Advanced Visual Analytics Methods for Literature Analysis

Daniela Oelke
University of Konstanz
Data Analysis and Visualization
Konstanz, Germany
oelke@inf.uni-konstanz.de

Dimitrios Kokkinakis
Språkbanken
Department of Swedish
University of Gothenburg
Gothenburg, Sweden
dimitrios.kokkinakis@gu.se

Mats Malm
Department of Literature,
History of Ideas and Religion
University of Gothenburg
Gothenburg, Sweden
mats.malm@lit.gu.se

Abstract

The volumes of digitized literary collections in various languages increase at a rapid pace, which results also in a growing demand for computational support to analyze such linguistic data. This paper combines robust text analysis with advanced visual analytics and brings a new set of tools to literature analysis. Visual analytics techniques can offer new and unexpected insights and knowledge to the literary scholar. We analyzed a small subset of a large literary collection, the Swedish Literature Bank, by focusing on the extraction of persons' names, their gender and their normalized, linked form, including mentions of theistic beings (e.g., Gods' names and mythological figures), and examined their appearance over the course of the novel. A case study based on 13 novels, from the aforementioned collection, shows a number of interesting applications of visual analytics methods to literature problems, where named entities can play a prominent role, demonstrating the advantage of visual literature analysis. Our work is inspired by the notion of *distant reading* or *macroanalysis* for the analyses of large literature collections.

1 Introduction

Literature can be studied in a number of different ways and from many different perspectives, but text analysis - in a wide sense - will surely always make up a central component of literature studies. If such analysis can be integrated with advanced visual methods and fed back to the daily work of the literature researcher, then it is likely to reveal the presence of useful and nuanced insights into the complex daily lives, ideas and beliefs of the main characters found in many of the literary works. Therefore, the names of all characters appearing in literary texts can be one such line of enquiry, which is both an important subfield of literature studies (*literary onomastics*) and at the same time the result obtained by a mature language technology (*named entity recognition*) which can be turned into a tool in aid of text analysis in this field. (Flanders et al., 1998) discuss that references to one type of names, namely that of people, are of intrinsic interest because they reveal networks of friendship, enmity, and collaboration; familial relationships; and political alliances. People's names can be an appropriate starting point for research on biographical, historical, or literary issues, as well as being a key linguistic and textual feature in its permutations and usage.

We argue that the integration of text analysis and visualization techniques, which have turned out to be useful in other scientific fields such as bioinformatics (Nature Methods, 2010), could be put to effective use also in literature studies. We also see an opportunity to devise new ways of exploring the large volumes of literary texts being made available through national cultural heritage digitization projects.

Digitized information and the task of storing, generating and mining an ever greater volume of (textual) data becomes simpler and more efficient with every passing day. Along with this opportunity, however, comes a further challenge: to create the means whereby one can tap this great potentiality and engage it for the advancement of (scientific) understanding and knowledge mining.

Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 35–44,
Avignon, France, 24 April 2012.

We apply a *supra-textual* perspective to the analysis of literary texts by encompassing a rather global visualization of a document. As a case study, we have analyzed a subset, 13 novels, of the Swedish Literature Bank[1] collection through two levels of inquiry by focusing on person names, their gender and their normalized, linked form, including mentions of theistic beings (e.g. Gods' names and mythological characters), and examining their appearance in sentences, paragraphs and chapters.

Our aim is to explore the usage of alternative visualization means that provide additional insight by showing the data at higher resolution levels or that permit an analysis of the development of the story in the course of the text. The employed visualization techniques are scalable enough to display several novels at once and therefore allow a literature scholar to compare different literary texts to each other. By combining advanced natural language processing techniques with visualization techniques, we aim to allow the user to rapidly focus on key areas of interest (based on name mentions) and provide the ability to discover e.g. semantic patterns in large collections of text. Therefore, our work is based on individual texts, by looking for certain patterns of variation based on a particular named entity type. Our work is also inspired by the notions of *distant reading* or *macroanalysis* applied to the analyses of literature collections which we find appealing for the research we describe. However, we do not consider such techniques to be used as a substitution for reading a book sequentially but as a useful supplement.

[1]The Swedish Literature Bank (*Litteraturbanken*, `http://litteraturbanken.se`) is a co-operation between the Swedish Academy, the Royal Library of Sweden, the Royal Swedish Academy of Letters, History and Antiquities, the Language Bank of the University of Gothenburg, the Swedish Society for Belles Lettres, and the Society of Swedish Literature in Finland. The Swedish Literature Bank also focuses on neglected authors and genres, effectively establishing a set of 'minor classics' alongside the canonical works. So far, mainly texts in Swedish are available, but over time, selected works will be offered in translation as well. Currently, the Swedish Literature Bank offers literary works either as searchable e-text, as facsimiles of the original edition, as PDF files or as EPUB files - often in more than one format. The texts are available free of charge and the software is developed as open source. The website is directed towards the general public and students and teachers at every level, as well as towards scholars. The digital texts are based on printed first editions or on later scholarly editions. They are carefully proof-read, thus establishing a basis for scholarly work. For the common reader, introductions and essays provide fresh perspectives on the classics.

