Lucie Flekova / Florian Stoffel / Iryna Gurevych / Daniel Keim

Content-based Analysis and Visualization of Story Complexity

Abstract Obtaining insights into the style and content characteristics of a novel can provide a benefit to a large number of users. Parents and teachers may be interested in finding appropriate books for children. Booksellers may want to assess the fit of a candidate's artwork into their portfolio or determine the target audience for their promotion activities. Literature scholars might discover particular stylistic similarities in writing patterns of different authors. For all of the above, manually reviewing the textual content of the books is a tedious and time-consuming task which can be achieved only to a limited level of detail. The combination of automated data analysis of literature and computer-based visualization techniques proves to be powerful in giving a quick overview as well as providing details of the visualized data.
In this chapter we define the umbrella term *Story Complexity,* and outline the text data analysis required to describe properties of literature contributing to the numerous aspects of this term. We introduce a multi-faceted *model of story complexity* by addressing numerous aspects of writing, which can pose difficulties to human readers attempting to follow a storyline in fictional literature. Approximations of these aspects are computed automatically with state of the art Natural Language Processing methods. We present the corresponding text data analysis methods, as well as giving examples of how the extracted data can be presented visually, so that the results of the data analysis can be perceived more effectively than by examining the extracted properties of text in a numeric way.

1. Introduction

Gaining an overview of a novel in terms of which aspects contribute to the difficulty, or ease, of following its story can be of benefit for multiple user groups. For example, a teacher may be interested in choosing appropriate reading material for the school class and can do so by considering the level of the language used, as well as the number of characters and parallel storylines, and the complexity and appropriateness of each character's behavior. Alternatively, an e-book merchant may consider acquiring a new series of books to their portfolio and

Published in: Noah Bubenhofer, Marc Kupietz: Visualisierung sprachlicher Daten. Heidelberg: Heidelberg University Publishing, 2018
DOI: https://doi.org/10.17885/heiup.345.474

wants to understand how well, based on previous demand, its writing style and content matches their existing customer base. A literature scholar, on the other hand, may be interested in a contrastive analysis of typical patterns in stories written by different authors or in different literary epochs. A common aspect for all these user scenarios is that the users are not necessarily interested in reading each of the novels in detail, but rather have an information need for an aggregated insight.

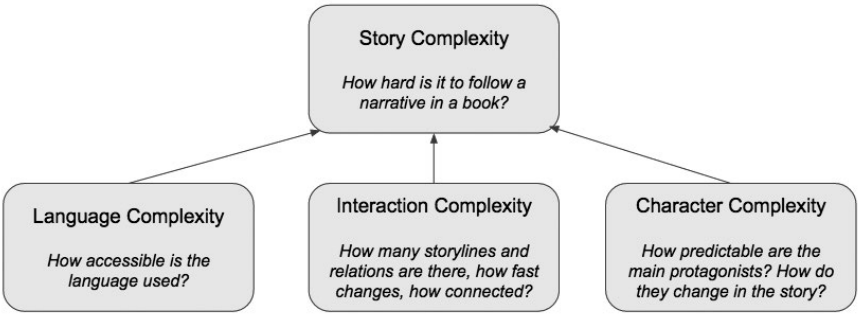

Figure 1: Our understanding of the story complexity and its underlying dimensions.

We approach Story complexity as a covering term that includes numerous aspects which we examine in individual sections of this chapter. In our model, we divide the complexity intuition into three broad areas, as illustrated in Figure 1. The first area is the complexity of the language used. This includes the lexical choice of an author (e.g. if many rare or expert terms are used, increasing the likelihood that the reader won't be familiar with them), the ease of reading on a surface level (e.g. usage of extremely long sentences or words), as well as some syntactic choices such as the dependency types or part-of-speech preferences (e.g. simplification of perception descriptions through interjections).
The second area is the complexity of character interactions. Does the story feature multiple main protagonists? Does it evolve at multiple locations or time periods in parallel? Are the characters in the parallel storylines interconnected in a complex way? How fast do the switches between storylines happen? Is the story linear in time?

The third area of complexity is the behavior of each of the main characters per se. Is the character flat and predictable or does it have both bright and dark sides? Does it develop dynamically as the story progresses? Does it repeatedly show the same emotions?

The overview of which feature types we use to analyze text for each of the three complexity areas is illustrated in Table 1. The list of features should not be treated as exhaustive - there are multiple other options, which can contribute towards understanding the given properties. In this chapter, we aim to provide a broad overview of the possibilities for analyzing a fictional story. A reader can take inspiration from these and expand this framework to meet their own specific analytical needs.

Table 1: Overview of the features we use in each of the complexity areas.

Complexity type	**Natural Language Processing methods used**	**Visualization methods used**
Language complexity	Readability measures, proportion of rare words, proportion of foreign words, density of entities, proportion of part of speech types, occurrence of topics discussed (topic lists, LIWC lexicon), emotions (NRC Emotion Lexicon), sentiment (Stanford Sentiment Analyzer), high-level verb and noun types (WordNet lexicographer files)	paragraph squares bar charts aligned bar charts flow charts
Interaction complexity	Named entity recognition and classification (person, location, organization), entity graph (strength of relationship metrics)	co-occurrence matrices
Character complexity	Speaker identification, emotion development analysis, sentiment development analysis, WordNet lexicographer files for individual character's activities	bar charts flow charts

The content of this chapter is structured as follows. First, we describe previous work related to story complexity in general and previous work related to story visualization (Section 2). Then we discuss each of the three complexity

dimensions, specific work related to these, and how to implement text processing and visualization for each of these dimensions: Section 3 discusses the complexity of the language, Section 4 the complexity of the plot, and Section 5 the complexity of an individual character. In Section 6 we draw conclusions from our work and discuss possible future directions.

2. Related work

This section describes previous work generally related to analyzing story complexity. Specific work related to individual complexity dimensions is discussed at the beginning of each of the following sections.

2.1 Related work in language research and human sciences

In language research, the concept of *narrative complexity* is used to examine the cognitive development of children (Greenhalgh & Strong 2001; Newman & McGregor 2006; Scott & Windsor 2000). Children use narratives to relate events, establish and maintain friendships, and express their thoughts and feelings (McCabe & Bliss 2003). Narratives are defined here as stories about real or imagined events that are constructed by weaving together sentences about situational contexts, characters, actions, motivations, emotions, and outcomes (Gillam & Pearson 2004). Evaluation of children's narratives includes both the overall story grammar, such as characters, setting or events, and detailed aspects, such as pronoun usage or cohesive ties (Petersen et al. 2008). Other measures include the overall story length and artful elaboration (Ukrainetz 2006).

Computer-assisted story analysis for literature has typically occurred at the word level of granularity, suitable for studies of authorial style based on patterns of word use (Burrows, 2004). However, many interesting questions in human sciences lie on a much higher level. Literary scholars explore aspects such as the communal harmony and discord in Russian novels (Lieber 2011), characteristics of fictional portrayals of physicists (Dotson 2009), personality traits of characters in Victorian novels (Johnson 2011), or differences in social interactions in urban and rural English novels (Eagleton 2005). A high-level representation of a story in terms of characters and their interactions is therefore desirable to support such analyses.

2.2 Computational analysis of narratives

Early experiments in representing a story automatically focused on characterizing the plot as a sequence of events. Halpin and Moore (2006) designed an automated system for evaluating a student's ability to rewrite a fictional story. They extracted a chronologically ordered sequence of events in a predicate-argument structure, representing the entities and their actions to predict the plot quality using a supervised machine learning system. They attained up to 56% accuracy in predicting the grades given by the teacher. Chambers and Jurafsky (2009) proposed unsupervised narrative schemas by performing an induction of situation-specific semantic roles and linked them to event chains. Exploiting these schemas, McIntyre and Lapata (2010) created a story generation system. Since it focuses on events, however, it cannot enforce a global notion of how the characters relate to one another, therefore the focus of novelistic plot structure to the level of individual events has been criticized (Elsner 2012).

Recent NLP experiments begin to prioritize the entity-centric models, as proposed by Lehnert (1981). He suggests focusing on the *plot units* as a knowledge structure for representing narrative stories and generating summaries. Plot units are fundamentally different from the story representations that preced them as they focus on the affect states of characters and the tensions between them as the driving force behind interesting and cohesive stories. This theory is followed e.g. by Goyal et al. (2010), who automatically identify the affect states of characters in short fables to represent the story. However, the scheme is very fine-grained and not suitable for larger texts such as novels. Elson et al. (2010) focus on the frequency of dialogue interactions between characters to automatically extract social networks from British 19th-century novels. Kazantseva (2011) suggests an aspect-based summarization model for short fictional stories, focusing on finding the typical attributes of the main characters without revealing the plot of the story. Elsner (2012) uses a set of 19th-century romance novels to identify relationships between the characters. He extracts frequencies of characters in different chapters, and the emotional language with which that character is associated in that chapter, and measures the strength of the relationships between character pairs based on their co-occurrence in a paragraph. He then compares the novels and shows similarities in terms of character emotions and relations. Chambers (2013) improves his previous induction of narrative schemas by learning entity-centric rules (e.g., a victim is likely to be a person). Bamman et al. (2014) and Smith et al. (2013) present latent variable models for unsupervised learning of latent character types in movie plot summaries and in English novels. Iyyer et al (2016) present an unsupervised neural network model for tracking dynamic relationships between fictional characters using latent vectorial features (embeddings) to represent the semantic concepts.

2.3 Related work in story visualization

One of the earliest approaches to applying information visualization techniques to literature based on the contents of documents was developed by Rohrer et al. (1998). Based on a set of principal components extracted from term frequency vectors, a density field is generated and transferred to a three-dimensional visual display using blobs that have directions and are computed from the density field information. The Compus system developed by Fekete and Dufournaud (2000) visualizes lexical and syntactic information from literature, in their case French letters from the 16th century. Tailored to the comparative analysis of different documents, Monroy et al. (2002) propose an information visualization system called ItLv (Interactive Timeline Viewer). Using this tool, the authors demonstrate the visual comparative analysis of books, which ranges from overview like displays down to the page level, thus providing different levels of detail and interactive drill-down capabilities. With the increasing number of input documents to visualize, overview tasks are becoming more important, which is also reflected in corresponding visualization techniques. DeCamp et al. (2005) visualize large document collections using an iconic display built out of the conceptual contents per document. It possible to quickly gain insights into similarities or dissimilarities of the visualized document corpus, as the authors demonstrate for a collection of patents. The work of Chen (2006) also visualizes large collections of documents, but concentrates on revealing patterns as well as connections in the data. Based on node-links diagrams, the author demonstrates the identification and visualization of co-citation networks in scientific literature. While the application is primarily motivated by scientific literature analysis, many of their concepts, in particular, the visualization techniques are also applicable to novels.

