What’s Being Said Near “Martha”?

Exploring Name Entities in Literary Text Collections

Romain Vuillemot[1] Tanya Clement[2] Catherine Plaisant[2] Amit Kumar[3]

[1]Université de Lyon
[2]University of Maryland
[3]University of Illinois, Urbana Champaign

Abstract

A common task in literary analysis is to study characters in a novel or collection. When dealing with large documents or collections automatic entity extraction, 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, the scholar uses word clouds and self-organizing graphs to review the vocabulary in the vicinity of one or more entities, to filter by part of speech, and to explore the network of other characters in that vicinity. 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: Design, Experimentation, Human Factors.

Index Terms:

1 Introduction

The development of digital libraries now gives scholars access to large bodies of literature. MONK 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. and used to choose regions to be compared.

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

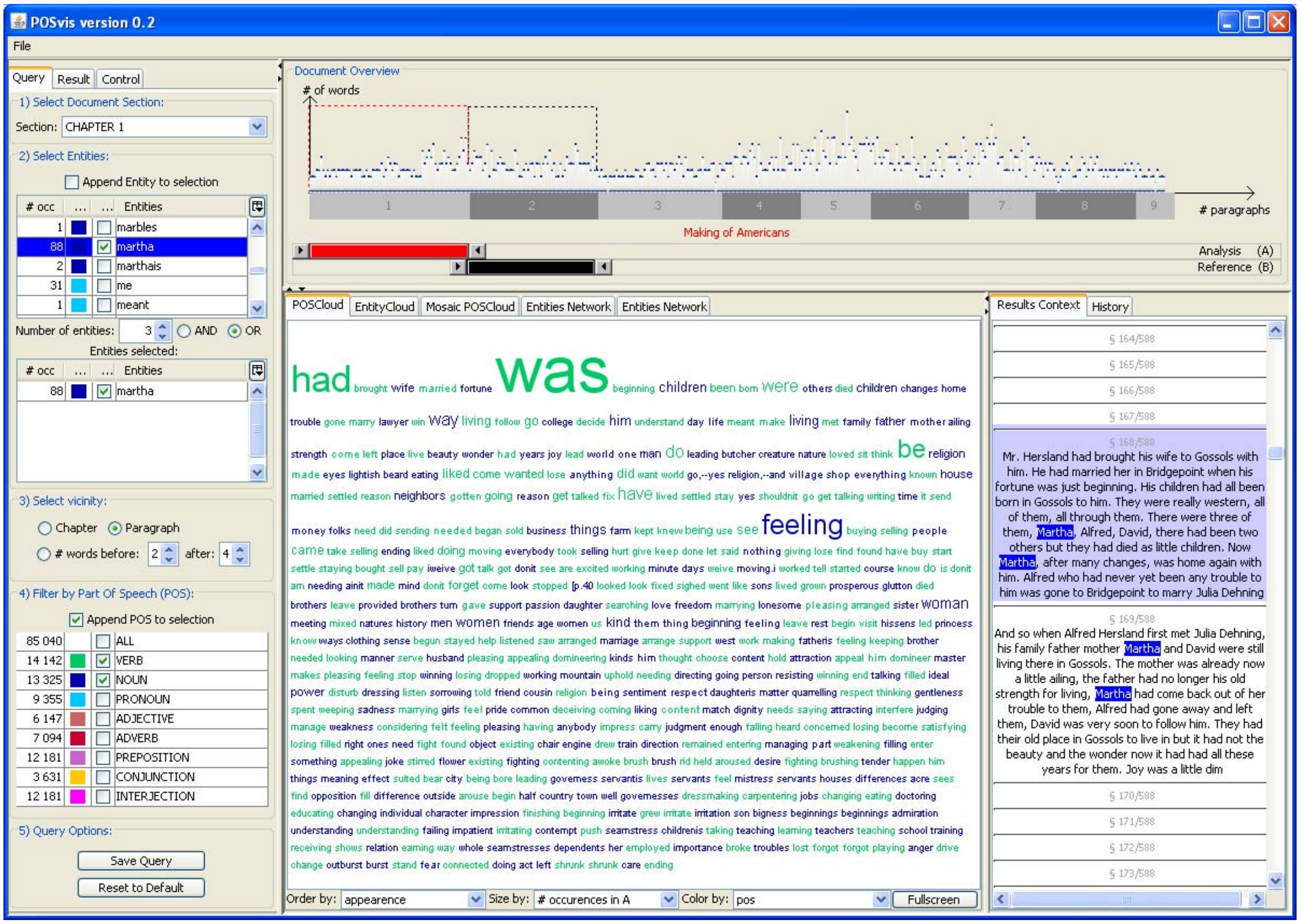
2 Motivation

We worked with Tanya Clement who received her PhD in the English department at the University of Maryland. As part of her research Tanya has been studying *The Making of Americans* by Gertrude Stein. The book is 9 chapters and 517,027 words long. 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 to understand how the identity and relationships of family members changed over time i.e. throughout the book. Because of the chaotic structure of the text, a reader, even expert, 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

POSvis follows Shneiderman's Information Visualization mantra of "overview first, zoom and filter, details on demand". Figure 1 shows the graphical interface, using *The Making of Americans* text collection loaded. The *Document Overview* panel (top strip) represents an overview of the document’s structure. The X-axis shows chapters. The Y-axis is proportional to the number of words in the sections. Filter controls are two range sliders, A (*analysis text*, in red 🟥) and B (*reference text*, in black ⬛), that permit to set the document scopes for comparisons.

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.


POSvis version 0.2
File
Query
Result
Control
Document Overview
# of words
1
2
3
4
5
6
7
8
9
# paragraphs
Making of Americans
Analysis (A)
Reference (B)
POSCloud
EntityCloud
Mosaic POSCloud
Entities Network
Entities Network
Results Context
History
had brought wife married fortune was beginning children been born were others died children changes home
trouble gone marry lawyer win way living follow go college decide him understand day life meant make living met family father mother ailing
strength come left place live beauty wonder had years joy lead world one man do leading butcher creature nature loved sit think be religion
made eyes lightish beard eating liked come wanted lose anything did want world go,--yes religion,--and village shop everything known house