2 Background

Computer-assisted literary criticism is a rather young field in literature analysis (Juola, 2008). Typically, researchers in literary studies use computers only to collect data that is afterwards analyzed conventionally. Yet, there are some cases in which the computer has already proven useful, e.g., for the analysis of prosody and poetic phonology or for comparing an author's revisions (from version to version). Computer-assisted studies have also been performed in the context of sequence analysis in the past, such as assigning quoted passages to speakers and locating them in the sequence of the text (Butler, 1992).

2.1 Distant Reading and Macroanalysis

(Moretti, 2005) coined the term "distant reading" in which "the reality of the text undergoes a process of deliberate reduction and abstraction". According to this view, understanding literature is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data. This way it becomes possible to detect possible hidden aspects in plots, the structure and interactions of characters becomes easier to follow enabling experimentation and exploration of new uses and development that otherwise would be impossible to conduct, e.g., quantifying the difference between prose styles.

Distant reading or its near synonym *macroanalysis* is a technique to analyze literature, as opposed to "close reading" of a text that is the careful, sustained interpretation of a brief passage of text where great emphasis is placed on the particular over the general, paying close attention to individual words, syntax, and the order in which sentences and ideas unfold as they are read. The most fundamental and important difference in the two approaches/terms is that the macroanalytic approach reveals details about texts that are for all intents and purposes unavailable to close-readers of the texts. Distant reading is in *no* way meant to be a replacement for close readings and in traditional humanities, as Moretti puts it (Schulz, 2011), "distant reading should supplant, not supplement, close reading".

2.2 Visual Analytics for Literature Analysis

Visual Analytics is "the science of analytical reasoning facilitated by visual interactive interfaces" (Thomas et al., 2005). The central idea of visual analytics is that by tightly integrating the human expert and the machine, the strengths of both can be leveraged in the analysis process. Visual Analytics has been applied successfully to many application domains in the past such as text analysis, geographical data analysis, security applications, (computational) biology or multimedia data.[2]

However, visual analytics is not often used in the context of literature analysis. Commonly, a text is read sequentially and then analyzed by the researcher bit by bit. Only during recent years some literary scholars have started to employ visualization techniques in their studies.

One of them is Franco Moretti, who advocated the usage of visual representations such as graphs, maps, and trees for literature analysis (Moretti, 2005). (Vuillemot et al., 2009) suggested the usage of word clouds and self-organizing graphs and presented a tool that allows to analyze a novel interactively with respect to several properties. In (Plaisant et al., 2006) a tabular representation that is enriched with visual symbols was used to present the results of an automatic algorithm for detecting erotic statements. (Rydberg-Cox, 2011) generated social network graphs of characters in Greek tragedies, based on information taken from linguistic dependency treebanks, which permit to visualize the interactions between characters in the plays. Furthermore, scatterplot views allowed the user to search for correlations between several variables of the meta data that comes with the novels. Rohrer et al. (1998) experimented with using implicit surfaces to compare single documents with respect to the most frequent terms and to visualize a document collection.

Pixel-based visualizations come with the advantage that the documents can be analyzed at a higher resolution level. The Compus system (Fekete and Dufournaud, 2000) used dense pixel displays to visualize the structure of richly annotated XML documents of books of the 16^{th} century. Keim and Oelke (2007) focused more on the analysis of documents with respect to certain text properties to compare authors with respect to their writing style or to learn more about the characteristics of a literary book. The two techniques also differ from each other in terms of how structural information is encoded and how they deal with the problem of overplotting that occurs if a pixel encodes several feature values.

[2] *Cf.* proceedings of the IEEE Conference on Visual Analytics Science and Technology (IEEE VAST), http://visweek.org/.

3 Named Entity Recognition

Named entity recognition (NER) is an important supporting technology with many applications in various human language technologies. It has emerged in the context of *information extraction* (IE) and *text mining* (TM). The automatic recognition and marking-up of names (in a wide sense) and some other related kinds of information - e.g., time and measure expressions and/or terminology - has turned out to be a recurring basic requirement. Hence, NER has become core language technology of great significance to numerous applications and a wide range of techniques (Jackson and Moulinier, 2007).