Going back to the highest level of detail of literature visualization, the actual text, Weber (2006) introduces a color scheme to visualize text documents that is generated from part of speech information of words. The authors show that the resulting visual display can be used to identify and distinguish different genres of text, as well as insights into the syntactic structure of the documents. Akaishi et al. (2007) propose visualization techniques for the display of narrative structures of a document. They concentrate on the visualization of terms and their relationships, which is demonstrated by the authors to be useful by displaying the structure of the analyzed document in terms of the contained topics.

Keim and Oelke (2007) proposed visualizing a variety of different text features that contain the syntax characteristics, surface properties, and vocabulary metrics. The resulting visualization reveals differences clear enough to characterize and identify authors, and at the same time allows insights into regularities and irregularities of the analyzed book.

A visualization system that reveals common patterns in the analyzed text documents was developed by Don (2007). It allows the exploration of frequent words and n-grams and integrates them in several linked visualizations, which are able to guide users to interesting parts of the explored documents.

Van Ham et al. (2009) developed Phrase Net, a technique to generate overview like visualizations from unstructured text documents. Based on node-link diagrams, it is tailored to relationships, which can be retrieved on the syntactic or lexical level of the input text. The output is suitable for comparing different aspects of the analyzed texts, as well as to give an overview of the contained relationships, for example between characters. The visualization of characters and their relationships is also part of the work by Regan and Becker (2009). Besides that, they also provide insights into the terms connected to characters in order to describe their personality. Noteworthy are also the insights into the design process that produces the amount of text that is included in the visual displays.

Besides the detection of emotions in text, Mohammad (2011) proposes different visualization techniques in order to visualize emotions based on a timeline, as well as using a word cloud to produce words for different emotion categories. The author also proposes techniques for visualizing associated entities with words expressing emotion.

To gain insights into the differences of text documents, Jankowska et al. (2012) use common n-gram classifiers to build up visual signatures. The visualization is compact and therefore suitable for comparing a number of different documents, or parts of documents, by plotting the signatures next to each other, revealing differences in the usage of n-grams over different documents. Continuing with the idea of providing compact signatures or fingerprints, Oelke et al. (2013) demonstrate that matrices of fingerprints can be used to compare character occurrences and co-occurrences, which is suitable for identifying networks of characters as well as their changes over the analyzed text documents.

Weiler et al. (2015) propose visualizations to track different aspects of text data streams, which can also be applied to documents. They identify three properties to track the evolution of topics in documents, being importance, emotion, and context. The authors combine these metrics in a visualization that emulatesa morphing shape over time, effectively communicating topics and their sequential changes.

Besides these feature based visualization techniques, a number of related works are trying to re-create hand-drawn story lines by means of computational methods. Work by Tanahashi et al. (2012, 2015) and Liu et al. (2013) present a set of techniques that provide layout algorithms in order to create a line-based visualization of story progression with respect to characters, events, and locations.

The different entities are also interconnected, which makes important interactions clear in the resulting visualizations.

3. Complexity of language used: analyzing stylistic and content features in book text segments

Each of the following subsections in this and the next two sections is structured in the following way: First, we describe previous research directly related to the specific problem. Next, we explain our methodology for deriving particular features from the story text, which can be helpful for obtaining insights into a story's properties. Finally, we discuss our reasons for implementing a particular visualization of the obtained features, and present cases which enableuser understanding of a story's complexity.

3.1 Expressing the reading ease of a text: readability measures, long and foreign words

Previous work

Traditional readability measures rely on two main features, beingword length and sentence length. They are computed by the average number of characters (or syllables) per word and the average number of words per sentence and are combined with manually determined weights resulting in a grade level as output. The most well known methods of this type are the Flesch–Kincaid Grade Level (Flesch, 1977) formula, which uses the average number of words per sentence and the average number of syllables per word to predict the grade level, the Automatic Readability Index (Smith and Senter, 1967), and the Coleman–Liau Index (Coleman and Liau, 1975). However, they have also been subject to criticism as they only capture surface characteristics of the text and can be misleading (DuBay 2004).

More recently, supervised learning algorithms have been used to automatically combine several text properties extracted from training data and used to associate them with the corresponding readability class. Feng et al. (2010) show that the density of entities (nouns and proper nouns) introduced in a text corresponds to a higher working memory burden for the reader, thus contributing to higher readability level. Pitler and Nenkova (2008) explore discourse level features from the Penn Discourse Treebank (Prasad et al. 2008) and report on their usefulness in predicting text readability.

Our features

Our framework computes the readability measures (Flesch–Kincaid Grade Level, Automatic Readability Index, Coleman–Liau Index) for each paragraph, as well as the ratios of each part of speech type and a proportion of named entities and foreign words in the text, using the OpenNLP (Morton et al. 2005) Tagger and Name Finder.

Visualization

The input data for the readability visualization is computed based on the paragraphs of the analyzed books, since the readability metrics mostly refer to a number of sentences or words, instead of single sentences. For each of the paragraphs, the information from the corresponding chapter is given.

The visualization of readability metrics for a given book is designed with the following goals in mind:

1. Provide an overview of the changes in readability metrics corresponding to the chapters or paragraphs of a book in a compact way
2. Keep the structure of the book visible, so that interesting values can be correlated to the corresponding unit of text.
3. Clearly indicate areas where the analyzed text is easy or hard to read

Goal one results in two requirements for the general construction of the visualization. First, paragraphs, as well as chapters, have to be indicated visually so that they can be easily seen. Second, the visual design has to be as compact as possible to enable effective communication of the readability metrics, while at the same time providing an overview of as much text (and data) as possible. Visually, these requirements have been met by a matrix-like visual design. Each paragraph is represented by a cell. A chapter, which is composed of a number of paragraphs, is represented by a group of cells. Compact representation also imposes requirements on the data preprocessing, which is done before the visualization is created. If all paragraphs are visualized with a cell, the resulting width and height of the created graphical depiction would be too large to satisfy the compactness constraint. As a consequence, the data is aggregated before it is visualized. The aggregation is based on a fixed window size, of which the arithmetic mean of the contained readability metrics is computed. In this way, outliers in both hard and easy to read directions should stay visible, but at the same time the overall data will reflect the properties of the aggregated paragraphs. In

addition, the windowing follows the logical structure of the text, which is given by chapters. If a window contains a chapter, the window size is reduced so that only paragraphs from the same chapter are aggregated, and finally, the aggregation window is enlarged again and moved to the boundary of the next chapter.

Goal two has in one main requirement, namely that the paragraphs and chapters can be easily perceived as such. Based upon the previous requirements, the structure needs to be resembled by positioning and aligningthe cells to represent a paragraph. One condition of representing the structure is determined by the fact that the cells that are referring to consecutive paragraphs should be placed next to each other. This condition is fulfilled by aligning the paragraphs on a common baseline in the order in which they appear in the single chapters. A visual overflow per row, which could happen when the width of cells exceeds the width of the visualization space, is solved by introducing a line break so that multiple rows can represent the same chapter. Having made sure that the paragraph alignment resembles the structure of the book, the final step in visually representing the structure of the book chapter is to visually indicate the affinity of paragraphs to the corresponding chapter. This is done by introducing a margin between the rows that refer to different chapters, and which is large enough to be perceived easily as the border between chapters. The result is a layout where rows of cells indicate paragraphs, line breaks are used to mitigate the overflow of rows being too wide for the visualization space, and rows referring to different chapters are separated clearly by a wide margin between them.

The requirement resulting from the last desired property, the visual indication of the readability metric, refers directly to the representation of cells. The displayed property is presented per paragraph and is represented by cells, the most prominent visual property of these being their area and color, which are used for the indication of the readability metric. Cells of paragraphs that are, according to the computed readability metric, easy to read, are filled with a light reddish color. In contrast, cells with low readability are filled with a darker red color. Readability values in between are mapped to a number of bins, each represented by a color interpolated between light red and the dark red tone. The result is a color map that assigns readability values to a color starting at light red (easy to read) todark red (hard to read), as well as the colors in between.

In the example given in Figure 2, Flesh Reading Ease is computed from the book *Harry Potter and the Sorcerer's Stone* by J. K. Rowling and visualized with the described technique. The overall impression is quite mixed and shows, except for some single cells (fifth and eleventh row), a mixed picture of the readability score.

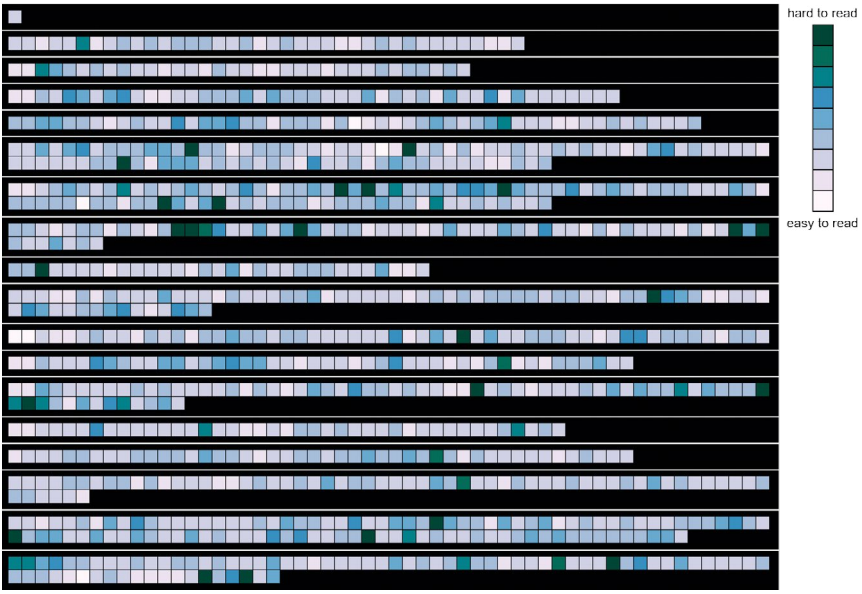

Figure 2: Flesch Reading Ease score per paragraph of *Harry Potter and the Sorcerer's Stone*.

