Automatic Narrative Structure Identification in Literary Texts

Nuette Heyns[1], Menno van Zaanen[2]

[1]North-West University,
South Africa
nuette.heyns@gmail.com
[2]South African Centre for Digital Language Resources,
South Africa
menno.vanzaanen@nwu.ac.za

Numerous tasks in the field of computational literary analysis require a high-level understanding of the structure of a text. The ultimate goal of this research is to create and leverage an annotated database to train a computer system capable of automatically annotating literary texts according to their structural components. By aligning human expertise with computational methods, this research aims to reshape our understanding and engagement with literature in the digital age, offering new insights and opportunities for exploration in the world of literary analysis. This research article introduces an approach to annotating literary texts according to their underlying structural elements. Our objective is to bridge the gap between human interpretation and computational capabilities by developing a set of guidelines for annotating literary texts based on their structural intricacies. This article also outlines the process of constructing a database of short stories annotated according to the guidelines defined in this article, which serves as a valuable resource for future studies. Additionally, it addresses the vital issue of inter-rater annotator agreement, ensuring the reliability and robustness of the proposed guidelines through multiple annotator involvement.

Keywords: narrative structure, computational literary analysis, scene segmentation, LDA

1 Introduction

In an era marked by technological advancements and digital transformation, the need for systematic and standardised methods of text analysis has become increasingly apparent. The ability to understand narrative trends by comparing different narratives on a large scale can benefit researchers in a wide range of fields.

Scholars from different disciplines, like linguistics, literary criticism, psychology, and sociology, have been studying story structures and their roles in our communication and understanding of the world. Recently, studies in economics, climate science, political polarisation, and mental health started using narratives to understand human behaviour (Piper et al., 2021). Fields related to computer science that also benefit from story structure analysis include video game development, bot development, summarization tools, automatic essay grading, or narrative popularity predictions.

Ouyang and McKeown (2014) refer to a couple of different uses of automatic narrative structure detectors. They argue that information about the narrative structure can provide a measure of the quality of the narrative. For example, orientating information is usually found at the start of a narrative. If it is placed too late in the narrative, it might be hard for the reader to understand the narrative. Ouyang and McKeown (2014) also argue that narrative generation systems can be improved by knowing the preferred structure of the narrative. For example, a narrative-generating system might present all orientating information at the start of a narrative, while human narrators might have preserved some information for later.

Structure theories are typically analysed using qualitative approaches or manual approaches that result in a fine-grained analysis. However, with the rise of digital text and computational processing power, large-scale analysis of texts has become possible. Although quantitative analysis of a text cannot replace the human expertise provided by a qualitative analysis, a quantitative approach can provide scholars with empirical evidence for a narrative structure theory and allows for large scale comparisons, among others.

The development of tools to analyse digital texts has been neglected mostly due to the lack of annotated data. Even the tools that have been developed are not always properly evaluated and tend to target a commercial market. Specifically, in an academic setting, the evaluation of tools is important to ensure the quality of the analysis generated by the tool.

This research will contribute to the field by creating comprehensive guidelines for annotating literature texts based on their structural intricacies. The guidelines are based on existing structure theories. We construct a database comprising three short stories annotated following our guidelines, thus establishing a valuable resource for future scholars.

Furthermore, we recognise the critical importance of human annotator agreement when applying structured annotations to literary texts. The article seeks to assess the inter-rater annotator agreement level by involving multiple annotators in the annotation process, ensuring the reliability and robustness of our guidelines.

Our objective extends beyond the theoretical domain into the practical application of these guidelines. Ultimately, this research will bridge the gap between human analysis and computational capabilities. We will use the annotated database to train a computer system that can automatically annotate literature texts according to their structural components. This development holds significant potential for literary analysis and applications in various fields. The system could also provide scholars with empirical evidence for a narrative structure theory.

2 Background

A narrative does not only refer to the narrative in a novel, but it extends to a wide range of fields. We use narratives when we communicate historical events, political agendas,

race, religion, identity and time (Wake, 2006). All of these notions explain the way the world is experienced through an individual's narrative thereof. This definition of the narrative does, however, not include all forms of text or spoken word. Prince (1982) explains that a narrative has specific traits that distinguish it from other forms. Prince (1982) claims that we have "internalised certain rules about what constitutes a narrative and what does not". In general, a narrative is the "representation of real or fictive situations and events in a time sequence" (Prince, 1982). A narrative can thus be described as specific elements following specific rules in a specific order. The study of these elements and the order in which they are found is called narrative structure theory.

As we have mentioned, the narrative consists of elements like situations and events. A situation refers to the circumstances or context in which events, actions, or interactions take place within a narrative. Situations provide the backdrop against which characters operate and events unfold. They encompass the setting, the characters' relationships, their motivations, and the broader societal or cultural context. Situations can be dynamic and subject to change as the story progresses, influenced by the characters' decisions and external forces.

Events typically refer to occurrences or happenings within the narrative that drive the plot forward, shape character development, and contribute to the overall progression of the story. On the other hand, the absence of events, or what Prince (1982) refers to as non-events, can also hold importance in narrative structure and meaning. Non-events, refer to instances where something expected or anticipated does not happen within the narrative. The absence of anticipated events can create tension, suspense, or surprise, depending on the context. Non-events can be just as meaningful as actual events in shaping the narrative structure and character development. Examples of non-events include a character's failure to act in a crucial moment, a planned meeting that does not occur, or the absence of an expected outcome. While events propel the narrative forward, non-events add nuance, complexity, and unpredictability to the storytelling process, enriching the reader's experience and inviting deeper interpretation.

