



Computational analysis of narrative structures

A. Heyns



orcid.org/0000-0002-0802-5005

Thesis accepted in fulfilment of the requirements for the degree *Doctor of Philosophy in Languages and Literature with Linguistics and Literary Theory* at the North-West University

Promoter: Prof. M. van Zaanen

Graduation: May, 2026

The bottom of the cover features a blue wave pattern, mirroring the purple wave pattern at the top.

Acknowledgements

I am deeply grateful to my husband, Wicus for your unwavering patience, steadfast support and boundless love throughout the long and demanding phases of this PhD. Thank you for making even the smallest moments joyful and full of light.

To my parents, Michael and Marita, thank you for all the advice. Thank you for your encouragement and belief in me from the very beginning and for the inspiring example you have set throughout my life.

I extend my sincere gratitude to my supervisor, Menno van Zaanen. Thank you for the insightful guidance and constructive feedback that shaped this work into its final form.

I am also thankful to Inge van de Ven. Thank you for the help during the early stages of this project. I am especially gratefull for the way you introduced me to your network, welcomed me into your office, and offered warm hospitality.

Abstract

This thesis explores the computational analysis of high-level narrative structures in literary texts, bridging narratological theory with natural language processing techniques. Narratives are examined at two core levels: the story (*histoire*), encompassing temporal-causal event sequences and narrative arcs, and the narrative discourse (*récit*), involving elements such as scenes, levels, anachronisms and focalisation. The research addresses key challenges in computational narratology, including the scarcity of standardised annotated data sets and fragmented computational approaches that hinder large-scale analysis.

A literature review maps existing terminology, methodologies, and data sets, revealing inconsistencies in definitions for terms like “plot”, “scene”, and “narrative level”. To identify narrative arcs, two systems using latent dirichlet allocation (LDA) are developed and evaluated against human annotations on whodunit short stories, demonstrating feasible detection of single and multiple transitions.

For narrative discourse, modular annotation guidelines are proposed and refined by addressing challenges in the annotation process. These guidelines enable the creation of a unified data set combining existing annotations with guidelines specifically designed for the mystery genre data. Fine-tuned RoBERTa (Robustly optimised BERT pretraining approach) models, trained on this data set, achieve promising results for automated discourse annotation, with the combined model yielding an average F1 score of 68.60 and high ROC-AUC (94.01), outperforming existing baselines.

Other contributions include novel arc detection systems, standardised guidelines, and a publicly available data set advancing reproducible computational literary studies. Limitations, such as data set size and genre-specific focus, suggest future expansions to diverse genres and advanced architectures for enhanced scalability and interpretability.

Keywords: computational narratology; narrative arcs; narrative discourse; annotation guidelines; narrative structure.

Preface

This thesis is based on the following articles of which I am the principal author:

- Heyns, N. (2024). Post-digital narrative analysis. In *The Routledge Handbook of AI and Literature*. Routledge.
- Heyns, N., & van Zaanen, M. (2021). Finding topic boundaries in literary text. *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*. https://dh2021.digitalhumanities.org.za/wp-content/uploads/2021/11/DHASA_book_of_abstracts.pdf
- Heyns, N., & van Zaanen, M. (2022). Detecting multiple transitions in literary text. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3375–3381. <https://aclanthology.org/2022.lrec-1.360.pdf>
- Heyns, N., & van Zaanen, M. (2024). Automatic Narrative Structure Identification in Literary Texts. *Digital Humanities Benelux Journal*, 6, 71–90. https://journal.dhbenelux.org/wp-content/uploads/2024/11/5_Heyns_individual.pdf
- Heyns, N., & van Zaanen, M. Refining narrative annotation for segmentation: Challenges and debates. In: *Comparative Literature Goes Digital DH2025 Mini-Conference*. 2025.
- Heyns, N., & Van Zaanen, M. (2025). Annotated mystery narratives data set. *Journal of Open Humanities Data*.
- Heyns, N., & van Zaanen, M. (n.d.). *Automatic narrative annotation* [Submitted to *Computational Humanities Research* (under review)].

