Interactive Text Mining Suite: Data Visualization for Literary Studies

Olga Scrivner and Jefferson Davis

Indiana University

Abstract

In recent years, there has been growing interest in visualization methods for literary text analysis. While text mining and visualization tools have evolved into mainstream research methods in many fields (e.g. social sciences, machine learning), their application to literary studies still remains infrequent. In addition to technological challenges, the use of these tools requires a methodological shift from traditional *close reading* to *distant reading* approaches. This transition also aligns digital humanities with corpus linguistics, which still "remains obscure" and not fully embraced by digital humanists [16]. To address some of these challenges, we introduce *Interactive Text Mining Suite*, a user-friendly toolkit developed both for digital humanists and corpus linguists. We further demonstrate that the integration of visual analytics and corpus linguistics methods helps unveil language patterns, otherwise hidden from a human eye. Making use of both linguistically annotated data and natural language processing techniques, we are able to discern patterns of part-of-speech uses in Medieval Occitan manuscript *Romance de Flamenca* and its English translation. Furthermore, visual analysis not only detects stylistic differences at a word level, but also at sentential and document levels. While preserving traditional *close reading* techniques, this toolkit makes it possible to apply an interactive control over documents, thus allowing for a "synthesis of computational and humanistic modes of inquiry" [18].

1 Introduction

In the past three decades, the digital humanities has evolved from Text Encoding Initiative and large-scale digital projects to a field in its own right [11]. This change has also entailed a shift in conceptual and methodological foundations. As Schnapp and Presner state in the Digital Humanity Manifesto 2.0, "the first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database". With the second wave, the focus has shifted to "qualitative, interpretive and emotive" aspects, concentrating on "digital toolkits in the service of the Humanities' core methodological strengths" [23, 2]. As the volumes of digital collections continue to grow, we are moving into the third wave, where the

emphasis is placed on search, retrieval, and analysis, focusing on “the underlying computationality of the forms held within a computational medium” [2]. With this shift, traditional methods become increasingly ineffective, leading to a transition from traditional *close reading* to *distant reading* analyses [21]. As Matthew Jockers affirms in his *Macroanalysis: Digital Methods and Literary History*, "massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making" [17]. Built from quantitative models and evolutionary theories, *distant reading* methods encouraged the use of graphs and maps to interpret textual data [22]. With recent advances in computing, these methods have further evolved into more sophisticated models involving machine learning algorithms for topic modeling and cluster analysis. Despite these advances, most commonly used computational methods in literary studies still remain *primitive* and are limited to word frequencies, concordances, and keyword-in-contexts [17]. First, many text processing tools require some programming skills, which take time to learn and are often challenging for literary scholars. Secondly, while some visualization tools (e.g. Voyant, Weka, and PaperMachine) provide graphical-user interfaces, social and humanities researchers seek more interactive and dynamic control of modeling, which can serve as “holistic support for exploratory analysis” [19].

In this paper, we propose to address these issues by integrating *micro* and *macro*analyses with a dynamic interactive interface in which a researcher has control over text analysis and visualization. To illustrate the application of such techniques for digital humanities, we analyze Medieval Occitan *Romance of Flamenca* translated in English by Blodget [5].

The remainder of this paper is organized as follows. In Section 2, we review existing *close reading* and *distant reading* visualization tools. In Section 3, we introduce *Interactive Text Mining Suite* and its functionalities. Section 4 describes our visualization analysis of *Romance of Flamenca*. Our conclusion and future development directions are presented in Section 5.

2 Close Reading and Distant Reading: Visualization Tools

The tradition of *close reading* is associated with American New Criticism developed in the 1930s [15]. The close textual analysis of individual texts was thought of as a *principle of order*, demonstrating that literature was “an autonomous mode of discourse with its own special ‘mode of existence’, distinct from that of philosophy, politics, and history” [9, 145]. In contrast, *distant reading*, introduced by Moretti [21], refers to as “the construction of abstract models” [21, 67]. These two terms, *close* and *distant* reading, are also denoted as *micro-* and *macroanalysis* [17].