Narrative structure describes how different parts of a narrative relate to one another. A paradigmatic approach to narrative structure focuses on the underlying structure of a text without taking into account the linear order of events (Propp, 1968). The structure of a text is not necessarily linear but can be affected by changes in location or time, or by actions or events (e.g., flashbacks or flash-forwards). The structure determines how the plot of a story unfolds and has an effect on how it is perceived. Connecting elements of a narrative through the structure of a text leads to the development of a story line.

A longstanding tradition in literary criticism is to analyse narratives according to plot elements or thematically similar units centred around the actions of characters. Propp (1968) argued that all Russian folktales can be characterised by the combination and organisation of 31 basic plot motifs.

However, there is a superfluity of different narrative structure theories all aiming to describe the complicated nature of narrative form. These theories are based on the claim that regularities can be found in narrative structures. These theories aim to organise narrative elements, including the introduction, rising action, climax, falling action, and resolution. We will look at several well-known structure theories to demonstrate the overlap between sections. Ultimately we will focus on two of these theories, the three-act and the four-part structure theories. We want to apply these theories to our text to evaluate which theory suits our suggested genre, short mystery stories

better. Short stories are generally believed to follow a three-act structure, whereas the mystery genre tends to follow a four-part structure. When combining short stories and the mystery genre, which will determine the structure of a text, if there is indeed a common structure?

2.1 Structure theories

Structure theories aim to describe the combination and/or order of events in a text, by dividing a text into different sections (sometimes referred to as acts or parts). Each section contains events with similar goals. Many structure theories often share common elements or concepts, resulting in overlapping sections. This means that while each theory may have its unique perspective or emphasis, there are areas where they intersect or coincide in their analysis of narrative structure. These overlapping sections indicate common ground among different theories, highlighting key aspects of storytelling that are widely recognised within the field.

2.1.1 The six-part structure

It is worth mentioning Labov (2013) further extended their three-act structure theory to include the abstract, resolution, and coda. This creates a six-part structure consisting of the *abstract* that serves as an introduction to the story and often contains a description of the core event. The *orientation* act introduces the setting and characters by providing background information. The *complicating action* consists of a chain of events that lead up to the core event. The *evaluation* of the story is where the author attempts to provide credibility to the story, present alternative outcomes or assign praise or blame to characters. The *resolution* extends the chain of events to a final resolution of the situation. The *coda* relates the events of the narrative to the present. We will not use this structure theory in this work as short stories do not generally include an abstract and coda.

2.1.2 Freytag's pyramid/five-point dramatic structure

Another expansion of the three-act structure was introduced by Freytag (1894). This structure is usually used in Greek tragedies or tragic contemporary tales. The *exposition* introduces the setting and characters. In the *rising action* the basic conflict and obstacles are introduced. During the *climax* there is a turning point that effects a change for the better or for worse. During the *fall of action*, unexpected incidences can create suspense, but overall, the conflict is resolved. During the *resolution* the protagonist achieves or fails their goal.

Similarly, Todorov Todorov (1971) identifies five stages in a narrative. *Equilibrium* shows a glimpse of the daily routine that a character follows at the start of a narrative. During the *disruption* stage, the daily life of the character is disturbed. *Recognition* is where the character realises the cause of the disruption. In the *repair* stage the character tries to solve the problem that causes the disruption. At the end of the narrative, the character returns to the *equilibrium stage* where they continue or adjust to daily life.

2.1.3 The seven-point structure

The seven-point story structure is a relatively new structure compared to the ones discussed until now. It is commonly identified as the structure used in the Sci-fi genre and was popularised by Wells (2010) after claiming that he used the structure in his

novels. The *hook* is the introduction to the setting and characters. The *call to action* is driven by an event that sets the protagonist on an adventure. The *first setback* is usually due to the introduction of the antagonist or the major conflict. The *turning point* is where the protagonist starts taking action. In the *second setback*, the conflict increases and the protagonist feels discouraged. Some new element is introduced in order for the protagonist to *solve the problem*. A *resolution* is reached when the major conflict is resolved and the antagonist is defeated.

2.1.4 The Hero's Journey

Campbell (2008) base the Hero's journey on common motifs and themes found in world mythology and folklore, suggesting that these narrative patterns resonate deeply with human experience and psychology. Campbell's work in identifying and popularising this structure has had a significant influence on storytelling across various mediums, from literature and film to video games and beyond. It outlines the stages a protagonist typically undergoes as they embark on an adventure or quest, facing challenges and transformation along the way.

The Hero's journey consists of 12 stages. In the *call to adventure* the hero is introduced to a problem or challenge that sets them on their journey. This could be a direct call or an unexpected event that disrupts the hero's ordinary life. Initially, the hero may hesitate or refuse the call to adventure due to fear, doubts, or obligations to their current life. This stage is known as the *refusal of the call*. After refusing the call, the hero will *meet the mentor* or guide figure who provides advice, tools, or wisdom to help them on their journey, resulting in the hero committing fully to the adventure or *crossing the threshold* and leaving behind their familiar world, entering into the unknown or unfamiliar territory. In the stage of *tests, allies, and enemies* the hero faces various challenges, meets allies, and confronts adversaries as they progress on their journey, learning important lessons along the way. The hero approaches a significant challenge or ordeal, often represented as a metaphorical "cave" or inner conflict that they must confront in the *approach to the inmost cave*. Here the hero faces their greatest trial or battle in what is known as the *ordeal*, undergoing a profound transformation or revelation as they overcome this obstacle. After overcoming the ordeal, the hero receives a *reward*, which could be an object, knowledge, or insight that will help them in their ultimate goal. The hero begins the *journey back* to their ordinary world, but they may encounter further obstacles or face the consequences of their actions. The hero experiences a final moment of death and rebirth in the *resurrection*, symbolising their ultimate transformation and growth. The story ends with the *return with the elixir* stage, where the hero returns to their ordinary world, bringing back the knowledge, experience, or boon gained from their journey, which benefits themselves and/or their community.