married settled reason neighbors gotten going reason get talked fix have lived settled stay yes shouldnt go get talking writing time it send
money folks need did sending needed began sold business things farm kept knew being use see feeling buying selling people
came take selling ending liked doing moving everybody took selling hurt give keep done let said nothing giving lose find found have buy start
settle staying bought sell pay iweive got talk got donit see are excited working minute days weive moving.i worked tell started course know do is donit
am needing ainit made mind donit forget come look stopped [p.40 looked look fixed sighed went like sons lived grown prosperous glutton died
brothers leave provided brothers turn gave support passion daughter searching love freedom marrying lonesome pleasing arranged sister woman
meeting mixed natures history men women friends age women us kind them thing beginning feeling leave rest begin visit hissens led princess
know ways clothing sense begun stayed help listened saw arranged marriage arrange support west work making fatheris feeling keeping brother
needed looking manner serve husband pleasing appealing domineering kinds him thought choose content hold attraction appeal him domineer master
makes pleasing feeling stop winning losing dropped working mountain uphold needing directing going person resisting winning end talking filled ideal
power disturb dressing listen sorrowing told friend cousin religion being sentiment respect daughteris matter quarrelling respect thinking gentleness
spent weeping sadness marrying girls feel pride common deceiving coming liking content match dignity needs saying attracting interfere judging
manage weakness considering felt feeling pleasing having anybody impress carry judgment enough falling heard concerned losing become satisfying
losing filled right ones need fight found object existing chair engine drew train direction remained entering managing part weakening filling enter
something appealing joke stirred flower existing fighting contenting awoke brush brush rid held aroused desire fighting brushing tender happen him
things meaning effect suited bear city being bore leading governess servantis lives servants feel mistress servants houses differences acre sees
find opposition fill difference outside arouse begin half country town well governesses dressmaking carpentering jobs changing eating doctoring
educating changing individual character impression finishing beginning irritate grew irritate irritation son bigness beginnings beginnings admiration
understanding understanding failing impatient irritating contempt push seamstress childrenis taking teaching learning teachers teaching school training
receiving shows relation earning way whole seamstresses dependents her employed importance broke troubles lost forgot forgot playing anger drive
change outburst burst stand fear connected doing act left shrunk shrunk care ending
Fullscreen
Save Query
Reset to Default