In our study involving 19th century fiction, we use a slightly adapted NER system to the language used in fiction around the turn of the twentieth century (Borin and Kokkinakis, 2010). Moreover, the nature and type of named entities vary, depending on the task under investigation or the target application. In any case, *person*, *location* and *organization names* are considered 'generic'. The system we applied implements a rather fine-grained named entity taxonomy with several main named entity types and subtypes but for our case study we chose to only use the type *person* which usually incorporates people's names (forenames, surnames), groups of people, animal/pet names, mythological names, theonyms and the like. Note that we haven't performed any formal evaluation of the entity or the gender annotation in this work. In previous studies, based on data from the same source and applying the same NER-tools (Borin et al., 2007), we have shown high figures on precision and recall (96-98%) on, particularly, person recognition.

3.1 Gender Attribution

Current NER systems are limited to the recognition of a small set of entity types without attempting to make finer distinctions between them. The system we use goes beyond this in the sense that it attempts to also automatically determine

the referential gender of all person entities. Referential gender relates linguistic expressions, both persons and groups of individuals, to "female", "male" or "gender-indefinite". This is an important constraint which contributes to better performance in subsequent language processing applications based on NER, such as anaphora resolution, by filtering-out of gender-incompatible candidates (Evans and Orasan, 2000). The approach to gender discrimination is based on applying a combination of the following heuristics:

- NER has a high accuracy in identifying person names, a large number of which are assigned gender. A pre-classified list of 16,000 common first names assigns gender to commonly used first names. This way a first distinction is already being made between entities that carry gender. The list has been acquired from various internet sites.

- Use of gender-marked pronouns in the vicinity of person entities (a simplistic form of pronoun resolution where simple decisions are made by matching a genderless person entity with a gender bearing personal pronouns, *han* 'he', *hans* 'his', *hon* 'she' and *hennes* 'her'). Also, various types of honorifics and designators, manually pre-categorized into gender groups, provide the evidence that is explored for the annotation of both animate instances but also their gender. Inherent characteristics for a large group of these designators (e.g., morphological cues), indicate biological gender. Examples of gender-bearing male designators are e.g. the nouns *baron* and *herr* 'Mr', and adjectives with suffix bearing gender, namely *-e*, such as *starke* 'strong', *hygglige* 'kind' and *gamle* 'old'; while female-bearing designators are e.g. *tant* 'aunt' and *fru* 'wife'. Gender is also captured using a simplified set of suffix matching rules, such as *-inna/innan*, *-erska/erskan* (typical suffixes for female) and *-man/mannen*, *-iker/ikern* (typical suffixes for male).

- Labeling consistency is a technique that operates over the whole annotated text. This module reviews the annotations made so far, in order to support gender attribution of unassigned cases based on unambiguous previous gender assignments. This is a simple but robust approach that does not rely on pre-compiled statistics of any kind. In order to capture such consistency we employ a two stage labeling approach. During the first stage, we note the instances of person entities with unknown gender, and search for a context where the same entity has been assigned gender (male, female) due to a gender-indicating context and for which no other occurrences of the same entity are found in the document with a different gender. If this is the case, then all occurrences of that entity are assigned the same gender throughout the document. During the second stage, the system investigates if there are any conflicting, ambiguous annotations for gender for which the local context and the supporting resources (e.g., first names' gazetteer) cannot decide the gender attribution. If this is the case and more than one possible annotation for gender is recorded, we choose the most frequently assigned gender label for the entity in question, in case of a tie we mark the gender as *unknown*.

3.2 Name Linking

Since the same name can be referred to in various ways, extracting named entities alone is not sufficient for many tasks. Therefore, mapping and linking multiple linguistic variations to a single referent is necessary. We apply a simplified form of co-reference resolution based on salient features and pattern matching that links (hopefully) all mentions that refer to a single person entity. Consider the aggregated occurrences for the name *O'Henny* appearing in the novel "Clownen Jac" [lb904603] (1930). All 92 occurrences of the figure *O'Henny* will be linked to the same individual since there is sufficient and reliable evidence which is based on gender match, no annotation conflicts (i.e. other individual named *Denny* or *Henny* with the same gender) and orthographic characteristics: *O'Henny* (58); *Denny* (19); *Denny O'Henny* (7); *Henny-Denny* (4); *Denny-Henny* (3); *Henny* (1).

4 Material

Prose fiction is just one type of textual material that has been brought into the electronic "life" using large scale digitized efforts. But it must be

considered an essential source within many disciplines of humanities (history, religion, sociology, linguistics etc.) and social studies and an invaluable source for understanding the movements of society by its ability to demonstrate what forces and ideas are at work in the society of its time. Prose fiction is complex and difficult to use not only because of interpretational complexity but also because of its limited availability.