2.1.5 The three-act structure

The three-act structure divides the story into three parts. This theory is often credited to Aristotle, however, the three-act structure we are familiar with today has been expanded on by multiple theorists. A popular version of this theory was proposed by Labov and Waletzky (1967). The first act is called the *orientation*. This act introduces the setting and characters as well as the basic conflict. The second act is used to describe the *complicating action*. Usually, the protagonist tries to solve a problem, only to raise the conflict further. The third act is the *evaluation* of the story, where the conflict is

resolved. Steele (1981) notes that many commentators have noticed the Aristotelian properties of classic detective stories. This is the main reason we chose to use the three-act structure theory for this work, as we will be annotating detective stories in this experiment.

2.1.6 The four-part structure

The four-part structure cannot be attributed to a single source, but several theorists have contributed to our understanding of the four-part structure. Similar to the five-part structure attributed to Freytag (1894) and the six-part structure, the four-part structure is also derived from the three-act structure mainly attributed to Aristotle, but it divides the second act into two parts. Most four-part structures can easily be placed into a three-act structure and vice versa. Mystery genres are also described as having a four-part structure. The *introduction and setup* introduces the protagonist and a crime. The protagonist decides to solve the crime. During the *discovery phase*, the protagonist looks for clues. The protagonist might encounter obstacles in this phase that can lead them to become puzzled or discouraged. In the *funnel* phase, new evidence comes to light and subplots are closed. In the *reveal*, the villain is confronted, the crime is explained and a conclusion is reached.

3 Previous and related work

Research on the computational analysis of narrative structure includes Jockers (2015) who aims to identify the change in sentiment over a text, while Finlayson (2016) and Droog Hayes et al. (2018) present a case for inferring Propp's functions using machine learning methods. Finlayson (2016) mentions that even though Propp described the functions in detail, there is still some ambiguity when annotating text according to functions. They argue that the description of the function groups is at times unclear, implied, or in disagreement with each other. This caused a greater variation in human annotations of the text. They solved this problem by creating panels that discussed the text until they were in agreement. Semantic role labelling was also performed by the annotators using the PropBank scheme. The first step of their algorithm is to extract a timeline from the annotations. They then associate each event with an agent and patient using the semantic role. This approach is extremely resource-intensive and requires a lot of human judgment.

A number of studies (Barth, 2021; Ketschik et al., 2019; Reiter et al., 2019; Wiren and Ek, 2021) have created guidelines for human annotation of narrative boundaries to train and test models to identify the start and end of scenes within a text. These studies define scene changes as the change in narrator, event, time, place, or the constellation of characters. They base their work on Jahn (2005) who in turn adapts an idea from Genette et al. (1980) where they distinguish between different levels of narration. According to Genette's theory, narrative levels occur when there are stories within stories. Their work aims to identify scene boundaries as well as the relation between sub-stories in a text in terms of their narrative level. Here our work differs in terms of what we consider a segment. Where these studies consider a segment to consist of a single scene, we propose a segment consisting of one or more scenes that have a common purpose. We are interested in the order in which events and scenes are clustered together to create a segment as well as the possible repetition of a segment in the same text or across multiple texts in the same genre.

Ouyang and McKeown (2014) introduced a system to automatically detect one of the narrative structure elements based on William Labov's theory of narrative analysis (Labov and Waletzky, 1967) as mentioned in Section 2.1.5. They present experiments where they detect the Complicating action element in Labov's theory with a 71.55 f-score. They make use of a corpus consisting of twenty oral narratives collected by Labov (2013) as mentioned in Section 2.1.1. Although this work has the same goal as ours, they rely on properties of speech data to identify boundaries. This technique would not work with text data.

In previous work (Heyns and van Zaanen, 2021, 2022), we describe a method to identify high-level textual transitions in text using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). A transition is found when there is a relatively large contrast between LDA topics before and after a position in the text. This means that a transition is not simply found when the LDA topic changes, but rather when a group of LDA topics that co-occur changes. For this study, we are considering the segments created by these transitions to function as a story act. We specify the number of segments to divide the text into. We then measure the distance between corresponding segments of different text. This creates a tool to measure the applicability of a structure theory on a text and decide where to place narrative transitions.

Similar studies (Gius et al., 2021) have shown that inter-rater annotation can vary greatly because of the subjective and context-dependent nature of a text. Reiter (2015) has found that summarising a text before attempting to annotate it, helped to create greater agreement between manual annotations.

4 Methodology

We will use three short stories from the mystery genre. In Section 5 we will discuss the development of a database where five human annotators will annotate the text. We will rate the agreement between human annotations. In Section 6 we will describe the system we will use to analyse the text. We will then compare the human annotations with the system annotations.