1) Select Document Section:

Section: CHAPTER 1

2) Select Entities:

☐ Append Entity to selection

# occ	...	...	Entities
1	🟦	☐	marbles
88	🟦	☑	martha
2	🟦	☐	marthais
31	🟦	☐	me
1	🟦	☐	meant

Number of entities: 3 ○ AND ● OR

Entities selected:

# occ	...	...	Entities
88	🟦	☑	martha

3) Select vicinity:

○ Chapter ● Paragraph

○ # words before: 2 after: 4

4) Filter by Part Of Speech (POS):

☑ Append POS to selection

85 040		☐	ALL
14 142	🟩	☑	VERB
13 325	🟦	☑	NOUN
9 355	🟦	☐	PRONOUN
6 147	🟥	☐	ADJECTIVE
7 094	🟥	☐	ADVERB
12 181	🟪	☐	PREPOSITION
3 631	🟧	☐	CONJUNCTION
12 181	🟪	☐	INTERJECTION

5) Query Options:

Order by: appearence Size by: # occurences in A Color by: pos

§ 164/588

§ 165/588

§ 166/588

§ 167/588

§ 168/588

Mr. Hersland had brought his wife to Gossols with him. He had married her in Bridgepoint when his fortune was just beginning. His children had all been born in Gossols to him. They were really western, all of them, all through them. There were three of them, Martha, Alfred, David, there had been two others but they had died as little children. Now Martha, after many changes, was home again with him. Alfred who had never yet been any trouble to him was gone to Bridgepoint to marry Julia Dehning

§ 169/588

And so when Alfred Hersland first met Julia Dehning, his family father mother Martha and David were still living there in Gossols. The mother was already now a little ailing, the father had no longer his old strength for living, Martha had come back out of her trouble to them, Alfred had gone away and left them, David was very soon to follow him. They had their old place in Gossols to live in but it had not the beauty and the wonder now it had had all these years for them. Joy was a little dim

§ 170/588

§ 171/588

§ 172/588

§ 173/588

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. Words found in the vicinity of Martha are summarized in a word cloud in the middle, and details on-demand are available on the right panel.

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). The *full screen* button makes the interface full screen to better explore results and communicate them (data export is available as PNG and XML).

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 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 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 we emphasize focused entities or POS by respectively highlighting them in blue 🟦 or in yellow 🟨. We will keep this bi-color code in further selections/focus, such as 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 🟪. Those colors can be changed in the *properties tab* and be more detailed regarding POS sub categories (e.g. to distinguish plural from singular nouns).

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. If the character *Martha* is selected, the results display will show where it frequently appears and 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, and enjoy the often elegant resulting displays. The result can be displayed as a word cloud of 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 ate available: words order, size and color. They can encode text variables such as order of appearance in selection (analysis or reference), cumulative count of occurrence in analysis or reference selection. Other customizations to increase salience such as count thresholds or colors assignments are possible using the *properties* panel on the left. Some POS can be excluded/included in the query panel, but also specific entities can be included or excluded (by category or individual entities). This is useful as some words may be particularly preponderant than others and overwhelm the visualizations.

