouds can be seen as resources to which we pass parameters for options and results are in XML-like file format. Resulting XML files have a unique URL which can be visualized in any web-browser and easily shared or plugged into another system. Finally, the URL is imported by the Java application and is coupled with a local custom CSS file, according to users color mappings. The social network results from a query to Lucene index, and filled into edges and links tables. If users want details on data, such as the original text, queries are performed directly to the Lucene which very quickly retrieve the document and highlight results.

4.1 Discussion of the architecture

Interface reactivity is crucial since we want users to iteratively perform queries -make them and remake them- and visualize results. The main bottleneck appears when an entity is added to the query: then the system has to generate new views on the dataset. These views are virtual tables that require time to be created but are very quick to interrogate. A view's lifetime is a user session long, and then it is deleted. Saving or pre-generating these views as cache is a viable optimization, but the trade off is that it needs lots of disk space. As a short term solution, batch processes can be triggered after user queries; users are then advised of the estimated time remaining to complete their tasks, and they get a dialog window notification when the job is done and the interface is ready to use.

Note that index creation would be the natural way to make queries faster, creating extra columns in database. With the varying windows (n words before, m words after) index creation of all variations is not practical (or it would require indexing all possible window size which results in a combinatory explosion). But pre-indexing on section or paragraph size windows might be a good compromise. Another further optimization is outsourcing independent time-consuming computations to clouds or distributed database. This requires making processes independent, launching them and finally gathering results, adding complex merging and checking constraints without being able to predict the response time.

5 EXAMPLE USE CASE

Before designing POSvis, our literary scholar partner (Tanya Clement) had used Wordle (<http://www.wordle.net>) to look at word frequencies in the different chapters. Wordle has been greatly appreciated by literary scholars who can simply load lists of word frequencies (instead of having to load large text documents) and see the results. Still it became quickly clear that raw word frequencies were not useful as they do not allow users to compare with a reference text to see what is really different in the text of interest. The next step was to load Dunning's log-likelihood values comparing *The Making of Americans* (the novel written by Gertrude Stein in 1925) with a set of 19th Century novels. By comparing Stein with different novels, Tanya noticed the strong prominence of the word "one". Looking again a simple frequencies at simple word frequency lists revealed that the frequency of *one* surges by the end of the book, but after reviewing the original text segments it became clear that the word *one* —unlike *he, she, I, we, or even you or it*—played many roles in the text (Figure 6), e.g. the role of a pronoun or an adjective, in the subject or object position. Plain word frequency information would not have been useful in determining the word's usage.

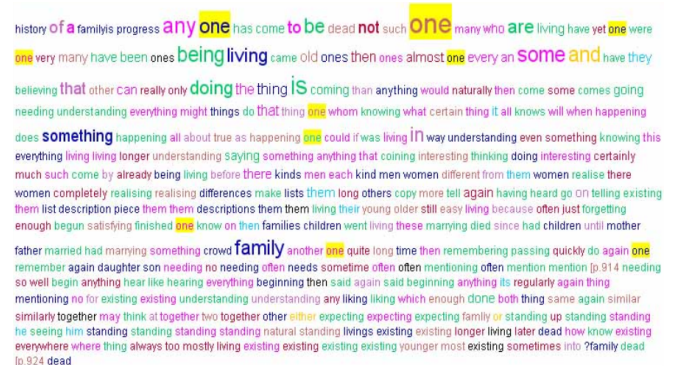


Figure 6. The word "one" plays many roles in Stein novel. Each Part of Speech version of the word "one" appears separately in this word cloud, highlighted in yellow. Hovering over each occurrence reveals roles such as pronoun or adjective, and in the subject or object position. The words are ordered "by appearance in the text" revealing that some roles are only introduced later in the text.