2.1 Close Reading

According to Jasinski [14], *close reading* helps unveil "words, verbal images, elements of style, sentences, argument patterns, and entire paragraphs" [14, 93]. In this textual analysis, scholars make use of color-coding, underlining and marginal comments. To render close textual analysis digitally, several recent projects have worked with color-coding, font size, glyphs, and connections, for example, Poem Viewer [1], PRISM [25], Juxta [26] and eMargin[1] [13]. Figure 1 offers a close reading of Shakespeare's Julius Caesar performed by eMargin, where words are colored, tagged and commented.

Figure 1: Example of Close Reading of the Shakespeare Play "Julius Caesar": eMargin tool.

2.2 Distant Reading

Distant reading takes a reader from the exhaustive interpretation of individual passages toward the global visualization of text collections. Drawing from quantitative history and geography, Moretti [21] uses graphs, maps, and trees to analyze historical novels. Since the publication of his work, a number of other visual methods have been put forward in literary studies: tag clouds, heat maps, timelines, network graphs as well as geographical maps [13, 7-9]. For example, word clouds have been used to analyze the style of *The Making of America* [6] and Federal Budget Speech of Australia [8], whereas heat maps and network graphs were used to look at the distribution and relationship of literary characters in novels [22].

Furthermore, advances in technology have made it possible to apply more complex quantitative and visual analyses to literary studies: topic modeling, summarization, and cluster classification, among others. Topic modeling identifies short

[1] http://emargin.bcu.ac.uk/

and informative descriptions of each text in a large collection. The main idea of this model is that text collections are "represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" [4, 996]. Topic modeling has been successfully applied to various text genres, e.g. news articles, scientific abstract, scientific papers, digital libraries, and twitter [4, 10, 12]. The common visualization of topics is a list of words associated with each topic and the correlation between topics and documents (see Figure 2). While there exist many tools and environments with topic algorithms, most of them require programming skills. As Blei [3] points out, the developing of interactive user interfaces with topic visualization is a future direction in the topic modeling field. Social and humanities fields also express a need for the use of topic modeling in exploratory literary analysis [18].

	Topic 1	Topic 2	Topic 3
1	irony	impact	czech
2	based	sentiment	english
3	classification	tweets	detection

(a) List of topics and words

```
     [,1] [,2] [,3]
[1,]    1    4    1
[2,]    2    3    2
[3,]    3    2    3
[4,]    4    1    4
```

(b) Topic-document correlation

Figure 2: Common Topic Visualization

The second technique–cluster classification–refers to the automatic algorithm that groups documents into subgroups. These subgroups, or clusters, "are coherent internally, but clearly different from each other" [20]. The common visualization of cluster is a dendrogram, where individual texts are grouped based on agglomerative and distance measures of their similarities, illustrated in Figure 3.

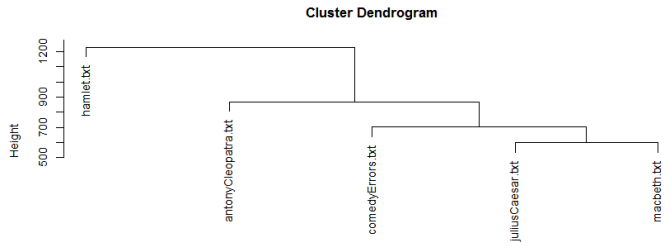


Figure 3: Cluster Dendrogram of Shakespeare's Plays

In a recent survey of close and distant reading visualization tools, Janicke et al. [13] observe a large body of work that combines both types of analysis. Despite the increasing interest in macroanalysis, close interaction between a reader and a text remains essential to humanities scholars. As Cole [7] state, there is "an urgent

desire in the literary community to embrace and explore the power of computation while at the same time prioritizing and protecting the relationship between literature and human readers".