The Swedish Literature Bank, and its sister project "the 19th Century Sweden in the Mirror of Prose Fiction", aims to change this by developing a large representative corpus which mirrors society at given points in time, chronologically selected in such a way that historical comparisons can be made. A substantial part of the material is all fiction, written in the original and published separately for the first time, that appeared in Swedish starting from the year 1800 and collected during consecutive twenty year intervals. The material provides a whole century of evolution and social, aesthetic, scientific, technical, cultural, religious and philosophical change. Out of this data we selected the literary production, 13 novels, of a single author, namely Hjalmar Bergman (1883-1931). The selected novels (followed by their id) are:

- Savonarola (1909); id=lb443177
- Amourer (1910); id=lb1611717
- Hans nåds testamente (1910); id=lb1611719
- Vi Bookar, Krokar och Rothar (1912); id=lb494265
- Loewenhistorier (1913); id=lb1631349
- Falska papper (1916); id=lb1525006
- Herr von Hancken (1920); id=lb1524996
- Farmor och Vår Herre (1921); id=lb1187656
- Eros' begravning (1922); id=lb1470072
- Chefen fru Ingeborg (1924); id=lb1524995
- Flickan i frack (1925); id=lb1470073
- Kerrmans i paradiset (1927); id=lb1317426
- Clownen Jac (1930); id=lb904603

5 Visual Exploration of the Data

In this chapter we report on our experiences with different visualization techniques that can be employed for analyzing novels with respect to the characters involved in the plot. Besides network representations two alternative, not as well known, visualization techniques are tested. Our goal is to learn about their strengths and weaknesses with respect to the task and identify challenges that are specific for the field. We show how visualization can be used to gain insight into literary work that otherwise would be much more laborious to get.

5.1 Network representation

Traditionally, persons in a novel are analyzed in terms of the relations that exist between them. Obviously, graph visualizations are well suited for representing this kind of information. Figure 1 shows a person network for the novel "Eros' begravning" ('Eros' funeral') (1922). Nodes represent characters of the plot and an edge is inserted between two persons if they co-occur in at least one sentence of the novel.[3] In such a representation it is easy to identify protagonists that are connected to many other characters (e.g., *Ludwig von Battwyhl* or *Olga Willman-Janselius*). Furthermore, it is possible to see clusters of characters. Figure 1 also shows that *Casimir Brut* is the person that connects the two main groups of characters of the novel, in the sense that he introduces one group of characters to another. The thickness of an edge encodes the number of times that two names co-occur which could be regarded as the strength of the relationship. A strong connection seems to exist between *Brita Djurling* and *Ludwig von Battwyhl* but also between *Hans Hinz Faber* and *Gruber*. It is interesting to see that *Gruber* is only weakly connected with other characters of the plot but almost exclusively occurs together with *Hans Hinz Faber*. Presumably, because *Hans Hinz Faber* was the faithful servant of *Gruber*.

The example shows that network representations can provide interesting insight with respect to the relationship between different persons in the plot. However, one question that this plot cannot answer is how these relationships evolve over the course of the novel.

5.2 Summary Plots

Summary plots are tabular representations in which each column represents a text unit (here:

[3]Note that using co-occurrence can be just considered an approximation. More advanced methods would be needed to ensure that all and only well-established relationships between characters are extracted.