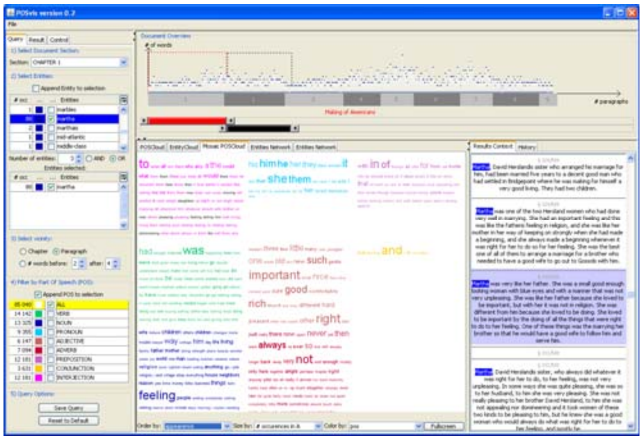


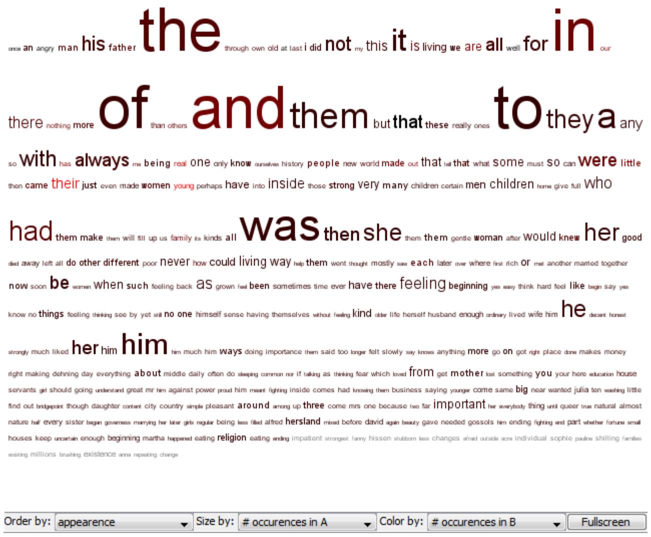
Figure 2 We introduce *mosaic word clouds* that shows word clouds according to categories. The color caption is available on the query panel (left panel).

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.

A single tag cloud can encode words issued from many categories, but this we lose comparisons within each category. For that purpose we introduced *mosaic word cloud* that shows 3x3 small word clouds (Figure 2) encoding the different categories. All the tag clouds share the same color, size and order encoding.

The multiple views are tightly coupled. For instance, if a word is selected on the tag cloud, lists on the left are automatically updated to reflect available choices (and users get quantitative frequency values), and on the right panel the context view shows quotes from the text where the word appears. This way ambiguity can be weaved and by clicking on quotes the full text is displayed in the *document content* tab, i.e. hiding the word cloud. This strong coordination helps exploring each result, and they can also be sequentially explored with key down/up to let users focus on their task with only one degree of freedom to control and be sure not to miss any entry.


once an angry man his father the through own old at last i did not my this it is living we are all well for in our
there nothing more of than others and them but that these really ones to they a any
so with has always me being real one only know ourselves history people new world made out that tell that what some must so can were little
then came their just even made women young perhaps have into inside those strong very many children certain men children home give full who
had them make them will fill up us family its kinds all was then she them them gentle woman after would knew her good
died away left all do other different poor never how could living way help them went thought mostly saw each later over where first rich or met another married together
now soon be women when such feeling back as grown feel been sometimes time ever have there feeling beginning yes easy think hard feel like begin say yes
know no things feeling thinking see by yet still no one himself sense having themselves without feeling kind older life herself husband enough ordinary lived wife him he decent honest
strongly much liked her him him him much him ways doing importance them said too longer felt slowly say knows anything more go on got right place done makes money
right making defining day everything about middle daily often do sleeping common nor if talking as thinking fear which loved from get mother lost something you your here education house
servants girl should going understand great mr him against power proud him meant fighting inside comes had knowing them business saying younger come same big near wanted julia ten washing little
find out bridgepoint though daughter content city country simple pleasant around among up three come mrs one because two far important her everybody thing until queer true natural almost
nature half every sister began governess marrying her later girls regular being less filled alfred hersland mixed before david again beauty gave needed gossols him ending fighting end part whether fortune small
houses keep uncertain enough beginning martha happened eating religion eating ending impatient strongest fanny hissen stubborn less changes afraid outside acre individual sophie pauline shilling families
existing millions brushing existence anna repeating change
Fullscreen

Order by: appearence ▾ Size by: # occurences in A ▾ Color by: # occurences in B ▾

Figure 3. Martha has been selected with chapter 1 as analysis text (*A*) and chapter 2 as reference text (*B*). Words order encodes appearance in *A*, font size encodes the number of occurrence in *A*, and color (from black to red) encodes the number of occurrences in *B*.

Multiple encoding allows to quickly getting what are the most salient words, and interactions can be possible by varying reference or analysis selections. Beneath lists selection are mandatory captions that communicate what is encoded with what.