3 Interactive Text Mining Suite

The purpose of *Interactive Text-Mining Suite*[2] (henceforth, ITMS) is to provide a dynamic exploration of text collections, while maintaining interaction between scholars and literary passages. The ITMS is built with R as a back-end and Shiny app as a front-end. In the back-end, Shiny app consists of two R scripts, namely server.R and ui.R. Server.R hosts all functions, whereas ui.R provides a graphical user interface. The use of Shiny web framework for text analysis has several advantages. First, as a web application, ITMS is platform-independent and does not require installation, as compared to other text mining tools. Second, as an R application, ITMS has access to a range of state-of-the-art text analytical, statistical, and graphical packages (e.g., lda, topicmodels, ggplot2, wordcloud, tm). Furthermore, Shiny app is designed to build a highly interactive and user-friendly interface. Finally, the performance of the application is not affected by the local system performance and memory, thus providing more optimal environment for data analysis.

The ITMS aims to bridge the gap between close reading and distant reading. The user has a dynamic bottom-up control of text selection and choices of exploratory analyses. In this approach, researchers can select a specific section of a text, or extract certain segments based on KWIC term selection. In addition, the ITMS allows to upload or extracts metadata (e.g. timestamp, location, language), which can be used for a chronological analysis.

At present, several text processing interactive functions are built into the ITMS, namely, stemming, stopwords, tokenization. At each step, the reader is able to access selected passages in order to decide which processing techniques to use. Finally, the reader can perform various text mining and visual methods. For example, users can analyze word distribution, generate word frequency graphs, perform cluster and topic analyses.

4 Case Study: Visualization of Medieval Romance Flamenca

For our study, we have selected 1000 lines from the annotated corpus of Medieval Occitan *Romance of Flamenca* [24] and its English translation [5]. While traditional visual tools are unable to perform text analysis using annotated corpora, our goal is to combine rich linguistic knowledge from annotated corpus with macro-analysis. First, we can perform a comparison between both documents at a word

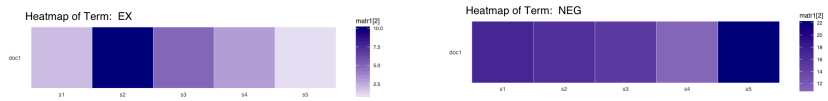
[2] https://languagevariationsuite.shinyapps.io/TextMining/

level using word cloud method. From this analysis, it becomes apparent that verb forms (VJ) dominate in the original text, whereas pronouns (PRP) prevail in its translation. Surprisingly, despite the nature of this novel, common nouns and proper nouns do not seem to be prevalent (see Figure 4).

(a) Occitan corpus (b) English translation

Figure 4: Part-of-Speech Word Cloud Analysis

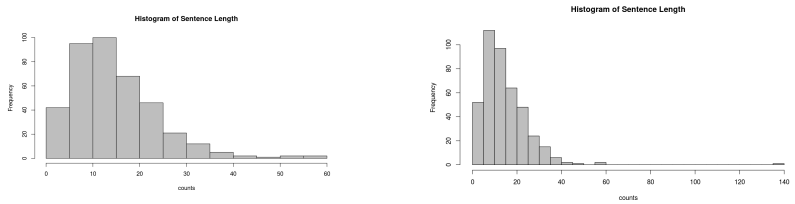
Second, we can examine the use of certain postags by using a keyword-in context search. To illustrate the potential of this method, we have queried our English text for existential (*EX*) and negation (*NEG*) postags. For example, the use of existential (*there is*) is concentrated in the second part of the novel and corresponds to the nuptial preparation (Figure 5a). The close examination of the context shows that this section provides many existential constructions describing the bounty of count Archambaut. On the other hand, the negation (*not*) is present across the entire selection; however, it seems to be more concentrated toward the end, which corresponds to the growing jealousy of Flamenca's husband (Figure 5b).