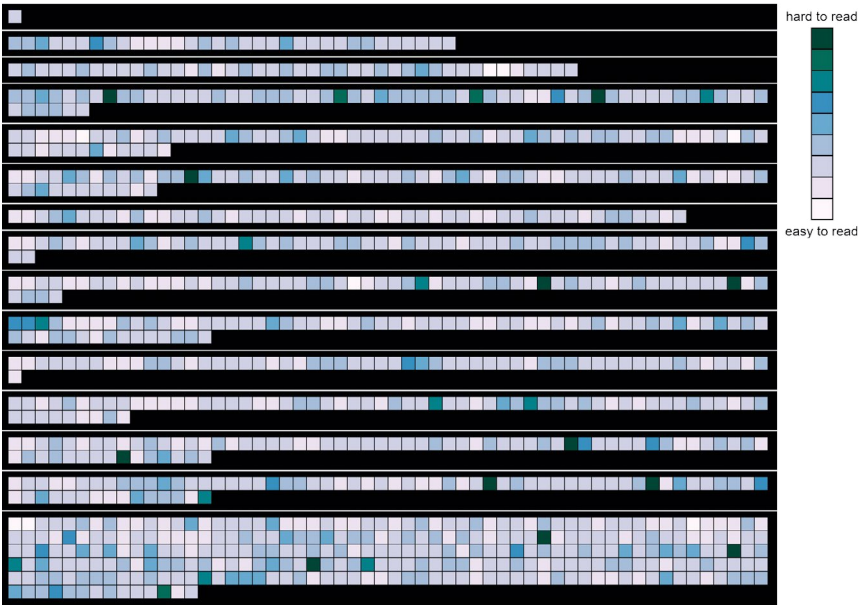

Figure 3: Flesch Reading Ease score per paragraph of *Harry Potter and the Chamber of Secrets.*

Comparing the first two volumes of the Harry Potter series with respect to Flesh Reading Ease, reveals that they are quite similar (Figure 2 and Figure 3). In both books there are very short passages that are hard to read, while the overall impression remains mixed and they are finally judged clearly as easy to read.

3.2 Analyzing the style and content of text

Previous work

Analysis of the style of books has been explored in the areas of authorship attribution (Zhao 2007), author profiling (Rangel et al. 2013) and computational stylometry (Daelemans 2013), mainly with the aim of differentiating between authors and author groups. Ashok et al. (2013) additionally show that stylistic features can predict the book success with high accuracy (84%). They find that less successful books rely on verbs that are explicitly descriptive of actions and emotions (e.g., "wanted", "took", "promised", "cried", "cheered", etc.), while more successful books favor verbs that describe thought-processing (e.g., "recognized", "remembered"). Additionally, less successful books rely more on topical words that could be almost cliche, e.g., "love", typical locations, and involve more extreme (e.g., "breathless") and negative words (e.g., "risk"). They also report that the prepositions, nouns, pronouns, determiners and adjectives are predictive of highly successful books whereas less successful books are characterized by the higher percentage of verbs, adverbs, and foreign words.

Our features

We measure numerous aspects that have proven useful in previous work that captures the style of an author. The following features are based on the OpenNLP part-of-speech tagger (Morton et al. 2005), using the maximum entropy model to annotate tokens with the Penn Treebank POS tagset: frequency of adjectives and adverbs in their comparative and superlative form; the frequency of personal and possessive pronouns, and frequency of exclamation and question marks. Additionally, we measure the contextuality score, which is considered an approximation of how formal or casual a given text is (Heylighen 2002), with the knowledge that some parts of speech types contribute to a more casual style (such as pronouns or adjectives) while others occur more often in a more formal text (such as nouns and determiners). The contextuality score reaches values between 0 and 100 and is calculated as follows:

$$\text{contextuality} = (\text{nouns} + \text{adjectives} + \text{prepositions} + \text{determiners} - \text{pronouns} - \text{verbs} - \text{adverbs} - \text{interjections} + 100)/2$$

where each part of speech type is expressed as its relative frequency compared to all words.

Additional insights into the overall characteristics of a given text can be obtained by exploring the topics that occur in each chapter. Sociolinguists commonly use the Linguistic Inquiry and Word Count (LIWC) lexicons for this purpose (Pennebaker 2003), which we also employ here. LIWC is unique in the sense that it provides not only a set of basic topical categories (such as family, money, friends, work) but also expressions of cognitive processes (insight, tentativeness, uncertainty...) or inner drives (achievement, inclusion ...). There are 69 categories in total. In addition, we use topical word lists from www.enchantedlearning.com, which enrich our set with additional categories such as school, computers, cars, politics or swear words.

Word lists such as those mentioned above are often criticized for being based only on the written form of the occurring expression, without taking into account additional information about its eventual polysemy or morphological variations. Therefore, we attempt to obtain more precise information about the categories of individual words using WordNet (Miller, 1996) semantic categories, sometimes also called lexicographer files or supersenses (Ciaramita and Altun 2006).

Wordnet supersenses are assigned to verbs and nouns on a WordNet synset level, i.e., taking into account the distinction between different senses of the same word. There are 26 categories for nouns, such as animal, person, artifact or process, and 15 categories for verbs, such as communication, motion, cognition or emotion. We retrieve the supersense for each verb or noun in the text by using its lemma and part of speech tag and mapping it to its most frequent WordNet sense.

The visualization of different stylistic features, which can be related, such as the LIWC dictionary words or WordNet senses, is mainly driven by the need to gain an impression abouth whether they occur, and if so, in what relation they stand with each other. To communicate the actual values of features, it must be possible to follow the values over the progression of a book, which is the first property that a visualization needs to have for this kind of data and requirements. Second, to be able to judge the domain of a set of features, as well as a region of concrete values, a comparison must be possible.

A bar chart is capable of adequately fulfilling the requirements resulting from the first property. The different bars are make clear that the displayed data is not coming from a continuously occurring feature, which is being measured at discrete stages. Additionally, the area of the bar, which can be filled with a color,

Table 2: Overview of the lexicons used in our experiments

Lexicon name	Reference	No. of words	No. of categories	Example of categories and content
Linguistic Inquiry and Word Count	Pennebaker 2003	10,555	64	Feeling: hard, press, warm Certainty: Fact, confidence, always
NRC Emotion	Mohammad 2011	8,265	8	Surprise: cheer, inspired, unexpected Joy: amuse, elegant, happily
NRC Sentiment	Mohammad et al. 2013	5,636	2	Positive: mighty, prestige, unconstraint
Hu & Liu Sentiment	Hu and Liu 2004	6,789	2	Negative: annoy, mistaken, worse
In-house emotion list	Wanner et al. 2011	416	16	Anxiety: cautious, fearful, nervous
In-house topic list	Enchanted learning.com	4,735	24	Politics: choice, quorum, voter School: math
WordNet lexicographer files – noun supersenses	Miller 1996	117,798	26	Animal: fish, cat Body: hand, leg Person: teacher
WordNet lexicographer files – verb supersenses	Miller 1996	11,529	15	Motion: fly, walk, swim Communication: talk, scream

can assist in the perception of the feature value. With each bar we represent a number of sentences, which are aggregated with respect to the logical borders of books, namely the chapters. The same reasoning as before also holds here, meaning that the arithmetic mean is the choice of aggregation method for the numeric feature values (since it is sensitive to outliers), which increases the possibility that the aggregate will have a value near to the outliers, as well as providing an effective way of preserving the feature values of the non-outliers. The order of the bars preserves the sequence of the represented aggregates, which allows conclusions based on the position of a bar with respect to the book, e.g. in the beginning, in the first half, or near the end. Also, the distance from neighboring bars is easy to perceive, because this can be done by comparing the different heights, and allows tracking of feature values, trend spotting, as well as tracking outliers during the progression of the book in question.

The second property effectively opens up the design in such a way that it is possible to perform the aforementioned analysis tasks for a set of features. Having the initial design based on a bar chart, a stacked bar chart adds comparison capabilities. However, the direct comparison between two or more features is negatively affected by a classical stacked bar chart, where the single bars are placed on top of each other in a single instance, leaving only one baseline in the chart, which grounds the perception of the lowest part of the bar in the chart. Because of this perception issue, each feature is still represented by a single bar chart instance. To make the charts comparable, they are placed on top of each other, and their scales are normalized accordingly. Their start and end, as well as the window size (resulting in the number of bars), are aligned. The result preserves a baseline for each represented feature, as well as allowing quick, but not exact comparisons of the feature values. An exact comparison is not considered a firm requirement, because of the different origins of the feature set as well as the language use, which may be fundamentally different for the measured properties of the text, already makes exact comparisons hard to interpret. To support navigation in the bar chart, a highlighter covering all charts follows the mouse.

Similar to the previous visualization, the data is aggregated in a window fashion which respects the logical borders of a book, which are determined by chapters. The window can be adjusted in order to give more detail, or aggregate to a high level, so that for very large window sizes the aggregated data corresponds to a whole chapter, while it is still possible to transfer to a high level of detail.

In Figure 4, the LIWC common verb classes “Future Tense”, “Past Tense”, and “Present Tense” are shown. The bar heights are globally normalized, which means they can be compared among themselves. From this viewpoint, it becomes clear that the first Harry Potter novel is written in the past tense, and there are only rare references to either the future or the present tense. Having a story line

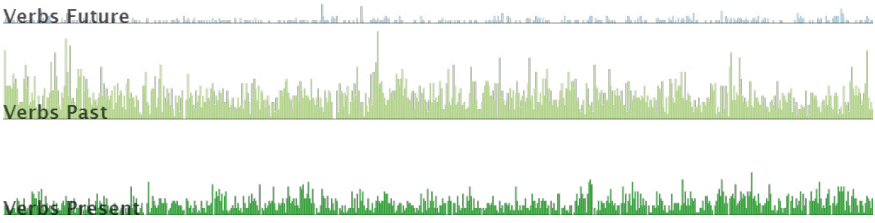

Figure 4: Visualization of the frequency of future, past and present tense in *Harry Potter and the Sorcerer’s Stone.*

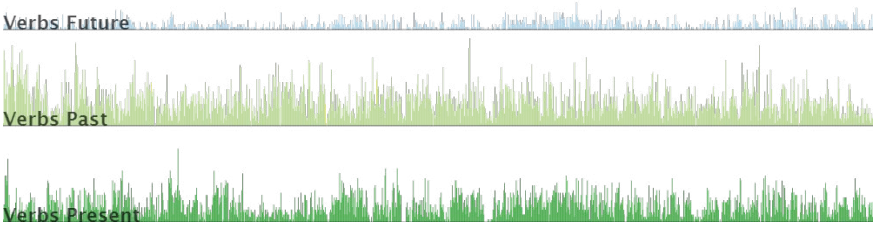

Figure 5: Visualization of the frequency of future, past and present tense in *The Lord of the Rings – The Fellowship of the Ring*.

jumping frequently between tenses can be a sign of a quite complex story with forward and backward references, but this is not the case here. For all other Harry Potter novels, these charts look similar. Comparing them with fictional literature from other authors, for example with the first volume of "Lord of the Rings" by J. R. R. Tolkien (Figure 5), it becomes apparent that a similar picture can be expected for other novels written in the past tense.