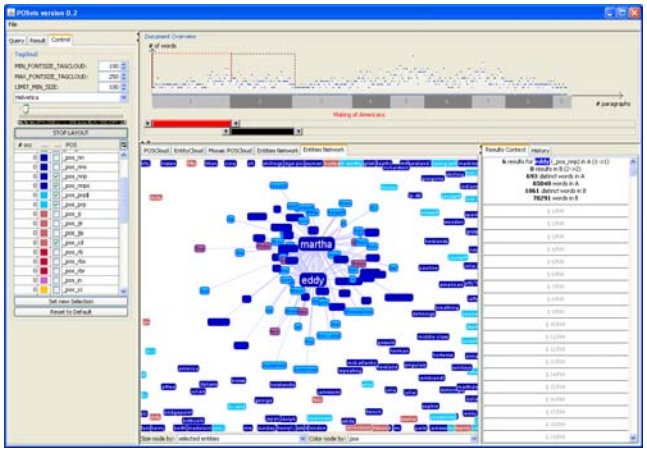


Figure 4. A second entity (Eddy) has been selected (either with the left checkbox or by double clicking on the name in the network panel) and is getting closer to Martha since they are both sharing name entities. We also selected new POS sub-categories as entities (left *Control* panel).Color coding is by POS.

3.2.3 Co-occurrence entities graph

When it comes to visualize interactions among multiple items, a network data structure fits well; but two main issues remain: how to construct the graph and how to interactively lay it out. The graph construction follows our definition (two entities are connected if they both co-occur within the same region).

We think self organizing networks permit an efficient use of the screen space and quickly get to a stable state. There exists many graph drawing or morphing techniques, some may fit better, but