All co-authors of the publications listed above have given their written permission for the articles to be included in this PhD thesis and for the thesis to be submitted for the degree of Doctor of Philosophy.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Problem statement	3
1.3	Project scope	3
1.4	Research purpose and questions	4
1.5	Objectives of the study	5
1.6	Contribution of the study	6
1.7	Research philosophy and design	6
1.8	Chapter layout	7
2	Literature Review	9
2.1	Introduction	9
2.2	Background	10
2.2.1	Structure	10
2.2.2	Plot	11
2.2.3	Narrative level	12
2.2.4	Event and scene	12
2.2.5	Social networks	12
2.3	Methodology	13
2.3.1	Source selection	13

2.3.2	Analysis	14
2.4	Results	14
2.4.1	Terminology	16
2.4.2	Annotation strategies	20
2.4.3	Data sets	23
2.4.4	Computational approaches	24
2.4.5	Applications	28
2.4.6	Term correlations	29
2.5	Discussion	31
2.6	Conclusion	32
I	Story	35
3	Automatic Narrative Arc Identification	37
3.1	Introduction	38
3.2	Background	40
3.2.1	Related work	40
3.2.2	Length of LDA documents	41
3.2.3	Evaluation of LDA	42
3.3	Methodology	43
3.3.1	System 1: Single boundaries	43
3.3.2	System 2: Multiple boundaries	48
3.4	Discussion	55
3.5	Conclusion	57
3.6	Future work	58
4	Comparing Arc Systems with Human Judgement	59

4.1	Introduction	59
4.2	Background	61
4.2.1	Structure theories	62
4.2.2	Previous and related work	64
4.3	Methodology	65
4.3.1	Data	65
4.3.2	System	70
4.4	Evaluation	73
4.4.1	Inter-rater annotator agreement	73
4.4.2	Human annotator versus computer annotator	74
4.4.3	Comparing structure theories	75
4.5	Discussion	77
4.6	Future work	77
4.7	Conclusion	78
II	Narrative Discourse	80
5	Narrative Discourse Annotation Guidelines	82
5.1	Introduction	82
5.2	Background	84
5.2.1	Narrative	84
5.2.2	Narrative levels	84
5.2.3	Anachronisms	85
5.2.4	Scene, non-scene and ellipsis	85
5.2.5	Narrative perspective	86
5.3	Existing data sets	87
5.4	Annotation guidelines	87

5.4.1	Narrative levels	88
5.4.2	Anachronisms	89
5.4.3	Scene, non-scene, and ellipsis	89
5.4.4	Narrative perspective	91
5.4.5	Punctuation and paratext	91
5.4.6	Genre-specific annotations	92
5.5	Annotation scheme	92
5.6	Data set	93
5.7	Annotation process	94
5.8	Discussion	94
5.9	Conclusion	94
6	Problems During Narrative Discourse Annotation	96
6.1	Introduction	96
6.2	Annotation challenges	97
6.2.1	Container principle	97
6.2.2	Scene and non-scene	98
6.2.3	Focalisation	99
6.3	Misleading character description	101
6.4	Annotation method	101
6.5	Discussion	101
6.6	Conclusion	102
7	Annotated Narrative Discourse Data	104
7.1	Introduction	104
7.2	Method	105
7.2.1	Steps	105

7.2.2	Sampling strategy	107
7.2.3	Quality control	109
7.3	Reuse potential	109
7.4	Conclusion	110
8	Automatic Narrative Discourse Annotation Systems	112
8.1	Introduction	112
8.2	Background	113
8.2.1	Existing annotated data	113
8.2.2	Existing methods	114
8.3	Methodology	115
8.3.1	Data sets	116
8.3.2	Model development	118
8.3.3	Baseline model	119
8.4	Implementation	120
8.5	Evaluation	121
8.6	Discussion	125
8.7	Future work	126
8.8	Conclusion	127
9	Discussion	129
9.1	Introduction	129
9.2	Objectives and key findings	129
9.3	Limitations	133
9.4	Contributions	133
9.5	Future work	133
10	Conclusion	135

10.1 Problem statement	135
10.2 Scope	136
10.3 Research purpose and questions	136
10.4 Objectives of the study	137
10.5 Contribution of the study	138
10.5.1 Practical contributions	138
10.5.2 Theoretical contributions	139
10.6 Conclusion	140
10.7 Ethics clearance	140
References	141
A Final tagset for narrative discourse	154
B Distribution of tags in narrative discourse annotated text	156
C Ollama prompt template	158

List of Figures

- 2.1 Overview of the publication selection process. 15
- 2.2 Correlogram showing how labels correlate to each other. 30
- 3.1 The relationship between the number of LDA topics and the number of possible transitions proposed by the system. 49