(a) Existential pos (b) Negation pos

Figure 5: Part-of-Speech Heatmap Analysis

Furthermore, we can examine stylistic similarities and differences at a sentence level. First, the peak of sentence length in both documents seems to be concentrated around 10 words and the overall distribution has a similar shape (see Figure 7). In contrast, there is a dissonance in their usage frequencies.

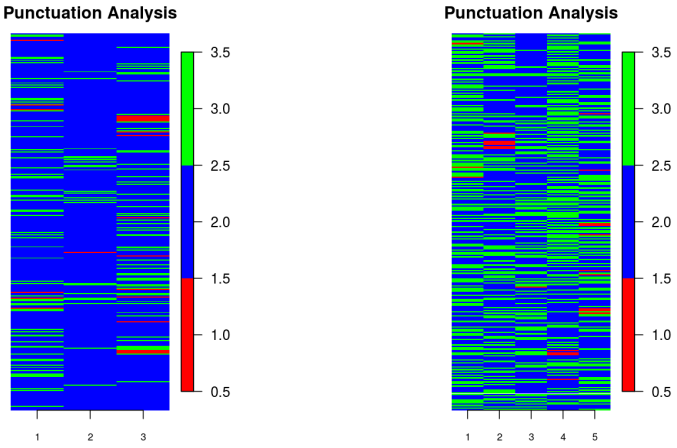


(a) Occitan corpus

(b) English translation

Figure 6: Stylistic Analysis of Sentence Length

Similarly, the choice of punctuation between original and its translation has a noticeable disaccord. Inspired by the Adam Calhoun's punctuation heatmap,[3] we assigned colors to specific types of punctuation in order to detect usage patterns. The heatmap analysis reveals that the original texts contain more quotation marks, hyphens, and parenthesis, as compared to the translated text.



(a) Occitan corpus

(b) English translation

Figure 7: Stylistic Analysis of Punctuation

Finally we can examine this entire novel at a document level by using cluster analysis and topic modeling methods. In order to visually detect similarities in story development, we have split the novel in six section, based on the story plot, namely marriage, jealousy, William's arrival, planning how to meet Flamenca, finding solution, first meeting with Flamenca, and escape from tower. Cluster analysis demonstrates the similarities between William's arrival and William's search for escape solution as well as between their first conversation and Flamenca's escape. Furthermore, topic visualization by means of word cloud help unveil several underlying themes for *love, jealous, Archambaut* and *William, prayer, Flamenca*.

[3]https://medium.com/@neuroecology/punctuation-in-novels-8f316d542ec4

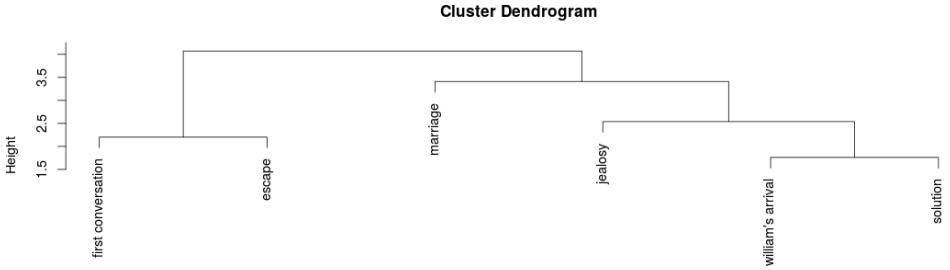


Figure 8: Cluster Analysis

Figure 9: Topic Analysis

5 Conclusion

In recent years, we have seen growing interest in the construction of global features and visual abstract models of text collections. Many scholars, however, have expressed the need for a more integrated approach–the "synthesis of computational and humanistic modes of inquiry" [19]. To incorporate this approach, the authors of this article have proposed to develop a *bottom-up* application for textual analysis and visualization. The current project, *Interactive Text Mining Suite*, aims to provide interactive control for text preprocessing and analysis. This method assists with a more meaningful and fine-grained exploration of corpus. Given the multifaceted nature of the genres of literary research, we have also designed our graphical user interface to reflect choice of studies: scholarly articles, literary genre, bibliographical metadata, and annotated corpora. Finally, the accessibility of our web application facilitates data analysis, as researchers are not constrained by memory limitation or platform dependency.