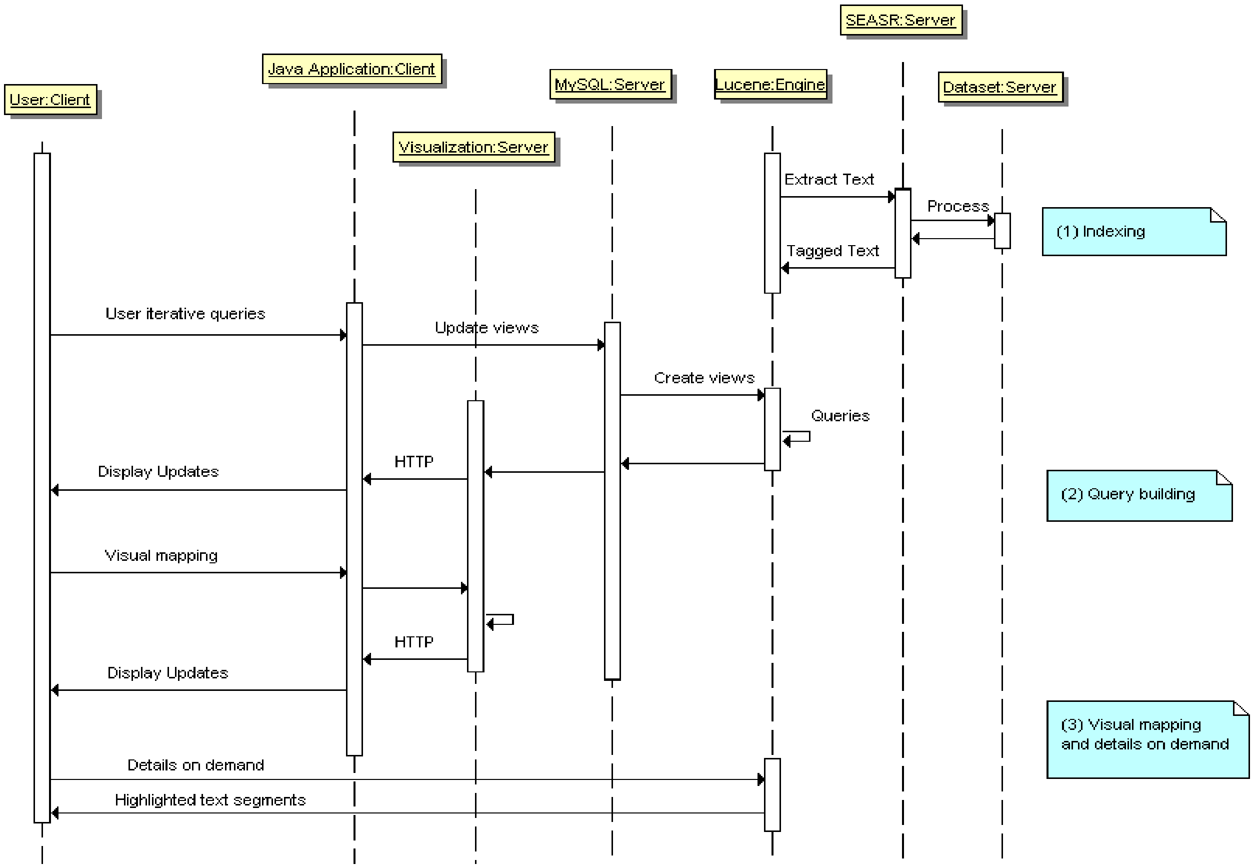


Figure 5 The system execution is threefold: 1) an offline indexing process extracts tagged chunks from text, keeping document structure 2) users select entities and POS of interest and gets tag clouds and social networks that can be customized during the last step 3) details on demand are available and retrieved by means of a query to the Lucene engine.

our focus was on real time force-directed layout computations with dynamic updates. Users can filter the network with strict co-occurrence (entities are connected to other entities that have to also appear). Like for word clouds, network visual attributes (entities/links color and size) can encode data attributes. The network holds the same coordination features as the tag cloud. An extra feature is that if users double-click on names, they are automatically added/removed from the query, which allows direct querying rather than shuttling back and forth from the query panel.

4 Architecture

The user interface was implemented as a Java application for the client interface. We used Prefuse [9] for interactive graph layout functions and internal data structures (tables). Regarding the backend part, we needed 1) to build an index of the texts and answer queries using that index, 2) rank according to criteria (counts, relevance) and 3) multi data source integration. 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 5). 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 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 cleaning up stop words or porter stemming can easily be included in the data flow, even by scholar using SEASR interface drag and drop intuitive interface. Then, while users construct their queries, views are created in a cache MySQL database based on words attributes. Word clouds are generated from the MySQL database by a visualization server, available as a RESTful Web Service published over HTTP (so no 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 to the Lucene which very quickly retrieve the document and highlight results.

4.1 Discussion on architecture design

During the architecture design, keeping interface reactive and client light have been our major concerns.

Interface reactivity is crucial since we want users to iteratively perform queries 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. Views lifetime is a user session long, and then are delete. 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 users queries, users are then advised of estimated time, and they get a dialog window to be notified when the job is done and the interface ready to use.

Note that even if index creation would be the natural way to make queries quicker, in our case index are not only based on a specific field only (chapter id, paragraph id...) but also on *n* word before and *m* word after varying windows that can't be indexed. Another further optimization is outsourcing independent time-consuming computations to clouds or distributed databases, but requires making processes independent, launching them and finally gathering results.

5 PILOT CASE STUDY

Before designing POSvis, our literary scholar partner had successfully 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 the list of word frequencies (instead of large text documents) and see the results. What became immediately evident in comparing the Wordles of different chapters was the prominence of a particular word that consistently scores a high value: *one*, which prompted our scholar to look closely at the use of that word. While POSvis was originally intended to be used with proper names, it was quickly extended to allow the use of entities which are not proper names. Comparisons revealed that the frequency of *one* surges by the end of the book. After reading the text segments it became clear that the word *one* —unlike *he*, *she*, *I*, *we*, or even *you* or *it*—plays many positions in the text, in the role of a pronoun or an adjective and in the subject or object position, therefore the possible surge in frequency. Our scholar proposed that the high frequency of *one* was the result of the confusion accomplished by the word's schizophrenic nature. Thus, 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 our scholar to see that the use of *one* appears to change as the text progresses. The analysis lend to a reading in which *one* may represent a singular subject position or multiple subject positions at once. With this information, an argument could be made that the discourse about identity formation in *the Making of Americans* is engaged in this multiplicity, not dissolved in indeterminacy, which let to a publication ([3], and inclusion in a 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 to spot as finding "Victor Hugo" in historical texts. While the problem of missed references was found acceptable (e.g. when a pronoun is used instead of the proper name), scholars asked to be allowed to add specific name entities (e.g. the "little prince" or "fox" when analyzing Saint Exupéry's story), or nouns such as "mother" when unambiguous. Multiple named entities may need to be grouped into a single one (e.g. "Mother" and "Martha") or all the names of family members could be aggregated into a single family entity.