5 Data

Gius et al. (2019) argues that manually annotated datasets are imperative for developing automatic scene or narrative detection systems. This work will take inspiration from Gius et al. (2019) by establishing rules to identify sections in a text. We will create a human-annotated dataset that will also benefit future studies in automatic scene and narrative detection. This dataset will serve as a baseline for comparing the automatically identified sections using the system described in Heyns and van Zaanen (2022).

5.1 Data collection

For now, we will focus on mystery short stories. Because we have to manually annotate the data, we will keep the dataset relatively small. We will be using three short stories from Agatha Christie. The length of the text can affect the ideal number of topics to use. For now, we will use short stories of similar length to eliminate that variable.

Table 1: Number of words (Word), number of segments (Seg), and average number of words per segment (Word/seg) for each short story.

Text	Word	Seg	Word/seg
The adventures of the Italian nobleman	2409	66	35.97
The jewel robbery at the grand metropolitan	3495	94	36.05
The mystery of Hunter's Lodge	2811	77	35.25

5.2 Data processing

Pre-processing is done on the text. Stopwords are removed using NLTK[1]. The texts are lowercased, lemmatized, and the punctuation is removed using spaCy[2].

Two of the texts (*The adventures of the Italian nobleman* and *The Mystery of Hunter's Lodge*) contain section breaks to indicate a lapse of time. This section break was removed during pre-processing, although human annotators would have had access to that visual clue.

To apply LDA to the text, we split the text into smaller segments. If a sentence consists of more than 30 words, it is used as a segment; if not, sentences are concatenated to create segments of 30+ words. This is done so that each segment is of a similar length, but it does not interrupt a sentence. Table 1 shows the short stories we chose to use as well as their length. The average segment length is 35.75 words.

5.3 Data annotation

We asked five human annotators to annotate the text. The purpose of the annotation is to identify the different sections in the story structure, as well as explain the reasoning behind grouping sections. The annotator should read through the short story and decide where to add a transition that will break the story into different parts. The annotation should include a summary of the section, following the reasoning of Reiter (2015) that summarising the text before annotating it helps create a greater agreement between manual annotations. For this study, we will only include events, simplifying the nature of a literary narrative by leaving out the nuanced information that a non-event can provide for a literary analysis.

There will be two rounds of annotation. For the first round annotators should divide the story into three parts by stipulating the line number where a boundary should be placed. The annotator should identify the line number where a phase starts and ends. Note that each phase (except the first) will start on the next line to where the previous phase ended. For example, if phase 1 ends on line 24, phase 2 will start on line 25. They should also give a summary of each part, followed by a motivation of their reasoning for placing a boundary on a specific line. For the second round, the same annotations should be given, but the annotator should divide the story into four parts.

5.3.1 First round: three-act structure

The three-act narrative structure is a framework that many storytellers use to construct a short story. The three parts generally consist of:

[1] `https://www.nltk.org/`
[2] `https://spacy.io/`

1. Exposition: The main character and setting are introduced. An event (in the case of a mystery, a crime/mysterious event) sets the story in motion.

2. Rising action: The protagonist is drawn into the event; for example, the protagonist starts to investigate the crime/mystery. Clues are revealed. During this phase, the protagonist might encounter obstacles that will cause the protagonist to become puzzled or discouraged.

3. Climax and resolution: The villain is confronted, the crime is explained, and a conclusion is reached.

5.3.2 Second round: four-part structure

The four-part narrative structure is a framework that many storytellers use to construct a mystery novel. The four parts will consist of:

1. Introduction: The protagonist and a crime or mystery are introduced. The protagonist decides to investigate the crime or mystery.

2. Discovery: In this phase, the protagonist investigates the crime or mystery. Clues are revealed. The protagonist might encounter obstacles that will cause the protagonist to become puzzled or discouraged.

3. Funnel: During this phase, new evidence comes to light or the protagonist examines old evidence with a new approach. Subplots are closed. The protagonist narrows the suspect down to one, but the villain is not yet revealed to the reader.

4. Reveal: The villain is confronted, the crime is explained, and a conclusion is reached.

Table 2 gives an example of a completed annotation form.

6 System

In previous work (Heyns and van Zaanen, 2022), we have proposed a system to automatically identify transitions in a text. Here, transitions are considered to be a location in the text where there is a relatively large change in the composition of topics. The text is first split into sentence-length segments. Topics are then identified in the text using LDA.

To identify the best transition points, the overall entropy of the distribution of LDA topics is computed and compared against the combined entropies of both parts in case a transition is inserted. This information is combined as information gain. Transitions are placed at the highest information gain position in the text. This will be the position with the most contrast between the LDA topics before and after a transition.

Note that the system does not simply identify a position where the LDA topics change, but rather where a group of LDA topics that co-occur, change.

We concatenate all the text in the corpus, that has been split into segments. We then train an LDA model that represents the entire corpus. Therefore an LDA topic is assigned to each segment, resulting in a sequence of LDA topics: $LDA(S) = \langle LDA(s_1), LDA(s_2), \ldots, LDA(s_n) \rangle$. Heyns and van Zaanen (2022) have experimented with different numbers of LDA topics. The sequence for each text in the corpus is then separated so that the transitions can be identified in each text. We rely on a

Table 2: Example annotation for The Kidnapped Prime Minister by Agatha Christie. Phase (P), start line (S), end line (E), summary and motivation are provided.