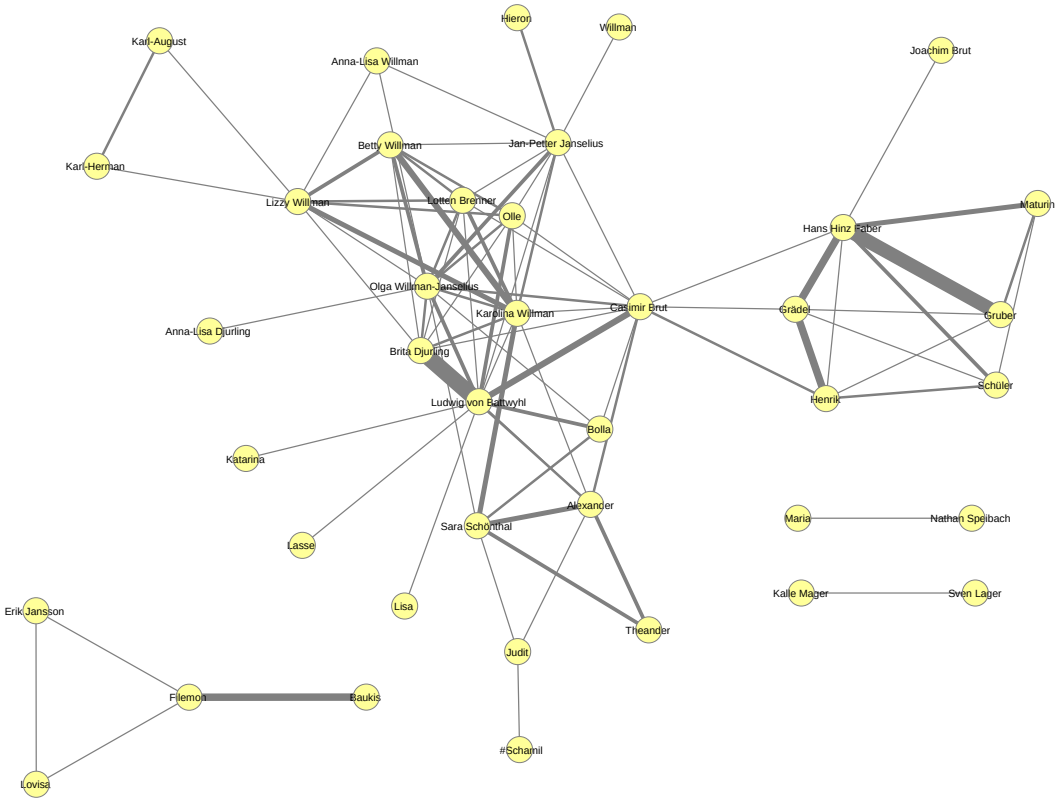

Figure 1: Network representation based on the co-occurrences of the person names in ”Eros’ begravning”

a chapter) and each line corresponds to a person of the novel. Table cells are colored according to the frequency of the character in the specific text unit (normalized with respect to the number of words in the chapter). The person names are sorted in descending order according to the overall frequency of the person name in the novel.

In such a heatmap-like representation it is easy to see which characters co-occur in a chapter but also how this develops in the course of the document. *Do always the same persons meet? Is there one main protagonist in the book that is almost always present or is the story line more complex in terms of characters?* Being able to answer this kind of questions provides the analyst with insight about the development of the story that would not be visible in a person network.

Figure 2 shows the summary plot for the novel ”Eros’ begravning” in which some interesting characteristics become apparent. For example, some person names are only mentioned in a specific chapter (see lines of *Hans Hinz Faber*, *Grädel*, *Schmil*, *Lisbeth* etc.). Besides, the chapters differ significantly with respect to the number of unique person names that are mentioned. The first and the last chapter are the ones in which most characters are mentioned whereas in the third chapter only four characters play a role.

A closer look into the text reveals that the novel consists of a ”frame text”, where different people meet and tell each other stories. The stories constitute chapters in the novel, and thus become a bit like short stories. The first chapter, which does not have a title, introduces a large number of people. This number of participating people then decreases during the course of the following stories (chapters), but towards the end of each chapter the discussion is returned to the overall story once again, where people are talking with each other about various things before the next story starts. Also, in the individual chapters there exist people who do not participate outside of a single chapter.

5.3 Literature Fingerprints

Summary plots allow literature scholars to see which characters co-occur in one chapter. However, they do not permit to analyze the usage of the person names within one chapter. In contrast to this, pixel-based visualizations avoid such ag-

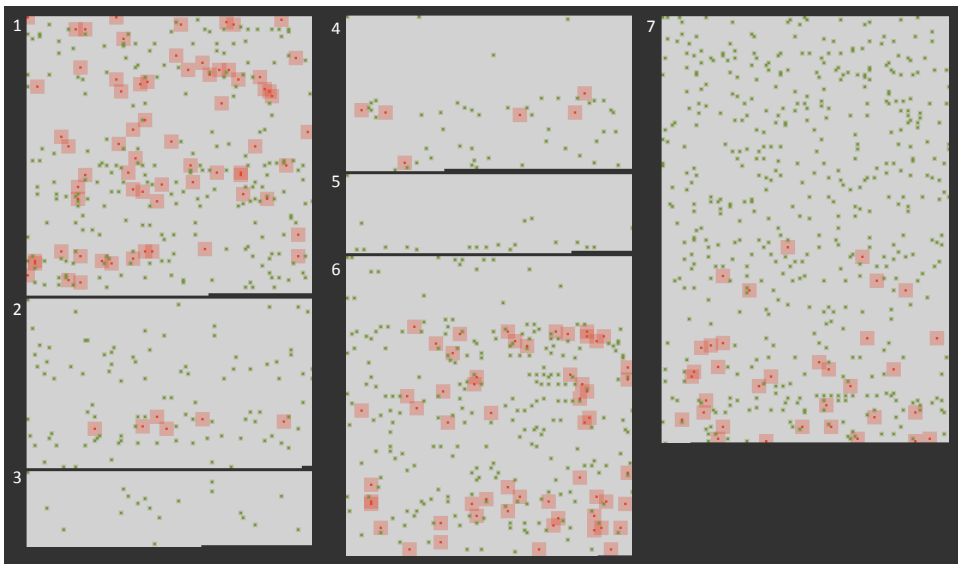

Figure 3: Literature Fingerprint for the novel ”Eros’ begravning”. Red pixels mark mentions of the protagonist ”Olga Willman-Janselius, green pixels highlight the position of other names.