The same visualization technique can be used to get an overview of the type of actions in a book. Appropriate for this are the extracted WordNet supersenses of verbs. In Figure 6, a subset of these features and their occurrences in *Harry Potter and the Sorcerer's Stone* is displayed. The feature values are globally normalized, which means that the height of the bars can be compared with each other. While analyzing this visualization, it becomes clear that two of the selected categories are dominating the verbs used, which are "telling, asking, ordering, singing", as well as "walking, flying, swimming". In contrast, verbs from other categories such as "fighting" or "eating and drinking" are only rarely used in the novel. When comparing this with the same data of the last book of the *Lord of the Rings* series (see Figure 7), three observations can be made. At first, there seems to be a much smaller focus on actions in the context of "telling, asking, ordering, singing" in *Lord of the Rings*. The same is true for verbs from the category "touching, hitting, tying, digging", but to a lesser extent (Figure 6).

The third observation seems toprovide the biggest difference between the two books, being is the increased occurence of words from the "eating and drinking" category at the end of the *Lord of the Rings*. This is due to the coronation of the character Aragorn, where the festivities are described. Similar passages are missing from the *Harry Potter* novel, which can be clearly seen when comparing Figure 6 and Figure 7.

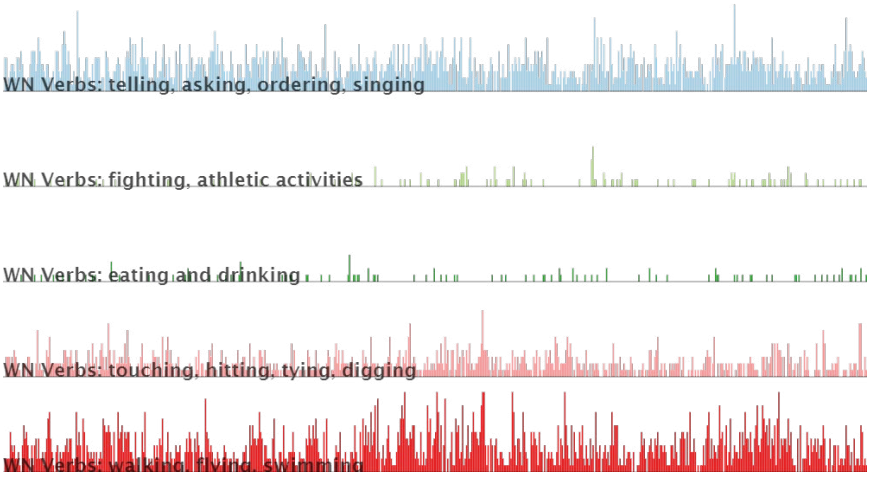

Figure 6: Visualization of a selection of Wordnet verb supersenses in *Harry Potter and the Sorcerer's Stone*.

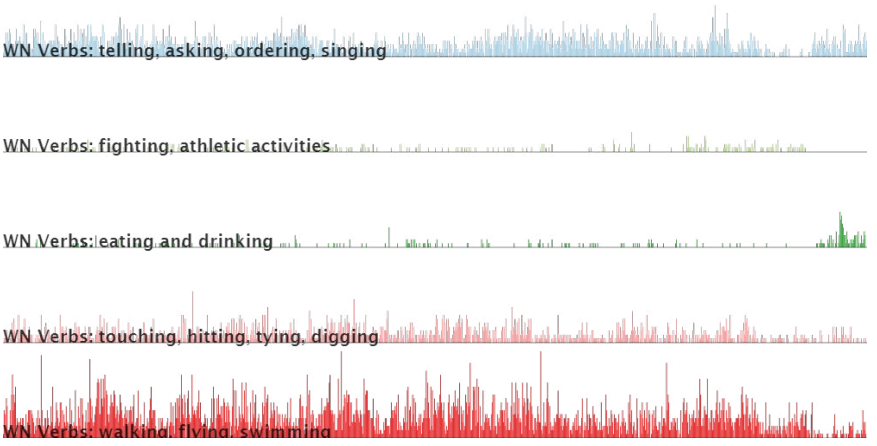

Figure 7: Visualization of a selection of Wordnet verb supersenses in *The Lord of the Rings – The Return of the King.*

3.3 Emotions and sentiment

Previous work

Ovesdotter Alm et al. (2005) set up a system to automatically predict the six basic emotions (anger, disgust, fear, happiness, sadness, surprise) in 22 children's fairy tales on a sentence level. They achieve an accuracy of 63%, and point out that simple bag-of-words models are prone to errors in texts enriched with frequent figurative expressions. Volkova et al. (2010) experiment with the German texts of Brothers Grimm fairy tales. They investigate how emotions are expressed in these stories and how people associate emotions with certain text fragments of these fairy tales. The authors define several positive and negative emotional categories, and then several annotators manually annotate the text passages that convey these emotions. They find that in most texts, positive emotions are expressed more frequently than negative ones. The authors observe a reasonably high inter-annotator agreement for emotions in thetext. Mohammad (2011) uses his *NRC Emotion Lexicon* to explore emotions displayed in fairy tales and novels. He explores how the frequency of words associated with certain emotions differs for different types of literary text and how they change through the course of a narrative. Moreover, he compares distributions of emotional words in novels and fairy tales, finding that fairy tales tend to have higher emotional density.

Our features

First we measured the positive and negative sentiment and the six basic emotions (happiness, sadness, fear, anger, surprise and disgust) using in-house word lexicons inspired by www.psychpage.com. Additional word lists based on the same website measure more fine-grained emotional states such, as anxiety, confusion, helplessness or love.

Sentiment lexicons, while widely used, have been a subject of criticism for capturing only very explicit expressions and, more importantly, out of their syntactic and semantic context. Therefore sentences such as "This movie was actually neither that funny nor super witty" would be incorrectly classified as positive based on the sum of its positive and negative expressions (2 + 0). This problem can be overcome by studying the compositional grammatical structures of the sentences. This has been done in the Stanford Sentiment Analyzer (Socher et al. 2013), using recursive neural tensor networks. In their system, the above-mentioned example is classified correctly as negative. We employ their trained model in our system as well, to predict sentiment score on a 5-point scale on sentence level.

Visualization

The visualization of emotions and sentiment is based on similar reasoning as the visual display of stylistic features. The corresponding visualization should effectively communicate the value of the corresponding feature, as well as its development, and allow comparative findings and insights. Keeping the same visual design as for stylistic features is motivated by the observation that the variety of emotion- and sentiment-related contexts contributes to the perceived degree of story complexity.

For the first property, a bar chart was used to effectively communicate feature values and their development. The feature values are double encoded in the bars by using their height as an indicator of the represented numeric value, as well as the area that is colored uniquely per emotion feature. It is possible to follow the different values, as well as to perceive changes over the progression of a book.

The comparison of different emotion features is enabled by stacking the bars of each feature in a single plot. This different approach is chosen, because in contrast to the stylistic features, the emotional context of a text passage represented by a single bar can be seen as a limited space, whereas a single emotion can dominate the perception of a chapter, for example, if words from a negative emotion context occurr more frequently than any other emotions. This is taken into account by effectively limiting the visual space to the height of one bar chart, and the different emotions, which are about to be compared, are shown as stacked bars in that limited height. Together with the colored area, representing the feature value, this technique ensures that any dominating emotion, and its assigned color, also dominates the perception of the limited area per-visualized text passage. This desirable property sacrifices the exact perception of the feature values and their comparison, but at the same time allows the emotional context and any dominating emotion to be followed effectively.

The visualized data is computed by a window over the sentences that reflects chapter borders and uses the arithmetic mean for feature value aggregation.

Inspecting Figure 8, which depicts the word counts of six different emotion word dictionaries (happiness, sadness, fear, anger, surprise, disgust), reveals that for the first Harry Potter novel, the number of sadness and anger words dominate the emotion categories as computed with the available in-house sentiment word dictionaries. There are outliers of the fear emotion in the beginning and the last third of the book, which are locally quite restricted and therefore describe a drastic, but limited change of the emotional tone of the book. Similar to the stylistic features, this general impression does not change much for further books in the Harry Potter series, but the number of outliers of a specific emotion increases (compare Figure 8 with Figure 9).

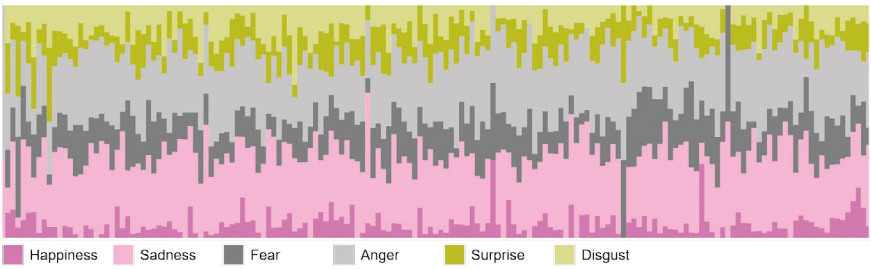

Figure 8: Emotion dictionary word counts from *Harry Potter and the Sorcerer's Stone.*

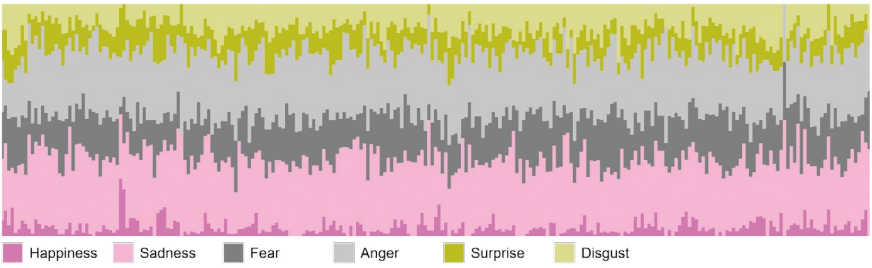

Figure 9: Emotion dictionary word counts of the last Harry Potter novel, *Harry Potter and the Deathly Hallows.*

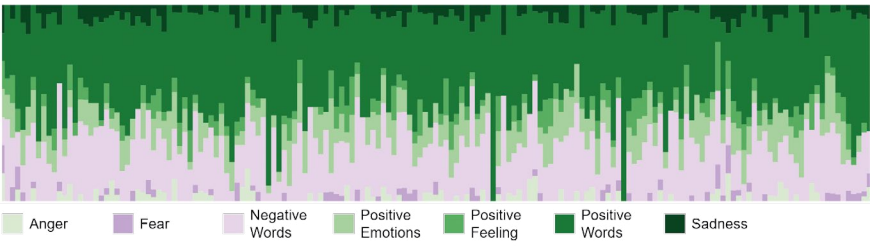

Figure 10: Emotion words counts from enchantedlearning.com for the first Harry Potter novel.