P	S	E	Summary	Motivation
1	1	10	The story begins with the Prime Minister of a country being kidnapped by a group of terrorists. We are introduced to the main characters, including the Prime Minister, the leader of the terrorists, and the members of the Prime Minister's security detail who are working to rescue him. We learn about the political tensions in the country and the reasons behind the kidnapping.	An event that sets the story in motion. The main characters are introduced. The political setting is introduced.
2	11	21	The main part of the story focuses on the efforts to rescue the Prime Minister. The security team follows clues and tracks down leads in their search for the Prime Minister, while the terrorists do everything in their power to keep him hidden. There are several close calls and tense moments as the two sides engage in a game of cat and mouse. As the story progresses, the stakes get higher and the tension increases, leading to a climactic moment where the security team finally locates the Prime Minister.	The protagonist is drawn into the event trying to rescue the Prime Minister. Clues are revealed.
3	22	50	The story concludes with the successful rescue of the Prime Minister and the capture of the terrorists. The Prime Minister is reunited with his family, and the country celebrates his safe return. The resolution also includes a moral lesson about the importance of strong security measures and the need to work together to combat terrorism.	The terrorists are captured, a conclusion is reached.

method from previous work (Heyns and van Zaanen, 2021) that has been expanded on in Heyns and van Zaanen (2022) to recognise multiple transitions in texts using an information gain-based method. We specify the number of transitions that we want to identify in the text beforehand and the algorithm finds the optimal place for each transition, where the most information gain is obtained. The transitions divide the text into sections. The number of sections that a text is divided into represents the structure of the text and can be compared to one of the story structure theories mentioned in Section 2.1 of this article.

For example, if a text is divided into three sections, then we can test the three-act story structure of that text. For each text sequence (with $x = 1 \ldots n-1$) every position between two segments (s_x, s_{x+1}) is considered as a potential transition point. The information gain at each potential position is computed and the position with the highest information gain is considered the real transition point. If multiple positions have the same information gain, the system selects the first position. If multiple transitions have to be identified, the system will divide the sequence in two at the transition point. The algorithm is repeated on the side with the least information gain.

The system developed in Heyns and van Zaanen (2022) uses no manual annotation, which is ideal when trying to analyse narratives on a large scale. However, it is difficult to evaluate the accuracy of proposed transitions without comparing them to human judgement.

6.1 Number of LDA topics

In previous work (Heyns and van Zaanen, 2021) we have found that increasing the number of LDA topics generally decreases the performance of the system. As the number of LDA topics increases, the system will, more often, propose the wrong transition leading to a misplaced section more frequently. We have previously (Heyns and van Zaanen, 2021) established that the system performs best with two LDA topics and gradually deteriorates as more topics are introduced.

We can, however, expect that the optimal number of topics may change according to the length of a book. Longer books will generally discuss more topics than a short story. However, for this experiment, we will be using the results of two LDA topics to make comparisons.

Uglanova and Gius (2020) argues that topic modelling is used as a statistical tool to break down the themes of a text, but in fact, it is only recognising the statistical patterns within a text's structure. Therefore, the topics relate to the structural elements of a text as opposed to the content of the text. van Zundert et al. (2022) further supports this idea by asserting that the topics generated through topic modelling, are more aligned with genre signals rather than semantic fields. van Zundert et al. (2022) found that contextual details from a literary text, such as the location, time, language community, or cues from the author, collectively influence the LDA topics. So, in fact, we might not be measuring narrative structure, but rather attributes that occur around a narrative change. We see words like [evening, long, night, sky, dark] that indicate time, and words like [door, house, window, world, side] indicating location, that co-occur in topic 1. Meanwhile, we find words like [hear, find, look, answered, understand, ask, remark] indicating actions and words like [inspector, police, friend, woman, doctor, father] indicating people that co-occur in topic 2. These topics are intertwined in the text, so by no means can we say a specific section is only focusing on one of these topics. The system we propose here simply finds the location where the order of these topics gives the most information gain. For example, for the three-act structure of *The*

jewel robbery at the Grand Metropolitan, topic 2 dominates the section by 69%, 65% of segments in section 2 are tagged as topic 2 and 57% of segments in section 3 are tagged as topic 3.

topic 1: [felt, evening, green, long, night, sky, mind, dark, great, fact, door, house, sound, good, strange, thing, window, world, side, life]

topic 2: [inspector, police, dead, interest, case, friend, voice, woman, doctor, father, hear, find, face, look, answered, round, understand, eye, ask, remark]

7 Evaluation

We compare the transitions identified by the system, with human-annotated transitions. We would like to know if the system agrees with human judgment. We compare the automatically identified sections with human-annotated sections using a distance metric. We compare the start and end points between the automatically annotated sections and the human-annotated sections.

To take into account the distance between the two transitions we use the Root Mean Square Error (RMSE). RMSE allows us to measure the distance between a predicted value and an observed value. In this case, we will see the computer-identified transition point as the predicted value and use the human-identified transition point as the actual value, to determine how well the computer and human annotators compare. This allows us to take the distance between the transition points into account instead of just ruling a transition as wrong when it is not the same. RMSE is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - p_r)^2}{n}}$$

where n is the number of observations, where each human annotation counts as a separate observation. p_i is the position of the proposed transition position (which can range from one to the number of lines in the text), and p_r is the position of the real transition (in this case, the line number of the human-identified transition). To calculate the RMSE for multiple texts, the RMSE is computed for each transition and these values are combined using the average. A higher RMSE value indicates a lower agreement between the predicted and actual value. The scikit-learn Python package[3] was used to calculate the RMSE.