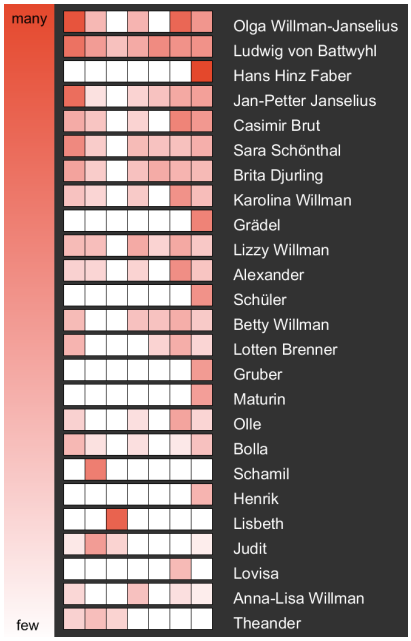

Figure 2: Summary plot for novel ”Eros’ begravning” (’Eros’ funeral’).

gregation and enable an inspection of the data on a much higher resolution level.

We use the literature fingerprinting technique (Keim and Oelke, 2007) to inspect the novel ”Eros’ begravning” in more detail. Each pixel represents one word. Pixels are arranged from left to right and top to bottom and are grouped according to chapters. The color of a pixel can be used to encode a value. In this case pixels were colored in red if they represent the name of the most frequent protagonist, *Olga Willman-Janselius*, and in green if another name was mentioned. The technique is scalable enough to display the whole book at this high resolution level. However, the colored pixels are sparse and would likely be lost in the sea of uncolored pixels. We therefore use semi-transparent halos around the colored pixels to increase their visual saliency. (For more visual boosting techniques for pixel-based visualizations see (Oelke et al., 2011)). In this visualization it is now possible to see where in the course of the novel the main protagonist, *Olga Willman-Janselius*, plays a role. Furthermore, it becomes obvious that there are parts in which almost no person name at all is mentioned. This is in line with the fact that the book tells several separate stories that are integrated at the end of each chapter into the overall story (see also explanation in section 5.2).

Alternatively, we also could have highlighted the positions of several names using one color per protagonist to compare their distribution. This way an analyst can learn about the relations between different characters. However, the number of different names that can be highlighted at the same time is restricted by the human ability to distinguish different colors easily (*cf.* (Ware, 2008)).

Figure 4 shows fingerprints for all 13 novels. Again each pixel represents a word but this time all words that neither are a name of a person nor of a theistic being are disregarded. This way a