Today the current POSvis implementation only allows the comparison of text regions using the document structure to select regions (e.g. comparing one chapter to another). A logical next step is to allow scholars to find entities and compare larger sets of texts, for e.g. scholars may want to compare the vocabulary in the vicinity of "Victor Hugo" in text written in different centuries or by male or female authors. The easy manipulation of such collections is part of the tools being developed by the Monk project and the design of POSvis should be integrated into MONK.

In the interface, the color coding of the Parts of Speech is problematic because of the large number of colors needed. Giving users control of the choice of color is helpful, but it might be necessary to allow them to group POS categories to limit the number of colors. While experimenting with the graph visualization, we found that being able to freeze the layout was important to avoid constant movement and to allow users to focus on the changes of highlighting and linking.

We know that the use of Dunnings log likelihood ratio is familiar to many of the literary scholars we work with but not all of them, therefore providing multiple word statistics and word cloud layout to compare texts might be needed, but this may in turn bring confusion as to which statistics is been used. For example when comparing two text collections with Dunnings; only words that appear in both analysis and reference can be compared. Other displays use plain word counts and duplicate the words (e.g. Many Eyes in Figure 7), first putting words in A, then in A&B, and finally in B, but there is a substantial amount of wasted space so fewer words can be shown overall. Tag clouds have been shown to be problematic and rarely more effective than plain lists [14], but they are greatly appreciated by humanity scholars who love the intuitive picture they give of the text. Further case studies need to explore the best design with this user population.

From a broader perspective, coming challenges on large textual data sets analysis can be summarized as Martin Wattenberg did in a Wired article[1]: *"The entire literary canon may be smaller than what comes out of particle accelerators or models of the human brain, but the meaning coded into words can't be measured in bytes. It's deeply compressed. Twelve words from Voltaire can hold a lifetime of experience."*

RELATED WORK

Tagclouds are one way to get a spatial, birds-eye view of keywords from a text document. Looking at a tagclouds gives an impression on the type of content. Their building is to filter large amounts of text using salience and interaction to find a better mapping with text abstraction attributes [16]. Tagclouds are scattered all over the web thanks to their easy construction and understanding, but their use tend to decline since there has not been much improvements and good uses of them expect a eye candy display of hyperlinks. Limits mostly appear in bad interaction design. Authors in [8] make a strong critique against tagclouds, complaining they are hard to compare and that the more letters they have the more they seem important. From our experience in reviewing websites or systems implementing tag clouds, another critique is that most of the time captions are missing and it is not possible to determine what is encoded with what, such as if colors or word orders have meanings.

In [1] authors studied the independent variables in tag clouds and ranked them from important (font size, font weight, intensity), less important (number of pixels, tag width, tag area) and to handle with care (color, position).

Optimizing tagcloud layouts with aesthetics has been tackled more recently with Wordle. The drawback with Wordle is that vertical or reverse words -even biggest ones- don't catch attention since they are hard to read. Comparisons among tags become even more complex. An innovative way is to show relationships between tags by Dynamic TagClouds[2] (available as a Wordpress plug-in) that is a Flash application that shows tags in a circle

[1] http://www.wired.com/science/discoveries/magazine/16-07/pb_visualizing (retrieved 03/2009)

[2] http://blog.figmentengine.com/2008/11/dynamic-tag-cloud-v12.html

(Figure 6). If the user’s mouse hovers over a tag, dashed links appear among the linked tags.