List of Tables

3.1	Sample from each of the two source texts.	45
3.2	RMSE results of each system for the range of LDA topics.	47
3.3	LDA topics with corresponding snippet numbers.	50
3.4	The minimum distance for the start and end snippets.	50
3.5	Lists <i>A</i> and <i>B</i> after the snippets of LDA topic two are added.	50
3.6	The final versions of list <i>A</i> and list <i>B</i>	51
3.7	RMSE results of the RSTI and MTI systems.	54
4.1	Information of the short stories used.	69
4.2	Example annotation.	71
4.3	The average RMSE between annotators.	74
4.4	The average RMSE between the baseline and the annotators.	74
4.5	RMSE per annotator compared to the random baseline.	75
4.6	The average RMSE per structure theory.	75
4.7	The average RMSE per section.	76
4.8	The RMSE between the system and a random baseline.	76
5.1	Properties of the annotated texts.	93
5.2	Comparison with the SANTA data set ranked by total word length.	93
7.1	Information on the narratives in the data set.	106
7.2	Information on the tags assigned in the data set.	107

8.1	Overview of the consolidated tags.	117
8.2	Best metrics per fold for the SANTA model.	122
8.3	Best metrics per fold for the mystery model.	123
8.4	Best metrics per fold for the combined model.	124
8.5	Average performance metrics for each model and the baseline.	124
A.1	Tags used during the annotation process.	154

Table of Abbreviations

AMR	Abstract meaning representation
BERT	Bidirectional encoder representations from transformers
CLA	Computational literary analysis
CLS	Computational literary studies
CR	Co-reference resolution
IAA	Inter-annotator agreement
LDA	Latent dirichlet allocation
LLM	Large language model
MRE	Most reportable event
MTI	Multiple transition identification
NER	Named entity recognition
NLP	Natural language processing
RoBERTa	Robustly optimised BERT pretraining approach
RSTI	Recursive single transition identification
SANTA	Systematic analysis of narrative texts through annotation
SDS	Scene description segment
STSS	Shared task on scene segmentation
STI	Single transition identification
STS	Scene transition segment

Chapter 1

Introduction

1.1 Introduction

In recent years, the rapid advancement of large language models (LLMs) has transformed how we interact with and understand textual information. From generating coherent summaries to answering complex queries, these models have become essential tools across domains. Yet, while LLMs excel at processing and summarising texts, or even generating creative outputs, their ability to deeply engage with the intricate structures of literary narratives, such as the unfolding of events, character dynamics or scene segmentation, remains under-explored. This is partly because LLMs operate by predicting the most likely next word or token based on patterns in massive training data, which allows them to mimic language fluently but often results in shallow or inconsistent handling of complex narrative elements without additional adaptation. Understanding the structure of literary narratives requires moving beyond surface-level linguistic analysis to examine how stories are crafted at a higher level. Narratives are not merely sequences of words or sentences, but complex systems of events, temporal shifts and character interactions that create meaning communication. To achieve high accuracy and consistency in modelling complex narrative structures such as plot arcs or temporal shifts, LLMs typically require fine-tuning on specialised, annotated data sets that capture precise narrative features. However, effective fine-tuning of models like LLMs demands a deep, domain-specific understanding of what constitutes narrative structure, so that annotated data can be created accurately to train them for tasks like identifying plot arcs or event timelines. This thesis seeks to address this challenge by using natural language processing (NLP) to explore the structural anatomy of narratives, offering new insights into how narratives are organised.

A narrative not only refers to the narrative in a story, 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 these notions explain the way the world is experienced through an individual's narrative thereof. However, this definition of the

narrative does 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, p. 179) claims that we have “certain intuitions (or 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, p. 179). This temporal-causal ordering defines the story (*histoire*). However, the narrative discourse (*récit*), the actual text, presents this story through selective, transformative mechanisms. Genette et al. (1980) rigorously separate these levels: the story is the reconstructed chronological content; the discourse is the signifying act that orders, paces, focalises and embeds it. Narratology is the systematic study of this interplay between story and discourse across forms and media. Within narratology, narrative structure theory narrows the focus to the formal architecture of sequence, segmentation and hierarchy, particularly how discourse signals and shapes macro-level story patterns such as thematic arcs.

Computational narratology is the application of computational methods to study narrative structures, elements and patterns in text. The field draws on narratological theories (for example, scholars like Genette et al., 1980, or Propp, 1968) and adapts them for computational analysis. Research on computational narratology can be broadly categorised into two levels of analysis: low-level linguistic analysis and higher-level narrative structure analysis.

Low-level linguistic analysis uses well-established NLP techniques to analyse text at the word or sentence level. Examples are sentiment analysis to track emotional arcs, topic modelling to extract themes or motifs, named entity recognition (NER) to identify characters or places, or syntactic parsing to study stylistic features.