Comparing this with a higher-level abstraction of emotion words, as can be seen in Figure 10, it becomes clear that the aforementioned observation is only true in the analyzed context.

Here, it can be seen that the number of words representing positive emotions, positive feelings, and positive words dominate the chosen emotional context, which is comprised of anger, fear, negative words, positive emotions, positive feelings, positive words, and sadness.

Depending on the chosen words, the visualization will differ and present a different picture of the emotional context.

4. Complexity of a plot: identifying characters in a literary text and relations between them

Beyond linguistic complexity, an important aspect influencing the ability to understand a story is the number of characters appearing in it and the complexity of their interactions. How many literary characters appear in a novel? Despite the seeming simplicity of the question, precisely identifying which characters appear in a story remains a difficult problem in literary and narrative analysis.

Previous work

Characters form the core of many computational analyses, from inferring prototypical character types (Bamman et al. 2014) to identifying the structure of social networks in literature (Elson et al. 2010; Lee and Yeung 2012; Agarwal et al. 2013; Ardanuy and Sporleder 2014; Jayannavar et al. 2015).

Our methodology

In order to identify individual characters in fictional literature, we have developed two character identification methods. A fast one to identify as many mentions of a character in the text as possible, and a precise one to assign direct speech to a particular speaker in the book. The second method is described in section 5. The first method, used here, is a two-phase process, where first a set of candidates is generated using several predefined rules, while in the second step the whole document to be analyzed is scanned for occurrences of candidates, in order to ensure no occurrences are overlooked. The rules contain heuristics based on common salutations (extracted from the English Wikipedia using DBpedia queries for Women's and Men's social titles, Military ranks, Academic

ranks, and Political titles). In addition, grammar based rules are in place that hint at a possible character based on specific parts of speech combinations, such as the identification of possessive constructions indicating a character. These include detection of possessive pronoun constructions (Figure 11, lines six and eight), as well as verbs in the 3rd person singular (Figure n, line ten) usually having a character in context.

Finally, the character candidate tokens are followed to capture full names and titles corresponding to characters (lines 13 to 18, Figure 11). This permits detection of characters such as "Lord Voldemort" or "Professor Dumbledore", while state of the art named entity detection produces results with the titles and salutations typically missing, e.g. with the Stanford Named Entity Recognizer (Finkel et al. 2005). The full attribution of characters, including any titles or salutations, properly reflects the books contents and allows different kinds of references to the same characters to be captured. We performed several experiments with the final set of heuristics and a number of different books written by different authors, to clarify if postprocessing to resolve coreferences is required. Based on the exemplary results and the fact that we kept the set of heuristics small, special treatment of coreferences has been omitted, as we found only very few, if any (below ten) false positives. Since we also identify other types of names, besides animated named entitiessuch as locations, the same visualization can be used

```
   Input  : A sequence of tokens T, a list of salutations and titles S
   Output: sequences T' ∈ T that are likely to denote a character
 1 T' ← {}
 2 for i ← 0 to |T| do
 3 |   t ← T[i], t' ← T[i − 1], t'' ← T[i + 1]
 4 |   if t ∈ S then
 5 |   |   push(T', T[i, follow_character(T, i)])
 6 |   else if POS(t) = NNP and POS(t'') = IN and ends(t, 's) then
 7 |   |   push(T', T[i, follow_character(T, i)])
 8 |   else if is_noun(t) and is_noun(t'') and ends(t, s') then
 9 |   |   push(T', T[i, follow_character(T, i)])
10 |   else if is_uppercase(t'[0]) and POS(t) = VBZ then
11 |   |   push(T', T[i, follow_character(T, i)])
12 return T'

13 define follow_character(T, o)
14 |   for i ← o to |T| do
15 |   |   t ← T[i]
16 |   |   if !is_uppercase(T[i][0]) and !is_hyphen(T[i][0]) then
17 |   |   |   break
18 |   return o
```

Figure 11: Heuristics to detect character names.

to explore the relation between characters and locations in the book. When the secondary characters mostly stay in one location, the book can be considered less complex.

Visualization

Thinking of co-occurrences as a node-link structure is obvious. Nodes represent the characters, and for each co-occurrence a link can be added to the graph, or the weight of an existing link between the two co-occurring character nodes can be increased. Visualizing character co-occurrences directly by means of a graph imposes huge perceptional challenges to the reader, because in fictional literature it can be expected that characters frequently co-occurr with each other, for example, because of interactions. A graph constructed as mentioned before can be visualized by utilizing a number of different graph layout techniques, which all optimize certain criteria, such as keeping the number of edge crossings low, imposing a high degree of visual symmetry, or keeping the average edge length below a certain threshold. For fictional literature it may be expected that the visualization of a graph, based on a suitable graph layout technique, could suffer from an overplotting of the edges or the nodes, making it difficult to perceive frequent co-occurrences or patterns in the co-occurrences.

Using adjaceny matrices is a technique that utilizes the same kind of data, i.e. node-link structures, where the nodes represent characters and links are used to indicate co-occurrences, and where the visualization is well known to scale for large numbers of rows and columns (representing characters), while also supporting the perception of visual patterns. In these matrices, the rows and columns represent nodes from the graphand the cells are used to map the connections of nodes in the graph. There is intentionally no overlap, which preserves any visual patterns and at the same time allows networks in the data to be identified.

To depict character co-occurrences with adjacency matrices (see Figure 12), an undirected, weighted graph is created from the co-occurrence data. Each node represents a character, for each co-occurrence of the character an edge is added to the data structure with the weight of one. In case an edge between two characters already exists, its weight is increased by one. Having constructed the graph, a subgraph representing the co-occurrences between the top n most occurring characters is extracted. The visualization of the adjacency matrix from this subgraph represents each character by a single row and a single column, which makes the result symmetric. For each character c, the co-occurrences with the other characters d are examined, and the corresponding cell in the row belonging to c and columns of characters from d are assigned a color on a color map ranging from light blue (few co-occurrences) to a dark blue (most co-occurrences),

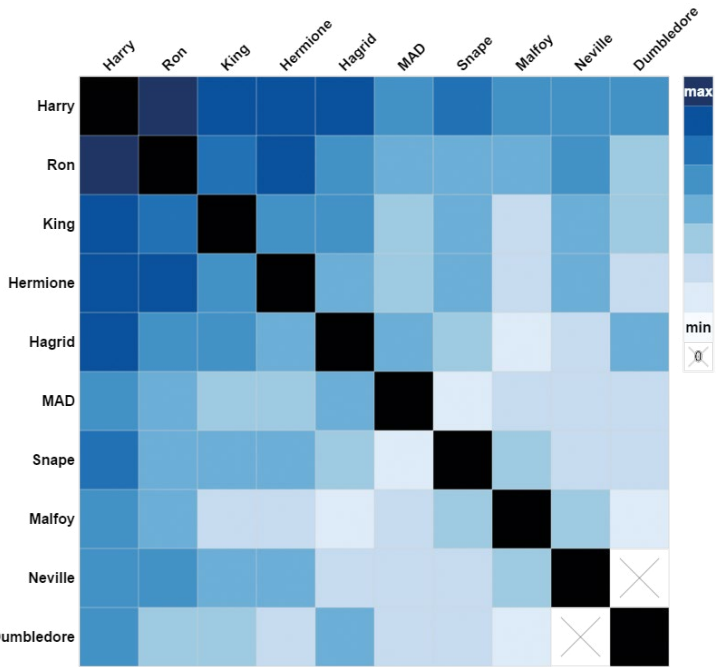

Figure 12: Co-occurrence of the ten most frequently occurring characters in Harry Potter and the Sorcerer's Stone.

that is used to fill the cell area. If a character does not interact with another one, which is part of the adjacency matrix, the visual indication is a cross on white ground. The rows and columns are ordered by the absolute occurrences of characters, ensuring that frequent characters and their co-occurrences are visible together starting at the top left of the matrix visualization. To indicate the symmetry of the matrix, the diagonal, which refers to co-occurrences of characters with themselves, the corresponding cells are marked in black.

Figure 13 shows the 15 characters that occur most often in the first Harry Potter novel. From top left to top right, or top left to bottom left, the characters are ordered descendingly according to their occurrences over the entire book. It can be seen that Harry (first row, first column), occurs together quite frequently with every single of the remaining top ten characters, as is indicated by the dark blue color of the cells. It can also be seen that Dumbledore does not occur at all together with Neville, as is indicated by the black x on white ground of the corresponding cell. Since the matrix is symmetric, the diagonal would indicate co-occurrences of the character with themselves, which is encoded with a cell marked in black. Figure 14 visualizes the same information of the first book of the Lord of the Rings series.