7.1 Inter-rater annotator agreement

We compare different human annotations of the text to assess the agreement between their annotations.

From Table 3 we can see that human annotators reach a greater agreement with the three-act structure theory than the four-part structure theory. This is unsurprising seeing that as the number of parts to annotate increases, the chances of making a mistake also increase.

To understand if the difference in agreement is significant, we added a random baseline where we divided the text into three and four equal parts. We then compare the RMSE between the baseline of each theory and the annotator's transition placement. Table 4 shows the average RMSE of all annotators compared to the baseline. Here we see that there is no real difference in the average annotation standards when compared

[3] `https://scikit-learn.org`

Table 3: The average RMSE between annotators for the individual texts as well as the average over the three texts. Results are provided for both the three-act and four-part analysis. The lower the RMSE, the higher the agreement is between annotators.

Text	Three-act	Four-Part
The adventures of the Italian nobleman	0.014	0.084
The jewel robbery at the Grand Metropolitan	0.147	0.317
The mystery of Hunter's Lodge	0.020	0.180
Average	0.061	0.193

Table 4: The average RMSE between the baseline and the annotators for the individual texts as well as the average over the three texts. Results are provided for both the three-act and four-part analysis. Higher RMSE values indicate larger distances between the location where the annotator placed a transition and where the random baseline placed the transition.

Text	Three-act	Four-Part
The adventures of the Italian nobleman	0.383	0.333
The jewel robbery at the Grand Metropolitan	0.557	0.564
The mystery of Hunter's Lodge	0.448	0.518
Average	0.463	0.472

to a baseline. Although if we look at each annotator individually in Table 5, we see a bigger difference in agreement between annotators for the four-part theory, than for the three-act theory. This difference is mostly just for the *The jewel robbery at the Grand Metropolitan*. Results from Table 3 also show that for both structure theories, the annotators struggled most with *The jewel robbery at the Grand Metropolitan*. If we look at the text, section breaks were added to *The adventures of the Italian nobleman* and *The Mystery of Hunter's Lodge* to indicate a lapse of time. This visual indication is where the annotators placed their transition in the three-act structure, and it also seemed to guide their transition choice in the four-part structure. However, *The jewel robbery at the Grand Metropolitan* does not contain such section breaks, and this can explain why the annotator agreement is lower for this text. Overall, the annotator agreement is high with an average RMSE value of 0.06 for the three-act structure and 0.19 RMSE value for the four-part structure.

7.2 Human annotator versus computer annotator

Table 6 shows the average RMSE between the human annotators and the computer annotator. Here we see a similar trend where *The jewel robbery at the Grand Metropolitan* shows a higher RMSE value that indicates a lower agreement between annotators and the system. Since the section breaks were removed during pre-processing, it could not have played a role in this case. This might have to do with the text itself. Perhaps the transition points were not as clear-cut in this text.

On the other side of the spectrum, if we look at the first text, we can see that the RMSE value is the lowest for both structure theories, which also agrees with the results of the inter-rater annotator agreement.

In general, the system performs similarly to human annotators with a 0.09 RMSE value for the three-act structure and a 0.144 RMSE value for the four-part structure.

Table 5: RMSE per annotator compared to the random baseline. RMSEs higher than the average are in red, indicating that the specific annotator added the transition further away from the baseline than other annotators. The blue RMSEs indicate the annotator added the transition closer to the baseline transition. This does not mean that the annotator is right or wrong for adding the transition at a specific place, but rather just indicates how close their annotation compares to a random baseline.

Text	Annotator				
	1	2	3	4	5
Three-act structure					
The adventures of the Italian …	0.383	0.386	0.393	0.367	0.386
The jewel robbery at the Grand …	0.500	0.634	0.513	0.524	0.614
The Mystery of Hunter's Lodge	0.473	0.440	0.440	0.446	0.440
Average	0.452	0.487	0.449	0.446	0.480
Four-part structure					
The adventures of the Italian …	0.302	0.305	0.308	0.282	0.469
The jewel robbery at the Grand …	0.542	0.412	0.644	0.809	0.411
The Mystery of Hunter's Lodge	0.585	0.355	0.632	0.408	0.608
Average	0.477	0.358	0.528	0.500	0.496

Table 6: The average RMSE per structure theory where a higher RMSE value indicates a lower agreement between annotators and the system.

Text	Three-act	Four-part
The adventures of the Italian nobleman	0.044	0.070
The jewel robbery at the Grand Metropolitan	0.159	0.224
The mystery of Hunter's Lodge	0.073	0.139
Average	0.092	0.144

Table 7: The average RMSE per section where a lower RMSE shows a higher agreement between annotators and the system per section.

Text	Section		
	1	2	3
Three-act structure			
The adventures of the Italian nobleman	0.058	0.021	–
The jewel robbery at the Grand Metropolitan	0.229	0.071	–
The Mystery of Hunter's Lodge	0.078	0.068	–
Average	0.122	0.053	-
Four-part structure			
The adventures of the Italian nobleman	0.068	0.139	0.021
The jewel robbery at the Grand Metropolitan	0.229	0.314	0.113
The Mystery of Hunter's Lodge	0.078	0.187	0.154
Average	0.125	0.214	0.096

7.3 Comparing structure theories

If we compare the three-act and four-part structure results with each other, both the human annotators and the system show better results for the three-act structure theory. The three-act structure seems to fit this type of text better.