Clement proposed that the high frequency of *one* was the result of the confusion accomplished by the word "one" schizophrenic nature. While POSvis had originally been intended to study proper names, it became important to allow users to analyze entities which are not proper names. By using POSvis, the progression of the manner in which the word "one" was used in terms of different parts of speech was documented, allowing Clement to see that the use of "one" changed as the text progressed. The analysis led to a reading in which "one" represents a singular subject position and multiple subject positions at once. With this information, Tanya could make the argument that "the discourse about identity formation in *The Making of Americans* is engaged in this multiplicity, not dissolved in indeterminacy", which led to a publication [3], and inclusion in her PhD thesis.

6 DISCUSSION AND FUTURE WORK

The literary scholars who provided feedback on the prototype could readily find potential use scenarios in their own work. Nevertheless their examples made it clear that character names are rarely going to be as easy as finding "Victor Hugo" in historical texts. While the problem of missed references was found

REFERENCES

- [1] Bateman, S., Gutwin, C., and Nacenta, M. 2008. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia* (Pittsburgh, PA, USA, June 19 - 21, 2008). HT '08. ACM, New York, NY, 193-202.
- [2] Clement, T. (2008). 'A thing not beginning or ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*. In *Literary and Linguistic Computing* (2008), 23.3: 361-382.
- [3] Clement, T., Plaisant, C., Vuillemot, R. The Story of One: Humanity scholarship with visualization and text analysis. In *Digital Humanities Conference* (2009). DH2009.
- [4] Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., and Plaisant, C. 2007. 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* (Lisbon, Portugal, November 06 - 10, 2007). CIKM '07. ACM, New York, NY, 213-222.
- [5] Dunning (2009). <http://wordhoard.northwestern.edu/userman/analysis-comparewords.html#loglike>. Retrieved 03/2009.
- [6] Eakin, E., Studying Literature By the Numbers, <http://www.nytimes.com/2004/01/10/books/10LIT.html>. Retrieved 03/2009.
- [7] Fekete J.-D. and Dufournaud, N., Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In DL '00: Proceedings of the fifth ACM conference on Digital libraries, New York, NY, USA (2000) 47-55.
- [8] Hearst, M. A. and Rosner, D. 2008. Tag Clouds: Data Analysis Tool or Social Signaller?. In *Proceedings of the Proceedings of the 41st Annual Hawaii international Conference on System Sciences* (January 07 - 10, 2008). HICSS. IEEE Computer Society, Washington, DC, 160.
- [9] Jerrey Heer, Stuart K. Card, and James A. Landay. prefuse: a toolkit for interactive information visualization. In *CHI*, pages 421-430, 2005.
- [10] Kaser, O., Lemire, D. Tagcloud drawing: Algorithms for cloud visualization. In *WWW*, 2007.
- [11] D. Keim, D. Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis, *IEEE Symposium on Visual Analytics and Technology* (2007) 115-122
- [12] Xavier Llorà, Bernie Ács, Loretta S. Auvil, Boris Capitanu, Michael E. Welge, David E. Goldberg, "Meandre: Semantic-Driven Data-Intensive Flows in the Clouds," *escience*, pp.238-245, 2008 In *Fourth IEEE International Conference on eScience*, 2008
- [13] D. Oelke, P. Bak, D. Keim, M. Last, G. Danon: Visual evaluation of text features for document summarization and analysis, *Proceedings IEEE Symposium on Visual Analytics and Technology* (2008) 75-82
- [14] POSvis website. <http://www.cs.umd.edu/hcil/textvis/posvis/>. Retrieved 03/2009.
- [15] Rivadeneira, W., Tag clouds: how format and categorical structure affect categorization judgment, Psychology PhD Thesis (2008).
- [16] Stasko, J., Görg, C., and Liu, Z. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization* 7, 2 (Apr. 2008), 118-132.
- [17] Wattenberg, M., Kriss, J., and McKeon, M. 2007. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1121-1128.
- [18] Wattenberg, M. and Viegas, F. (2008). Tag clouds and the case for vernacular visualization. In *Interactions*, 15.4: 49-52.
- [19] Wattenberg, M. and Viégas, F. B. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov. 2008), 1221-1228