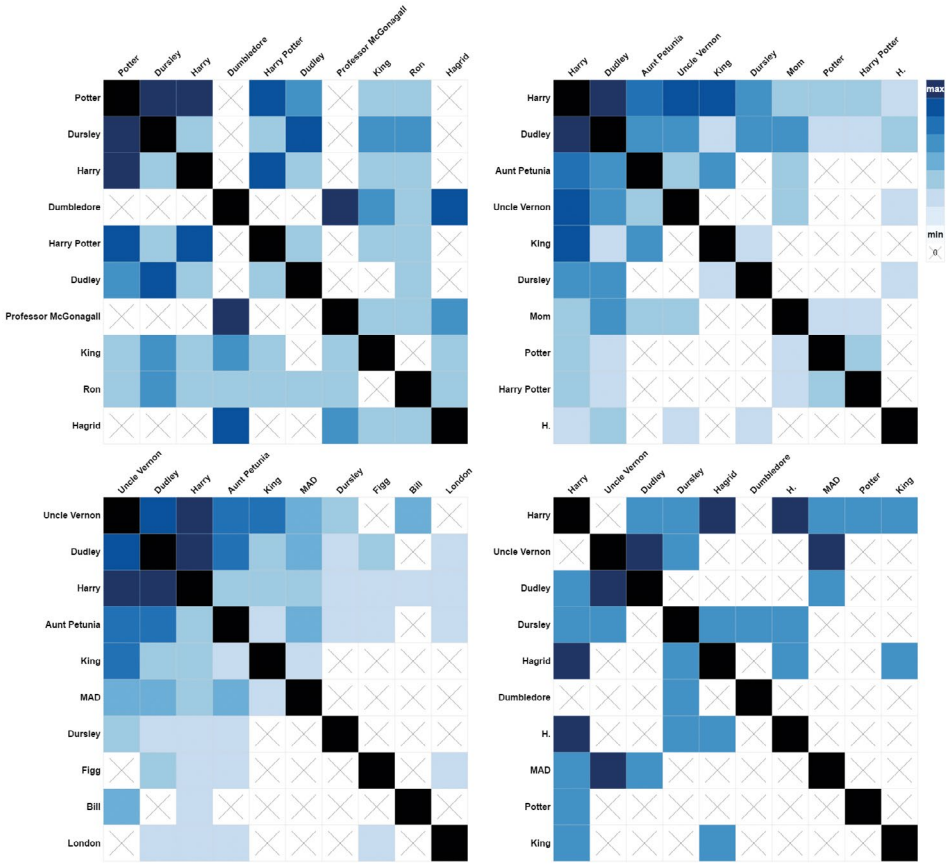

Figure 13: Visualization of the top ten character co-occurrences in the first four chapters of *Harry Potter and the Sorcerer's Stone*.

This matrix-based visualization concept can be applied to a much higher level of detail. In Figure 14, the first four chapters of the first volume from the Lord of the Rings series are shown, the top left matrix depicts co-occurrences of chapter one, top right chapter two, bottom left chapter three, and bottom right chapter four. Simply by examining the names of the top occurring characters per chapter, it can be seen that quite drastic changes in the involved characters also imply a change in the story line. It can also be observed that the number of co-occurring characters in Lord of the Rings is quite different to the Harry Potter novel visualized in Figure 13. In the latter almost every top occurring character has interactions with the others, while in the former this is not the case, as we can observe that certain characters, such as Gandalf, Gollum, or Took have only a limited number of co-occurring characters. For both Harry Potter

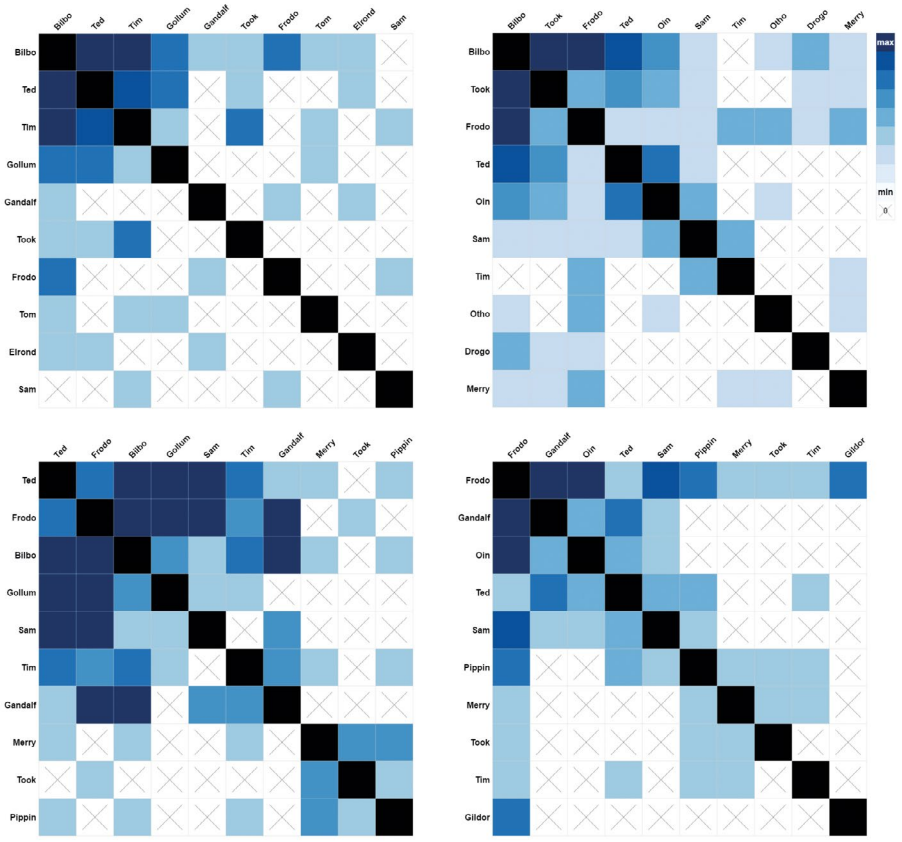

Figure 14: Visualization of the top ten character co-occurrences in the first four chapter of *Lord of the Rings – The Fellowship of the Ring*.

and Lord of the Rings, a group of characters is clearly in focus per chapter. On the top left there is usually a group of four of five characters, which co-occur quite frequently (dark blue) with themselves. In some cases, in particular the main character, they are part of the pattern and co-occur with almost every character of the top ten in the visualized chapter.

5. Complexity of the characters: analyzing the book from an individual character's perspective

5.1 Identifying direct speech and automatically assigning speakers

According to character-driven (as opposed to plot-driven) literature theory, the protagonists in a novel are the central aspect of the story. Characters possess multiple layers of personal traits which are exposed as the story develops. Our aim in this work is to gain an insight into each character's complexity by analyzing the concepts and style in their direct speech produced. In order to explore what the characters are discussing and in which manner, we first need to identify the direct speech segments in the text and assign them to the appropriate speaker. This is a difficult problem since direct speech utterances in modern literature are rarely in the traditional format, such as *"direct speech," said John.*

Related work

Numerous publications study how quotes can be attributed to a speaker, however, only a few of them deal with literary texts. Elson et al. (2010) propose a dialogue attribution method specifically designed for novels. They extract a feature vector for each pair, consisting of a candidate speaker and a quote. They use such information as the distance between the candidate and the quote, and the number of appearances of the candidate in the book. Several classifiers are then used to discriminate between vectors that belong to speakers of a given quote and other characters. Using a corpus of 19th and early 20th-century fiction by six authors, they achieve an overall accuracy rating of 83%. O'Keefe et al. (2012) conduct further experiments using the same corpora. In contrast to Elson et al., they perform the attribution without the use of annotated data. The best result of 53.3% is obtained by a simple rule-based method. He at al. (2013) present another supervised dialogue attribution method, using an unsupervised actor-topic model, which is used to predict likely speakers based on topic distribution of relevant text. Accuracy of between 80% and 86% is achieved.

Elson et al. (2010) further use the dialogue method to construct social networks for book characters. The network is represented as an undirected graph with nodes representing characters and weighted edges describing their relationships. The weight of an edge between a pair of character nodes is set according to the total word length of quotes spoken by one of the characters, in cases where there is a quote by the other character within 300 words. The networks thus extracted are used by the authors to compare the degree of connectedness and structure of the network for books with different settings, providing results that refute popular hypotheses from literary studies. Agarwal et al. (2012)

manually extracted a social network in Alice in Wonderland using social events. Two specific kinds of social events were used: *interactions*, in which both parties are aware of the event, and *observations*, in which only one party is aware of the event. Vala et al. (2015) propose a new character identification technique, bootstrapping characters from seeds of names found with the Stanford Named Entity Recognizer (Finkel et al. 2005) and Stanford coreference resolver (Recasens et al. 2013), or entities denoted as *animated* in WordNet. They achieve an F-score of up to 75%.

Our methodology

The most challenging task in building the direct speech data set is assigning direct speech utterances to the correct speaker. We benefit from the epub format of the e-books, which defines a paragraph structure in such a way that only the indirect speech chunk immediately surrounding the direct speech is considered:

> John turned to Harry. "Let's go," he said.

Given the large amount of text available in the books, we focus on precision rather than coverage and discard all utterances with no explicit speaker (i.e., 30-70% of the utterances, dependent on the book), as the performance of current systems on such utterance types is still fairly low (O'Keefe et al. 2012; He et al. 2013; Iosif and Mishra 2014). Conventional coreference resolution systems, which we tried, did not perform well on this type of data and were therefore not used in the final setup. We adapt the Stanford Named Entity Recognizer (Finkel et al. 2005) to consider titles (Mr., Mrs., Sir...) as a part of the name and to treat the first person "I" as a named entity. However, identifying only the named entity PERSON in this way is not sufficient. In our evaluation sample consisting of a Game of Thrones book "Pride and Prejudice" (the former annotated by us, the latter by He et al. (2013)), 20% of utterances with explicitly named speaker were not recognized. Of those correctly identified as a Person in the adjacent indirect speech, 17% were not the speakers. Therefore, we implemented a custom heuristics (illustrated in Figure 15), which additionally benefits from the WordNet semantic classes of verbs, enhancing speaker detection by recognizing the nouns. With this method, we retrieve 89% of known speakers, of which 92% are assigned correctly. Retrieved names are grouped based on string overlap (e.g. Ser Jaime and Jaime Lannister), excluding the match on the last name, and corrected for non-obvious groupings (such as Margaret and Peggy).

To quickly get an insight into the extracted direct speech of characters, word clouds can be used. They can enable a quick overview of the words used, while

Algorithm 1 Assign speaker

1: *nsubj* ← subjects in adjacent indirect speech
2: **if** *count*(*nsubj*(*i*) = PERSON) = 1 **then** *speaker* ← *nsubj*
3: **else if** *count*(*nsubj*(*i*) = PERSON) ≥ 1 **then** *speaker* ← the nearest one to *directSpeech*
4: **else if** *directSpeech* preceded by VERB.COMMUNICATION **then** *speaker* ← the preceding noun(s)
5: **else if** *directSpeech* followed by VERB.COMMUNICATION **then** *speaker* ← the following noun(s)
6: **else if** *directSpeech* followed by gap & VERB.COMMUNICATION **then** *speaker* ← the noun(s) in gap
7: **else if** *directSpeech* preceded by gap & VERB.COMMUNICATION **then** *speaker* ← the noun(s) in gap
return *speaker*

Figure 15: Our method for assigning a speaker to a direct speech utterance.

at the same time expressing the importance of the words, typically measured by their frequency. To do so, two visual variables are commonly used: the size of the words and their color.