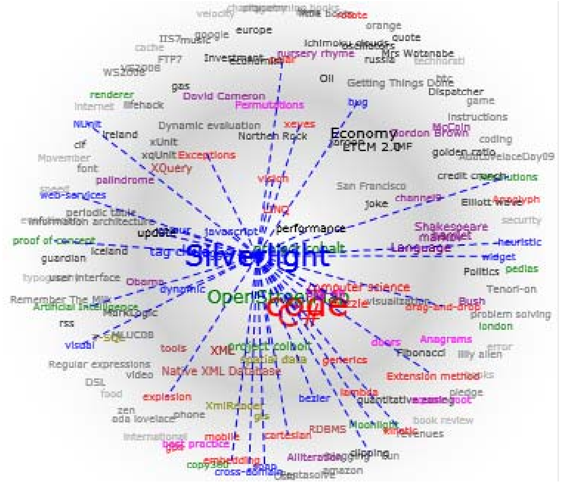


Figure 6. Dynamic tagclouds show relationships among tagclouds when the mouse hovers a tag.

Many Eyes [15] provides visualizations of users’ uploaded datasets and facilitates sharing among a community (example Figure 7). Major limits are interactivity (such as ranking or attributes color coding) and Java implementation, which makes it difficult to export raw data for further analysis. A recent visualization, Phrase Net (Figure 8) shows co-occurrences in uploaded texts.

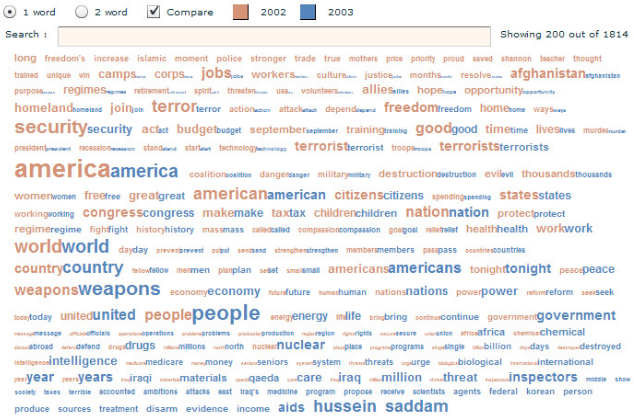
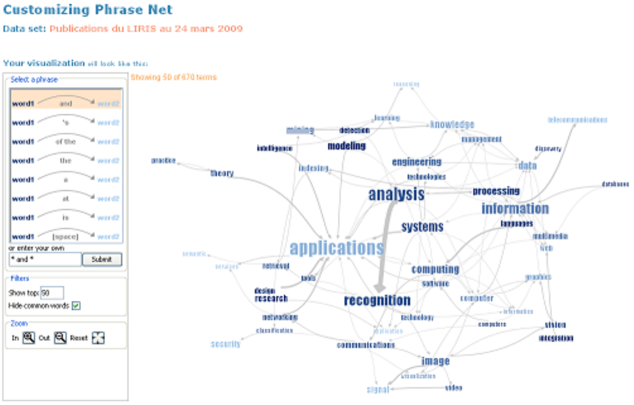


Figure 7. Comparison tagclouds from Many Eyes shows first words in A only, then in both A&B, and finally in B only. Data set compares US presidential State of the Union address from 2002 and 2003[3]

Pixel based visualizations have been used to present “fingerprints” of the texts, to facilitate analysis and comparisons (e.g. for opinion analysis and document summarization) [13] [11]. The visualization of richly tagged collections has been shown to be useful to literary scholars in their analysis (e.g. Compus for historical research [7]).


Customizing Phrase Net

Data set: Publications du LIRIS au 24 mars 2009

Your visualization will look like this:

Showing 50 of 670 terms

Figure 8. The Phrase Net visualization at Many Eyes shows co-occurrences in uploaded texts. A selection list (left vertical strip) gives details about relationships type, but characteristics are strictly limited to linking words.

CONCLUSION

We described an interface called POSvis that uses 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. Visualizations showed 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 reported on an early case study with a humanity scholar and discussed future works.

ACKNOWLEDGMENTS

We appreciate support from the Andrew W. Mellon foundation as part of the MONK project. This work was also support by Région Rhône-Alpes (France) mobility grant support.

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, abstract to appear 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

[3] http://manyeyes.alphaworks.ibm.com/blog/2008/04/01/a-cloud-of-comparisons/

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] Jerey 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] Rivadeneira, W., Tag clouds: how format and categorical structure affect categorization judgment, Psychology PhD Thesis (2008).

[15] 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.

[16] Wattenberg, M. and Viegas, F. (2008). Tag clouds and the case for vernacular visualization. In *Interactions*, 15.4: 49-52.