7.4 Comparing sections

We can see that there are specific sections that are easier to identify than others. We expected the first section to have a lower RMSE value, due to the first section always starting with the first sentence of a text. Because the start point is consistent only the transition point will cause a fluctuation in the distance between transitions. For later sections, the start position will depend on the transition point in the previous section. Therefore, it is expected that there will be more variation in later sections. However, we see that this is only the case in two of the texts. For the three-act structure, all sections can be identified similarly easily, except for the first section of *The jewel robbery at the Grand Metropolitan* where the RMSE value is disproportionately high compared to the other sections.

For the four-part structure, the first section is easier to identify for two of the text. Here we again see that *The jewel robbery at the Grand Metropolitan* has a higher RMSE value for the first and second sections. In the third section, it seems to compare well with the other two texts. The first and third sections have a lower average RMSE than the middle sections, implying it is easier to identify the introduction and resolution of a crime than to determine exactly when the investigation starts and ends. This information has the potential to tell us something about the specific text, how the plot unfolds and how it differs from other texts in this genre.

The hierarchical approach used is "local", which may cause a preference for narratives featuring a two-part structure, allowing for further subdivision of one part. Consequently, this preference may lead to a sub-optimal initial split for a three-act structure. To assess potential bias towards either structural theory, the system is compared against a random baseline. Table 8 shows that the system does place three-act

Table 8: The RMSE between the system and a random baseline to evaluate the system's bias for a specific structure theory. A higher RMSE value indicates a lower agreement between the system and baseline.

Text	Three-act	Four-part
The adventures of the Italian nobleman	0.411	0.311
The jewel robbery at the Grand Metropolitan	0.491	0.457
The mystery of Hunter's Lodge	0.451	0.442
Average	0.451	0.302

structure transitions slightly further away from the baseline transitions than for the four-part structure.

8 Discussion

Computational literary analysis relies heavily on the ability to identify the high-level structure of a text. To do this, we first need to identify transitions within the text. It is important to refer to existing theories and research when deciding on what type of transitions to identify.

Some work in this field (Barth, 2021) aims to identify scene transitions. Other work (Jockers, 2015) base their transitions on the sentiment in the text. We follow Heyns and van Zaanen (2022) and Ouyang and McKeown (2014) in using structure theories to identify transitions in a text.

In this article, we review specific theories behind the structure of a text, that help provide clear and actionable guidelines for understanding and analysing narrative structure computationally. According to these theories, we create guidelines for annotating a text according to its structural elements. We also create a small dataset, using these guidelines and comparing the inter-rater annotator agreement. In contrast to previous work using this system, we now use human-annotated data to test the system.

The task of annotating data according to its structure is time-consuming, but this experiment showed promising results by proving that the system can identify transitions similar to human annotators.

9 Future work

Overall the system reflects very similar results to that of human annotators when identifying the overarching structure. However, the structure may becomes more complicated. For instance, when there are sub-stories told within the narrative. These sub-stories can also follow a structure different from the main narrative, or they may or may not have an effect on the main story line. Although we can use the same method to identify more sub-plot sections in a story, confirming that a section is a sub-plot proves a little more challenging.

Furthermore, structure theories often overlap in sections that share a common purpose as shown in Section 2.1. There also exist multiple theories with a similar number of sections, but the description/purpose of each section might be different. To that end, it might be beneficial to not only divide the text into sections but also identify the purpose of each section.

In Section 5.3 we refer to the fact that we are only taking into account the events that take place in a narrative and that this simplifies the structure of a narrative by excluding the significance of non-events or dialogue-text. Dialogue-text specifically can reveal conflicts through characters' speech and interactions. The pace may also be affected by dialogue, as dialogue tends to have a faster pace compared to narrative text which may involve more descriptive passages and a slower pace. For now, this falls outside of the scope of this project. But it remains an interesting concept to explore in future work.

In the current work, we create segments on a sentence level to avoid the unlikely event of a transition mid-sentence, identified by the system. However, this might not be the optimal length for the LDA model to identify a topic. We can consider concatenating multiple sentences to create a segment, but the usefulness of this will need to be addressed in future research that focuses on the fine-tuning of the model or the use of LDA for this purpose.

There is evidently still a substantial amount of work remaining on this topic. We will leave this for future investigation. As for the near future, we will move on to investigating the characteristics of these plot points.

10 Conclusion

In conclusion, this article presents a computational approach to annotating literary texts based on popular structure theories. This approach is motivated by the need for systematic and standardised methods of literary analysis in the digital age. In the article, we have outlined a set of guidelines for annotating literary texts based on their structural elements, and have described the process of constructing a collection of short stories annotated according to these guidelines. This collection serves as a valuable resource for future studies, and it has been used to assess the inter-rater annotator agreement, ensuring the reliability and robustness of the proposed guidelines.

The ultimate goal of this research is to leverage the annotated database to train a computer system capable of automatically annotating literary texts according to their structural components. By aligning human expertise with computational methods, this research aims to reshape our understanding and engagement with literature in the digital age, offering new insights and opportunities for exploration in the world of literary analysis. The system shows promising results by reflecting similar results for the automatically identified structure and the human-annotated structure.