Having a data set with direct speech from the assigned speakers, we decided on an approach that visualized the differences of two characters in terms of their direct speech. To do so, the words have been lemmatized and counted. This is done for two characters, e.g. Harry and Hermione from the Harry Potter novels. To understand where the differences between the characters lie, these two sets of words are then subtracted from each other, which results in two disjoint sets with no overlap.

To construct the visualization, we join the two sets and assign each word an importance score based the number of occurrences. This importance score is reflected in the size of the words, leaving the color to indicate another dimension of the data set. In our case, we decided to indicate the character that spoke the word by their color. Having set the size and colors of the words, the construction of the word cloud follows the classical wordle technique, along a spiral from the origin of the visualization canvas.

The example in Figure 16 shows the differences in direct speech of Harry and Hermione in “Harry Potter and the Sorcerer’s Stone”. White color indicates words attributed to Harry, and red the ones spoken by Hermione. We can see

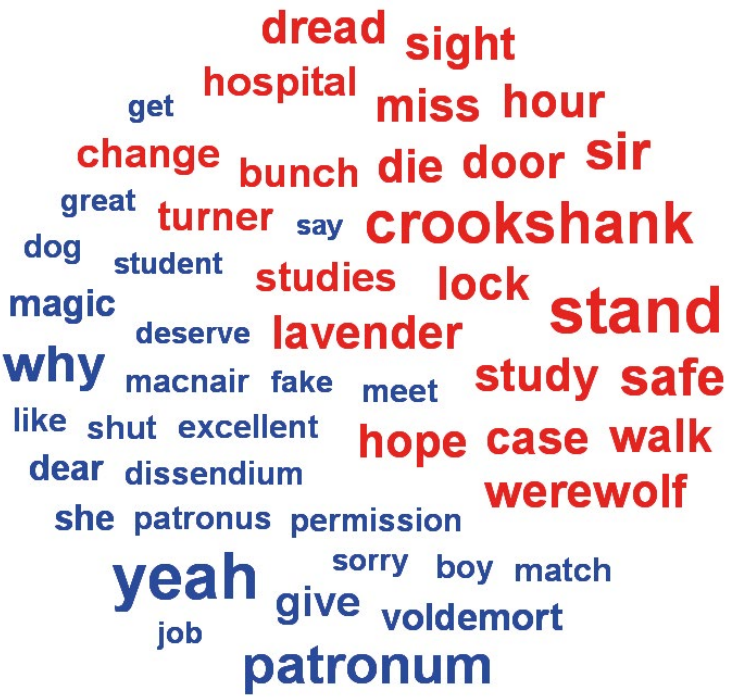

Figure 16: Word cloud displaying the most frequent words of direct speech from Harry (blue) and Hermione (red). Words occurring in direct speech of both characters have been removed from the data set.

that Hermione talks more about the studies and is also more concerned about safety. In contrast to Harry, she also mentions her cat, Crookshanks, multiple times. Harry, as is typical for the book, mentions Voldemort by name more often. He also uses comparatively more spells and often asks 'why'.

5.2 Extracting features on individual level: character's emotions, topics, actions

Related work

Nalisnick et al. (2013) analyze sentiment between characters in Shakespeare's plays. Their method is based on the assumption that a line of speech is directed towards the speaker of the previous line. A sentiment lexicon was then used to extract sentiment from the character's speech and measure the change of

sentiment between pairs of characters throughout the course of a play. Flekova and Gurevych (2015) use text classification techniques to predict personality traits of literary characters. Characters are classified as either 'introvert' or 'extrovert' types. This data is used to train SVM classifiers to predict these traits in unseen characters. Extracting features from the dialogue of the characters, a variety of features is used, including lexical, semantic and stylistic features, and the use of emotional language.

Methodology

For the purpose of analyzing individual characters, we use the same stylometric features as described in section 3, with the difference of applying them only on direct speech utterances of each protagonist separately rather than on the entire text of a book.

Visualization

To follow the emotional context of a character throughout a book, a visualization should provide the following insights into the emotional context of a character: what is the dominating emotion, and which emotions change and how drastic are these changes.

To determine the emotional context, the book is analyzed using a sliding window. For each window, the words from emotional categories, such as negative or positive emotions, the number of occurrences of words from these categories are counted and attributed to each character occurring in the window.

Compared to the overview visualization introduced in section **3.3**, the space-filling idea is discarded in favor of a flow- like visualization metaphor. This eases the task of perceiving the emotional change of neighboring emotional context, since, as well as the estimate of the amount of change, the area occupied by the emotion flow visualization also changes in relation to the overall amount of emotions. Each of the stacks is placed next to each other, as they occur in the book, to reflect the emotional change in the story. The transition between each of the stacks is displayed along a b-spline interpolation, which smoothes radical changes in the data (here: in the emotional changes), but still preserves a truthful transition between the contexts. For the emotional context, the areas corresponding to the different emotion categories are filled with distinct colors, which enables readers to easily follow an emotion category, as well as to effectively estimate the share of emotion categories per context.

In Figure 17, the emotion words connected to the main character, Harry, in the novel "Harry Potter and the Sorcerer's Stone" are shown. It is clearly visible that the number of words with a negative connotation dominate, while there is also a varying, but noticeable amount of "anger" words. In particular, it is striking that positive word classes, such as "joy" or "trust" occur only rarely in Harry's context.

In Figure 18, the emotion words co-occurring with Hermione are displayed. Besides the obvious insight that Hermione is appearing later in the book than Harry, the similarities between hers and Harry's emotion word context is quite clear.

Figure 19 shows a direct comparison of the emotional context (in terms of words from the NRC emotion dictionaries) of the two characters Strider (top) and Aragorn (bottom). For Strider, the context is indicated as being dominated by positive emotions, as well as a quite large extent of fear and negativeness. Aragorn, as the same character is referred to later, is missing a large amount of the positive extent, and his emotional context shows a larger influence of fear. This is in line with the story flow of the first Lord of the Rings book, where Aragorn is joining the fellowship of the ring and encounters, together with them, the Ringwraith. In contrast, Strider has a positive function as he offers help to the Hobbits in Bree, which can be seen in his emotional flow (Figure 19 top).

6. Conclusions and future work

In this chapter, we presented methods for extracting characteristics and features from book chapters, which can be used to approximate the components of story complexity from different aspects. In our model, we suggested that story complexity consists of three core areas: the complexity of the language used; complexity of the plot; and the intrinsic complexity of individual characters. We presented a range of Natural Language Processing techniques that enable initial insights into each of these areas and which can be further built upon.

Information visualization has been introduced as the method to make the different kinds of data visible and intuitively comprehensible to readers. Each of the visualization techniques is designed according to the characteristics of the available data, e.g. stylistic information of character co-occurrences, or the count of emotion words in a reference unit, for example, paragraphs. The visuals shown in this chapter are already highly specialized and tailored to the available set of features that in our opinion contribute to the whole ensemble, which we label as *story complexity*.

The next step in visualizing the different aspects would be the combination of different data, e.g. the co-occurrences of characters together with the emotion

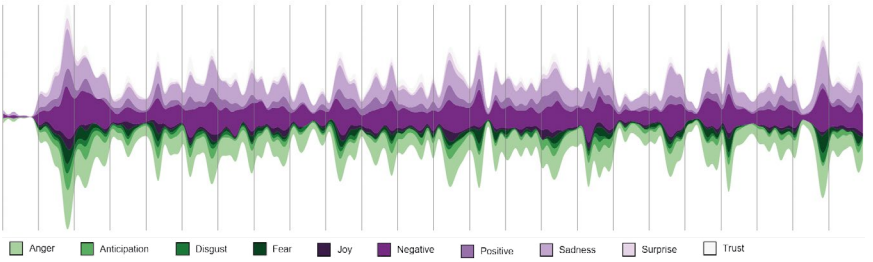

Figure 17: The flow of positive and negative emotions in “Harry Potter and the Sorcerer’s Stone” in the case of Harry. The vertical lines are for orientation purposes only.

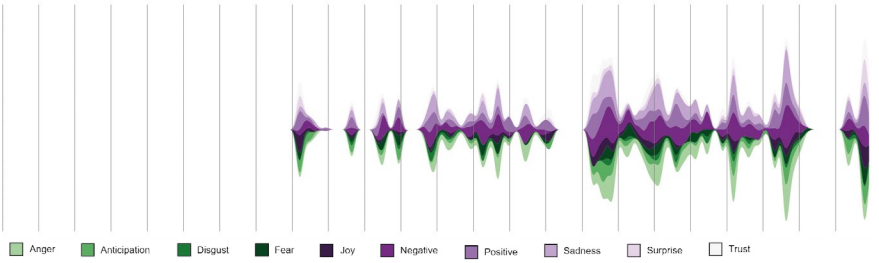

Figure 18: Emotion words in the context of Hermione during the first book of the Harry Potter series. The vertical lines are for orientation purposes only.

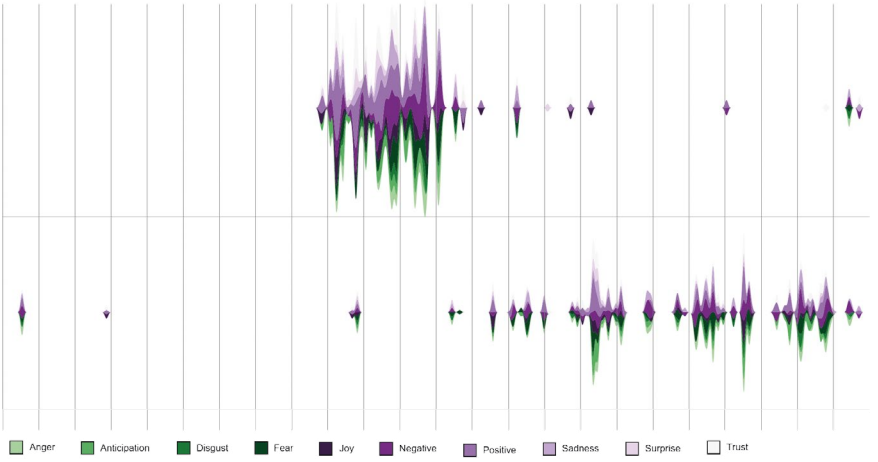

Figure 19: NRC emotion word categories appearing in the context of the character Strider (top) and Aragorn (bottom) in the first book of the Lord of the Ring series. The vertical lines are for orientation purposes only.

words, so that besides the fact that two characters occur together, a qualitative measure can be assigned. This could provide insight into books where a character changes his emotional context, and which gives a hint that his emotional profile might be more complex than that of other characters. In addition, more integrated views can open up new design spaces. More complex information spaces, such as projections based on the extracted features from multiple characters, can also give informative representations of the data, e.g. because of groups of entities or even the shapes formed by the entities.