Higher-level narrative structure analyses draws on narratological concepts like story versus discourse (Genette et al., 1980) and these concepts are computationally challenging because they require abstracting beyond surface-level text to interpret narrative logic. Examples of higher-level structure analyses include methods to model plot structures or event sequences, including how events unfold over time, algorithms that divide text into scenes or narrative units, and analysing character interactions and their roles in plot progression. Many of these high-level tasks use the better-researched, low-level linguistic analyses as stepping stones. Even though there is a growing body of research on higher-level narrative structures, the research is often more fragmented, domain-specific or limited by the complexity and inconsistency of formalising narrative concepts computationally, which in turn limits the annotation of data sets.

The motivation for computationally analysing literary texts on such a massive scale stems from the limitations of traditional manual methods, which are time-intensive, highly subjective and often restricted to small samples of works. By applying computational tools to vast corpora spanning thousands or even millions of texts, we can uncover patterns and insights that would otherwise remain hidden, like how storytelling styles change over time or reflect cultural differences and that deepen our understanding of literature. This

approach also makes literary analysis more accessible, enabling anyone to explore vast collections without years of manual study.

Moreover, computational analysis supports practical applications, like improving AI-generated stories, creating smarter educational tools, and helping creators in fields like film or gaming to craft better narratives. Unlike LLMs, which excel at generating text or producing content-focused summaries, an understanding of narrative structure enables structurally informed summarisation or narrative creation that captures plot dynamics and character arcs.

These capabilities not only enrich digital humanities research but also bridge traditional narratology with computational innovation. As a result, the field of computational literary studies has seen a growing interest in computational approaches aimed at studying literary narrative texts, opening avenues for interdisciplinary advancements.

This thesis develops computational models for higher-level narrative analysis by jointly modelling story (the what and in what order, including event sequences and story arcs) and discourse (the how of telling, encompassing levels, focalisation, anachronisms, and temporal pacing) in literary texts following Genette et al. (1980). In addition to developing computational models, this thesis also standardises annotation guidelines and combines various data sets in an effort to advance computational narratology, providing consistent and reproducible methods to study stories across varied texts and genres.

1.2 Problem statement

The computational analysis of literary narrative structures, specifically high-level narrative structures such as plot, scene, or temporal event sequences, is limited by a scarcity of standardised annotated data sets and fragmented machine-learning approaches. Existing data sets vary widely in annotation standards due to differing research goals, hindering comparative analyses. Similarly, techniques for detecting narrative structures are often incompatible, lacking a unified framework. These gaps restrict large-scale study of narrative structures, impeding advancements in computational narratology.

1.3 Project scope

Given the problem statement above, this thesis restricts the initial scope of the research to identifying narrative structure, with an emphasis on developing tools and resources to identify and annotate high-level structural elements.

The scarcity of annotated data sets is addressed by conducting a thorough survey of existing data sets and their purpose, and a standardised framework is proposed for

annotating collections. This thesis aims to combine existing data sets into one data set that will follow the standardised framework as well as add additional data to the data set.

The scope of this thesis is limited to narratives from classic literature, particularly focusing on the mystery genre due to its clear and formulaic narrative structure. Classic whodunit mysteries from the Golden Age (1920–1950), such as works by Agatha Christie, provide a good baseline. This thesis will not cover all genres or all periods of literature. Where necessary, additional features will be annotated to create a data set accommodating a range of narrative elements relevant to computational analysis. This data set will be in the public domain and can then be used in future studies.

This thesis recognises that literary texts vary widely in their nuances, such as genre-specific features or unique narrative elements, and researchers may wish to investigate these aspects in targeted ways. To accommodate this, while facilitating data sharing and combination for larger data sets, modular, standardised annotation guidelines are proposed that allow tags to be easily added or removed as needed.

This thesis aims to tackle the limitation of competing techniques by investigating what machine-learning techniques have been used to automatically detect high-level narrative structure in literary texts. The overlapping similarities are identified in an effort to consolidate these techniques. Based on these results, this thesis will then investigate how these or additional techniques can be used or improved upon.

This thesis will not focus on low-level linguistic features (syntax, coreference, entity recognition, sentiment, etc.) unless they directly serve a high-level narrative element. Theoretical or interpretative literary analysis also fall outside of the scope of this thesis.

1.4 Research purpose and questions

This research addresses the identified problem by proposing a standardised, modular annotation scheme and data set as well as a computational framework that will enable large-scale, empirical investigations into narrative patterns and support advancements in computational literary studies. The primary research questions guiding this thesis are:

RQ 1 *What computational approaches have been developed for analysing high-level narrative structures in literary texts?*

This question is answered in Chapter 2.