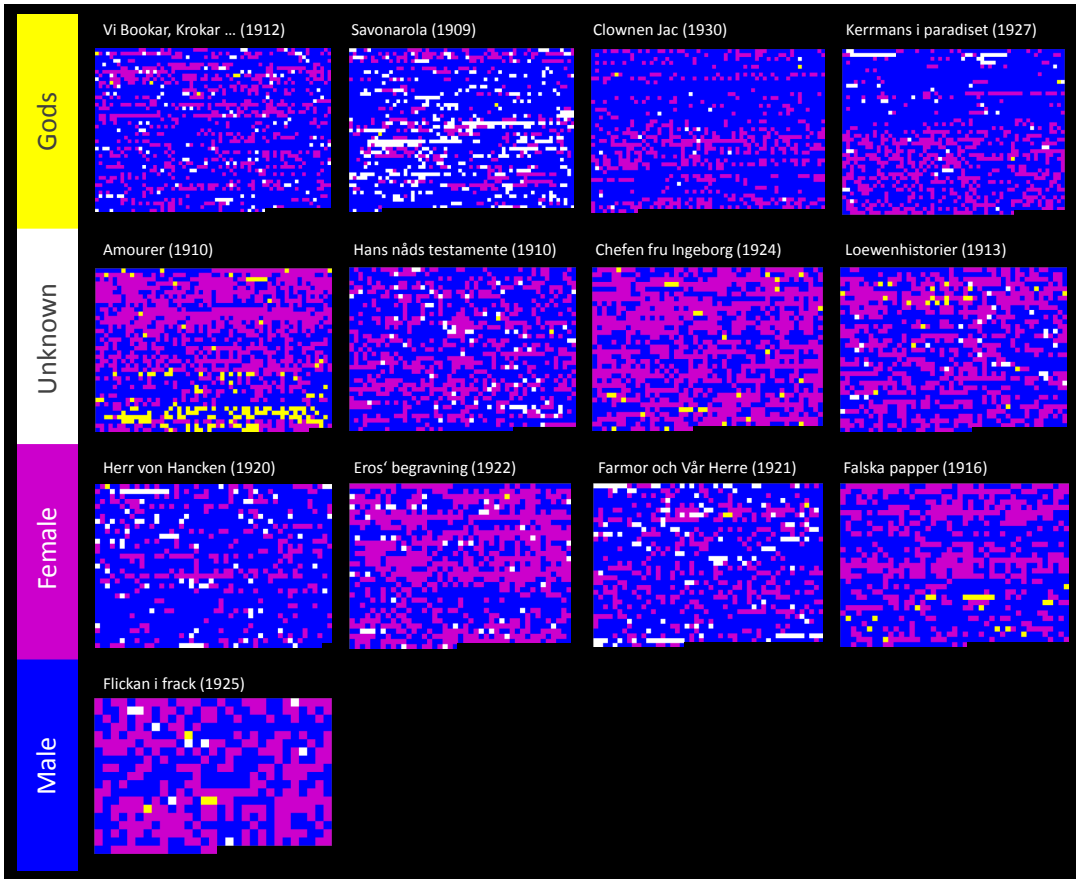

Figure 4: Fingerprints for the 13 novels. Color is used to encode the three categories male, female, gods.

focus is put on the order in which the mentions of the three categories (male, female, gods) appear. Words that the algorithm recognized as a name but could not assign to one of the categories are marked as unknown and are colored in white.

Some interesting patterns become visible in the visualization. One book (first one in the second row) sticks out because of its high number of mentions of theistic beings. "Amourer" [lb1611717] (1910) is a collection of short stories. The last story, "The False Cristoforo", varies the theme of *Christopher*, who carried Jesus Christ across the river which results in the peak of names of theistic beings that can be observed at the end of the book.

Another interesting observation is that in the beginning of the book "Kerrmans i paradiset" [lb1317426] (1927) (last one in first row), male characters are clearly dominant which is almost reversed in the book's second part. A closer look into the book reveals that this is because the book is divided into two main parts. The first part is more about prestige and position in society, i.e., social games with other men, while the second part is more personal and relates clearly to women. The summary plot of the book (Figure 5) reveals that there are not fewer male characters involved in the second part of the book but overall they are less frequently mentioned. At the same time, female characters that had in the first part of the book only a minor role become more dominant in the plot.

5.4 Discussion

Each of the visualization techniques that we experimented with has its strengths and weaknesses if used for the analysis of a novel with respect to its characters. Person networks come with the advantage that they can show relationships between characters. This way clusters of persons that form a group within the story become visible. In contrast to this, summary plots can only show co-occurrence within a chapter (or smaller text unit). But their strength is to show the development of the set of persons involved in the plot. In such a tabular representation it is easy to compare the in-

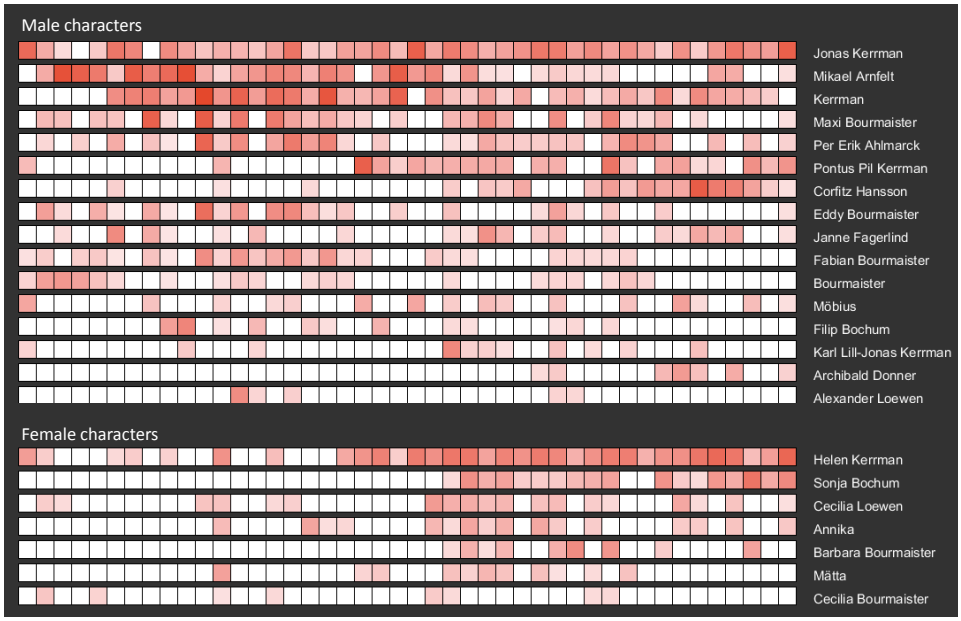

Figure 5: Summary plot for the novel ”Kerrmans i paradiset”. Lines are grouped according to the two categories male / female and are sorted within each category according to the overall frequency of the characters.

volvement of different characters across the document. Even more details are provided by the literature fingerprinting technique. Because the technique is very scalable, every single word can be visualized. Coloring is used to encode text properties of interest. Again, the development of the characters across a document is visible, this time even within single chapters. However, compared to the summary plot technique, fewer person names can be distinguished.