In this specific research, a three-act structure shows a higher agreement between annotators as well as between human annotators and the system. This can, in part, be because fewer transitions had to be assigned, and therefore, there was less chance for a mistake. But we also compare the human and system annotations to a baseline which did not show a significant difference, meaning that both human annotations and the system do not show any bias towards either structure theory. From this, we can conclude that the three-act structure is indeed a better fit for the specific texts used. On a more general level, the evaluation showed that the system has a high agreement with human judgement and can be applied to a wider number of structure theories and texts.

The automation of literary annotation is still a relatively new area of research, but it has the potential to transform the way we study and engage with literature. It can enable us to conduct large-scale analyses of literary texts that would be impossible or impractical to perform manually. It can also help us to identify patterns and trends

that may be difficult to discern with the naked eye. The work presented in this article provides a valuable foundation for future research in this area but still poses important questions for future work.

References

F. Barth. Annotation guidelines for narrative levels and narrative acts v2. *Journal of Cultural Analytics*, 6(4):98–139, 2021. doi: 10.22148/001c.30701.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

J. Campbell. *The Hero with a Thousand Faces*. Bollingen series. New World Library, 2008. ISBN 9781577315933. URL `https://books.google.co.za/books?id=I1uFuXlvFgMC`.

Maximilian Droog Hayes, Geraint Wiggins, and Matthew Purver. Automatic detection of narrative structure for high-level story representation. In *Conference: 5th AISB Symposium on Computational Creativity*, April 2018.

M. Finlayson. Inferring Propp's functions from semantically annotated text. *The Journal of American Folklore*, 129:55, 2016. doi: 10.5406/jamerfolk.129.511.0055.

G. Freytag. *Freytag's Technique of the drama: an exposition of dramatic composition and art. An authorized translation from the 6th German ed. by Elias J. MacEwan*. Scott, Foresman, Chicago, 1894.

G. Genette, J. E. Lewin, and J. D. Culler. Narrative discourse: an essay in method. *Comparative Literature*, 32:413, 1980.

E. Gius, F. Jannidis, M. Krug, A. Zehe, A. Hotho, F. Puppe, J. Krebs, N. Reiter, N. Wiedmer, and L. Konle. Detection of scenes in fiction. In *Book of Abstracts of the Digital Humanities conference*, Utrecht, Netherlands, July 2019.

Evelyn Gius, Carla Sökefeld, Lea Dümpelmann, Lucas Kaufmann, Annekea Schreiber, Svenja Guhr, Nathalie Wiedmer, and Fotis Jannidis. Guidelines for detection of scenes, January 2021.

N. Heyns and M. van Zaanen. Finding topic boundaries in literary text. In *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*, 2021.

N. Heyns and M. van Zaanen. Detecting multiple transitions in literary text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3375–3381, 2022.

Manfred Jahn. Narrative levels. In *Narratology: A Guide to the Theory of Narrative*, chapter N2.4. English Department, University of Cologne, 2005.

Matthew L. Jockers. Syuzhet: Extract sentiment and plot arcs from text, 2015. Available at: `https://github.com/mjockers/syuzhet`. Accessed 30 January 2023.

N. Ketschik, B. Krautter, S. Murr, and Y. Zimmermann. Annotation guideline no. 4: Annotating narrative levels in literature. *Journal of Cultural Analytics*, 4(3), 2019. doi: 10.22148/16.055.

W. Labov. *The Language of Life and Death: The transformation of experience in oral narrative*. Cambridge University Press, Cambridge, U.K., 2013.

W. Labov and J. Waletzky. Narrative analysis: oral versions of personal experience. In *Essays on the Verbal and Visual arts, Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society*, pages 12–44, Seattle, W.A., 1967. University of Washington Press.

Jessica Ouyang and Kathy McKeown. Towards automatic detection of narrative structure. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4624–4631. European Language Resources Association (ELRA), 2014.

Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.26. URL https://aclanthology.org/2021.emnlp-main.26.

Gerald Prince. Narrative analysis and narratology. *New Literary History*, 13(2):179–188, 1982.

V. Propp. *Morphology of the Folktale*. Texas Press, 1968.

N. Reiter, M. Willand, and E. Gius. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. *Journal of Cultural Analytics*, 4(3), 2019. doi: 10.22148/16.048.

Nils Reiter. Towards annotating narrative segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL*, pages 34–38, Beijing, China, 2015. The Association for Computer Linguistics.

T. Steele. The structure of the detective story: Classical or modern? *Modern Fiction Studies Lafayette, Ind*, 1981.

Tzvetan Todorov. The 2 principles of narrative. *Diacritics*, 1(1):37, 1971. doi: 10.2307/464558.

Inna Uglanova and Evelyn Gius. The order of things. a study on topic modelling of literary texts. In *Workshop on Computational Humanities Research*, 2020. URL https://api.semanticscholar.org/CorpusID:227913848.

Joris van Zundert, Marijn Koolen, Julia Neugarten, Peter Boot, Willem Robert van Hage, and Ole Mussmann. What do we talk about when we talk about topic? In *Workshop on Computational Humanities Research*, 2022. URL https://api.semanticscholar.org/CorpusID:254045843.

P. Wake. Narrative and narratology. *The Routledge Companion to Critical and Cultural Theory*, pages 23–36, 2006.

Dan Wells. Dan Wells on story structure. https://www.youtube.com/playlist?list=PLC430F6A783A88697, February 2010.

M. Wiren and A. Ek. Annotation guideline no. 7 (revised): Guidelines for annotation of narrative structure. *Journal of Cultural Analytics*, 6(4):164–186, 2021. doi: 10.22148/001c.30703.