There are several developments that we see in the future for our project. Given its design flexibility and back-end structure written in R, this toolkit can be easily augmented with additional features. For example, our exploratory analysis can be enhanced with dynamic network graphs and dynamic diachronic mapping (e.g.

igraph and GoogleViz packages). Another development can be stylometric analysis provided by a recent R package *stylo*,[4] such as genre and authorship identification.

References

[1] A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen. Rule-based Visual Mappings - With a Case Study on Poetry Visualization. *Computer Graphics Forum*, 32(3 PART4):381–390, 2013.

[2] David M Berry. The computational turn: Thinking about the digital humanities. *Culture Machine*, 12:1–22, 2011.

[3] David Blei. Probabilistic topic models. *Communcations of the ACM*, 55(4):77–84, 2012.

[4] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[5] E.D. Blodgett. *The Romance of Flamenca*. Garland, New York, 1995.

[6] Tanya Clement, Catherine Plaisant, and Romain Vuillemot. The Story of One: Humanity scholarship with visualization and text analysis. *Relation*, 10(1.43):84–85, 2009.

[7] Katherine Coles and Julie Gonnering Lein. Solitary mind, collaborative mind: Close reading and interdisciplinary research. 2013.

[8] Stephen Dann. Analysis of the 2008 federal budget speech: Policy, politicking and marketing messages, 2008.

[9] Gerald Graff. *Professing Literature: An Institutional History*. University of Chicago Press, 1989.

[10] Thomas L Griffiths and Mark Steyvers. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 pages 5228–35, 2004.

[11] Susan Hockey. The history of humanities computing. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A companion to Digital Humanities*, pages 3–19. Blackwell Publishing, Oxford, 2004.

[12] Liangjie Hong and Brian D. Davison. Empirical Study of Topic Modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88, 2010.

[4] The authors would like to thank an anonymous reviewer for this suggestion.

[13] Stefan Jänicke, Greta Franzini, Muhammad F. Cheema, and Gerik Scheuermann. On Close and Distant Reading in Digital Humanities : A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis) (2015)*, pages 1–21, 2015.

[14] James Jasinski. *Sourcebook on Rhetoric*. SAGE Publications, 2001.

[15] Paul Jay. *The Humanities "Crisis" and the Future of Literary Studies*. Palgrave Macmillan US, 2014.

[16] Kim Jensen. Linguistics in the Digital Humanities: (Computational) Corpus Linguistics. *MedieKultur: Journal of media and communication research*, 30(57), 2014.

[17] Matthew L. Jockers. *Topics in the Digital Humanities: Macroanalysis : Digital Methods and Literary History*. University of Illinois Press, Urbana, IL, USA, 2013.

[18] Lauren Klein and Jacob Eisenstein. Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication*, 4(3), 2013.

[19] Lauren Klein and Jacob Eisenstein. Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives, 2013.

[20] Christopher. Manning. *An introduction to Information Retrieval*. 2009.

[21] Franco Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

[22] Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. Advanced visual analytics methods for literature analysis. *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 35–44, 2012.

[23] Jeffrey Schnapp and Peter Presner. *Digital Humanities Manifesto 2.0*. 2009.

[24] Olga Scrivner, Sandra Kübler, Barbara Vance, and Eric Beuerlein. Le Roman de Flamenca: An Annotated Corpus of Old Occitan. In *the 3rd Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, 2013.

[25] Brandon Walsh, Claire Maiers, Gwen Nally, Jeremy Boggs, and P.P. Team. Crowdsourcing individual interpretations: Between microtasking and macrotasking. *Literary and Linguistic Computing*, 29(3):379–386, 2014.

[26] Dana Wheeles and Kristin Jensen. Juxta commons. In *the Digital Humanities 2013*, 2013.