From the Natural Language Processing perspective, literature still poses many challenges. Most of the text annotation models are focused on modern languages, such as those found in newspaper articles or even social media, and their adaptation to narratives which use notably more figurative language, such as more infrequent word expressions and sometimes an unusual syntactic structure, is challenging. For example, a named entity recognition model trained on Wikipedia is likely to produce very poor results when tried on classical novels. Development of more advanced methods tailored specifically to literature processing is required and exceeds the scope of this chapter.

7. References

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370.

Akaishi, Mina, Yoshikiyo Kato, Ken Satoh, Koichi Hori. 2007. "Narrative based Topic Visualization for Chronological Data." In *11th International Conference Information Visualization* IV: 139–144.

Ashok, Vikas Ganjigunte, Song Feng and Yejin Choi. 2013. "Success with style: Using writing style to predict the success of novels." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Chambers, Nathanael, and Dan Jurafsky. 2009. "Unsupervised learning of narrative schemas and their participants." In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:* Volume 2, 602–610.

Chen, Chaomei. 2006. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." *Journal of the American Society for Information Science and Technology* (57) 3: 359–377.

Ciaramita, Massimiliano and Yasemin Altun. 2006. "Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger." In

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06), 594-602.

Daelemans, Walter. 2013. "Explanation in computational stylometry." In *International Conference on Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh, 451–462. Berlin: Springer.

DeCamp, Philip, Amber Frid-Jimenez, Jethran Guiness, and Deb Roy. 2005. "Gist Icons: Seeing Meaning in Large Bodies of Literature." In *IEEE Info Visualization 2005 Conference.*

Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. "Discovering interesting usage patterns in text collections: integrating text mining with visualization." In *CIKM,* 213–222.

Dotson, Daniel. 2005. "Portrayal of physicists in fictional works." *CLCWeb: Comparative Literature and Culture* 11 (2):5.

DuBay, William H. 2006. *The Classic Readability Studies.* Costa Mesa, Cal: Impact Information.

Eagleton, Terry. 2005. *The English Novel: An Introduction.* Blackwell, Oxford.

Eder, Jens, Fotis Jannidis, and Ralf Schneider, Eds. 2011. *Characters in fictional worlds: Understanding imaginary beings in literature, film, and other media.* Berlin: de Gruyter (Revisionen, 3).

Elsner, Micha. 2012. "Character-based kernels for novelistic plot structure." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics: 634–44.

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. 2010. "Extracting social networks from literary fiction." In *Proceedings of the 48th annual meeting of the association for computational linguistics,* 138–147.

Fekete, Jean-Daniel, and Nicole Dufournaud. 2000. "Compus: visualization and analysis of structured documents for understanding social life in the 16th century." *ACM DL,* 47–55.

Feng, Lijun et al. 2010. "A comparison of features for automatic readability assessment." *Proceedings of the 23rd international conference on computational linguistics: Posters.*

Flesch, Rudolf. 1948. "A new readability yardstick." *The Journal of Applied Psychology,* 32 (3): 221–233.

Flesch, Rudolf. 1979. *How to write plain English.* Harper and Brothers, New York: Harper and Brothers.

Gillam, Ronald B., and N. Pearson. 2004. *Test of narrative language.* Austin, TX: PRO-ED.

Goyal, Amit, Ellen Riloff, and Hal Daumé III. 2010. "Automatically producing plot unit representations for narrative text." In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 77–86.

Greenhalgh, Kellie S. and C. J. Strong. 2001. "Literate language features in spoken narratives of children with typical language and children with language impairments." *Language, Speech, and Hearing Services in Schools*, 32 (2): 114–125.

Halpin, Harry, and Johanna D. Moore. 2006. "Event extraction in a plot advice agent." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,* 857–864.

He, Hua, Denilson Barbosa, and Grzegorz Kondrak. 2013. "Identification of speakers in novels." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1312– 1320.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. "Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships." In. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1534–1544. https://doi.org/10.18653/v1/N16-1180

Jankowska, Magdalena, Vlado Keselj, Evangelos E. Milios. 2012. "Relative N-gram signatures: Document visualization at the level of character N-grams." In *IEEE VAST,* 103–112.

Jayannavar, Prashant Arun, Apoorv Agarwal, Melody Ju and Owen Rambow. 2015. "Validating literary theories using automatic social network extraction." In *Proceedings of the NAACL-2015 Workshop on Computational Linguistics for Literature,* 32–41.

John Burrows. 2004. "Textual analysis". In *A Companion to Digital Humanities,* edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 324–347.

Johnson, John A., Joseph Carroll, Jonathan Gottschall, Daniel Kruger. 2001. "Portrayal of personality in Victorian novels reflects modern research findings but amplifies the significance of agreeableness." *Journal of Research in Personality* 45 (1): 50–58. https://doi.org/10.1016/j.jrp.2010.11.011

Kazantseva, Anna, and Stan Szpakowicz. 2010. "Summarizing short stories." *Computational Linguistics* 36 (1): 71–109.

Keim, Daniel A., and Daniela Oelke. 2007. "Literature Fingerprinting: A New Method for Visual Literary Analysis." In IEEE Symposium on Visual Analytics Science and Technology, 115–122.

Lee, John, and Chak Yan Yeung. 2012. "Extracting networks of people and places from literary texts." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.*

Lieber, Emma K. 2011. *On the Distinctiveness of the Russian Novel: The Brothers Karamazov and the English Tradition.* Diss. Columbia University.

Liu, Shixia, Yingcai Wu, Enxun Wei, Mengchen Liu, and Yang Liu. 2013. "StoryFlow: Tracking the Evolution of Stories." *IEEE Trans. Vis. Comput. Graph.* 19 (12): 2436–2445.

McCabe, Allyssa, and L. S. Bliss. 2003. *Patterns of narrative discourse: A multicultural, life span approach.* Boston: Allyn & Bacon.

McIntyre, Neil, and Mirella Lapata. 2010. "Plot induction and evolutionary search for story generation." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 1562–1572.

Mohammad, Saif. 2011. "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 105–114.

Mohammad, Saif. 2011. "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales." In *Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 105–114.

Monroy, Carlos, Rajiv Kochumman, Richard Furuta, Eduardo Urbina. 2002. "Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Documents." In *Visual Interfaces to Digital Libraries,* 39–49.

Morton, T., J. Kottmann, J. Baldridge, and G. Bierner. 2005. Opennlp: A java-based nlp toolkit.

Newman, Robyn M., and K. K. McGregor. 2006. "Teachers and laypersons discern quality differences between narratives produced by children with or without SLI." *Journal of Speech, Language, and Hearing Research* 49 (5): 1022–1036.

Oelke, Daniela, Dimitrios Kokkinakis, Daniel A. Keim. 2013. "Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature." *Comput. Graph. Forum* 32(3): 371–380.

Pitler, Emil, and Ani Nenkova. 2008. "Revisiting readability: A unified framework for predicting text quality." *Proceedings of the conference on empirical methods in natural language processing (EMNLP): 186–195.*

Rangel, Francisco, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. "Overview of the 2nd author profiling task at PAN 2014." In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers,* 898–927.

Rangel, Francisco, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. "Overview of the author profiling task at PAN 2013." In *CLEF 2013 Conference on Multilingual and Multimodal Information Access Evaluation,* 352–365.

Recasens, Marta, Marie-Catherine de Marneffe, and Christopher Potts. 2013. "The Life and Death of Discourse Entities: Identifying Singleton Mentions." In *Proceedings of NAACL-HLT* 2013*:* 627–633.

Regan, Tim, Linda Becker. 2010. "Visualizing the text of Philip Pullman's trilogy 'His Dark Materials'" In *NordiCHI,* 759–764.

Rohrer, Randall M., John L. Sibert, and David S. Ebert. 1998. "The Shape of Shakespeare: Visualizing Text using Implicit Surfaces." *INFOVIS,* 121–129.

Scott, Cheryl M. & Windsor, J. 2000. "General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities." *Journal of Speech, Language, and Hearing Research* 43 (2): 324–340.

Senter, R., and E. Smith. 1967. *Automated Readability Index*. Aerospace Medical Research Laboratories.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng and Christopher Potts 2013. "Recursive deep models for semantic compositionality over a sentiment treebank." In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*: 1631–1642.

Tanahashi, Yuzuru, Chien-Hsin Hsueh, and Kwan-Liu Ma. 2015. "An Efficient Framework for Generating Storyline Visualizations from Streaming Data." *IEEE Trans. Vis. Comput. Graph* 21(6): 730–742.

Tanahashi, Yuzuru, Kwan-Liu Ma. 2012. "Design Considerations for Optimizing Storyline Visualizations." *IEEE Trans. Vis. Comput. Graph.* 18(12): 2679–2688.

Ukrainetz, Teresa A., L. M. Justice, J. N. Kaderavek., S. L. Eisenberg, R. B. Gillam, and H. M. Harm. 2005. "The development of expressive elaboration in fictional narratives." *Journal of Speech, Language, and Hearing Research* 48: 1363–1377.

Vala, Hardik, David Jurgens, Andrew Piper, Derek Ruths. 2015. "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the Difficulty of Detecting Characters in Literary Texts." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 769–774.

van Ham, Frank, Martin Wattenberg, Fernanda B. Viégas. 2009. "Mapping Text with Phrase Nets." In *IEEE Trans. Vis. Comput. Graph.* 15(6): 1169–1176.

Wanner, Franz, Johannes Fuchs, Daniela Oelke, and Daniel A. Keim. 2011. "Are my Children Old Enough to Read these Books? Age Suitability Analysis." *Polibits* 43: 93–100.

Weber, Wibke. 2007. "Text Visualization – What Colors Tell About a Text."In *Information Visualization* IV: 354–362.

Weiler, Andreas, Michael Grossniklaus, Marc H. Scholl. 2015. "The Stor-e-Motion Visualization for Topic Evolution Tracking in Text Data Streams." IVAPP: 29–39.

Zhao, Ying and Justin Zobel. 2007. "Searching with style: Authorship attribution in classic literature." In *Proceedings of the thirtieth Australasian conference on Computer science,* 62: 59–68.