Obviously, a combination of the three techniques is advisable for analyzing novels with respect to the persons involved. But our comparison of the three techniques also allows us to identify a missing type of visualization: One that is able to show the development of the story in terms of the characters involved and at the same time is able to display their relationships.
Furthermore, the techniques lead to interesting insight but these newly generated hypotheses need to be checked in the text. A tighter integration of the actual text source into the visualization tools could therefore be a valuable extension.

6 Conclusions

The combination of robust text analysis with visual analytics brings a new set of tools to literature analysis, provides powerful insights on document collections, and advances our understanding of the evolution of human behavior, society, technological advancement and cultural trends. As a matter of fact, (Michel, 2010), introduced the term ”Culturomics”, i.e. the application of high-throughput data collection, digital book archives and the like, and analysis to the study of human culture and we believe that novel insights towards this direction can be gained by combining such technologies. In this paper we have shown that quantifiable data such as (person) names can be identified, extracted, and visualized in novel ways.

In the future we intend to further extend the capabilities for visual literature analysis. One research goal is the development of a visualization technique that allows to investigate the development of a story across a novel but at the same time shows the relationships between the characters. Furthermore, we believe that interactive visual analysis tools (instead of static visualizations) open up additional possibilities for literature scholars to explore the large volumes of digitized literary collections that are nowadays available.

Acknowledgments

This work was supported by the Zukunftskolleg of the University of Konstanz and the Centre of Language Technology in Gothenburg.

References

Yevgeni Berzak, Michal Richter, Carsten Ehrler and Todd Shore. 2011. Information Retrieval and Visualization for the Historical Domain. *Language Technology for Cultural Heritage - Theory and Applications of Natural Language Processing.* Pp. 197–212. Springer.

Lars Borin and Dimitrios Kokkinakis. 2010. Literary Onomastics and Language Technology. *Literary Education and Digital Learning. Methods and Technologies for Humanities Studies.* Pp. 53–78. IGI Global.

Lars Borin, Dimitrios Kokkinakis and Leif-Jran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. *Proceedings of the ACL Workshop: Language Technology for Cultural Heritage Data (LaTeCh)*. Prague. Pp. 1–8.

Christopher S. Butler. 1992. *Computers and Written Texts*. Basil Blackwell.

Richard Evans and Constantin Orasan. 2000. Improving anaphora resolution by identifying animate entities in texts. *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC) 2000*. Lancaster, UK. Pp. 154–162.

Jean-Daniel Fekete and Nicole Dufournaud. 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*. San Antonio, Texas, United States. Pp. 47–55, ACM.

Julia Flanders, Syd Bauman, Paul Caton and Mavis Cournane. 1998. Names proper and improper: Applying the TEI to the classification of proper nouns. *Computers and the Humanities*. 31(4), pp. 285–300.

Peter Jackson and Isabelle Moulinier. 2007. *Natural language processing for online applications: Text retrieval, extraction and categorization*. Amsterdam: John Benjamins.

Patrick Juola. 2008. Killer applications in digital humanities. *Literary and Linguistic Computing*. 23(1): 73–83.

Daniel A. Keim and Daniela Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*. Pp. 115–122.

Jean-Baptiste Michel *et al.* 2010. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331 (6014): 176. "http://www.sciencemag.org/content/early/2010/12/15/science.1199644".

Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. R. R. Donnelley & Sons.

Nature Methods. 2010. Visualizing biological data. *Supplement to Nature Publishing Group journals.* 7 (3s): S1-S68.

Daniela Oelke, Halldor Janetzko, Svenja Simon, Klaus Neuhaus and Daniel A. Keim. 2011. Visual Boosting in Pixel-based Visualizations. *Computer Graphics Forum.* 30 (3): 871-880.

Catherine Plaisant, James Rose, Bei Yu, Loretta Auvil, Matthew G. Kirschenbaum, Martha Nell Smith, Tanya Clement and Greg Lord. 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*. Pp. 141-150, ACM.

Randall M. Rohrer, David S. Ebert, and John L. Sibert. 1998. The Shape of Shakespeare: Visualizing Text using Implicit Surfaces. *Proceedings of the 1998 IEEE Symposium on Information Visualization*. Pp. 121-129.

Jeff Rydberg-Cox. 2011. Social Networks and the Language of Greek Tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science.* 1(3): 1-11.

Kathryn Schulz. 2011. The Mechanic Muse - What Is Distant Reading? The New York Times - Sunday Book Review. Page BR14. "http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html".

James J. Thomas and Kristin A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* National Visualization and Analytics Center.

Romain Vuillemot, Tanya Clement, Catherine Plaisant and Amit Kumar. 2009. What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections. *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*. Atlantic City, New Jersey, USA. Pp. 107–114.

Colin Ware. 2008. *Visual Thinking for Design*. Morgan Kaufmann.