RQ 2 *Which under-explored aspects of high-level narrative structure are both foundational to overall story organisation and amenable to automated detection?*

This question is answered in Chapter 2.

RQ 3 *What are the primary challenges in annotating high-level narrative elements, across diverse literary texts?*

This is addressed in Chapter 2 and Chapter 6.

RQ 4 *Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?*

This question is answered in Chapter 4, Chapter 5, and Chapter 6.

RQ 5 *Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?*

This question is answered in Chapter 4 and Chapter 5.

RQ 6 *Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?*

This is addressed in Chapter 3, Chapter 4, and Chapter 8.

RQ 7 *How effectively can existing and enhanced machine-learning methods, applied to an annotated data set, identify and analyse high-level narrative structures?*

We intent to answer this question in Chapter 8

By answering these research questions, this thesis develops standardised, modular annotation guidelines and a consolidated data set, while evaluating and enhancing machine-learning methods. This addresses the scarcity of standardised data sets and fragmented approaches that limit the computational analysis of literary narrative structures.

1.5 Objectives of the study

The objectives of the study are the following, considering the research questions:

- Investigate narratological concepts and existing computational approaches.
- Identify computationally under-explored aspects of high-level narrative structure identification.
- Create a standardised annotation format.
- Create an annotated data set to support model training and evaluation, with a focus on integrating existing data sets into a publicly accessible data set.

- Apply computational techniques (for example, fine-tuned models) to automate the identification of narrative structure.

1.6 Contribution of the study

The study makes both practical and theoretical contributions.

- Practical contributions
 - Standardisation of narrative level annotation schemes specifically with regard to high-level narrative structure.
 - Creation and public release of an annotated literature data set incorporating high-level narrative structure information (identified in Part II), clearly labelled and contextualised within narratives.
 - The development of new tools, for example by applying topic modelling and machine-learning techniques to automate the identification of structural components of narratives, that can be used by the literary research community to aid them in their research in the field of narratology.
- Theoretical contribution
 - This study refines the theoretical understanding of high-level narrative structures by formalising narratological concepts for computational analysis, providing a clearer framework for how these structures can be systematically represented and studied across diverse literary texts.
 - Ultimately, this study combines human understanding of narratives with computational tools to uncover new insights about how narratives are structured and how computational methods can better analyse literature, while creating a clear, reliable method for future research in computational narratology.

1.7 Research philosophy and design

This thesis adopts an empirical epistemological stance that researches quantifiable phenomena in narrative structures through computational methods. Computational tools are used to test hypotheses derived from narratological theory (e.g., Genette et al., 1980) in a data-driven manner. While interpretive elements inform annotation guidelines (e.g., subjective judgements on scene boundaries), the core focus is on reproducible, empirical validation using computational models and annotations.

The high-level methodological framework employs a mixed-methods design:

1. Qualitative synthesis of narratological literature and annotation processes to develop guidelines and datasets in Chapters 2, 5, 6, and 7;
2. Quantitative development and evaluation of computational systems in Chapters 3, 4, and 8.

1.8 Chapter layout

The chapters in this thesis can be outlined as follows:

Chapter 1: Introduction This chapter introduces the research problem and the following chapters are presented.

Chapter 2: Literature review This chapter provides an overview of existing work in computational literary studies relating to narratology. Key sub-fields within computational narratology are identified, available annotated data sets are surveyed, and existing methods for high-level narrative structure analysis are investigated, establishing the foundation for addressing the identified research gaps.

Part I: Story In Part I, the focus is on the story level (*histoire*), particularly the identification of narrative arcs (this concept is defined in Chapter 2) to explore how computational methods can detect and describe these high-level patterns in texts. This approach lays the foundation for understanding the temporal-causal backbone of narratives before integrating finer-grained discourse elements.

Narrative arc identification is a central yet under-explored area in computational narratology. This part develops and evaluates LDA-based systems for detecting both single and multiple arc boundaries, benchmarked against human annotations, to establish reliable, reproducible methods for large-scale arc analysis across literary corpora.

Chapter 3: Automatic narrative arc identification Different systems to automatically identify narrative arcs are studied and implemented, addressing an under-researched area identified in Chapter 2.

Chapter 4: Comparing arc systems with human judgement To compare the arc identification systems, Chapter 4 compares the system with human judgement after guidelines were written and the text was annotated. The arc identification systems developed in Chapter 3 are then compared with the human annotated data from this chapter.

Part II: Narrative discourse In Part II, the focus shifts to narrative discourse, with an emphasis on narrative levels and scenes (these concepts are defined in Chapter 2) to explore how they contribute to and are organised within these larger arcs. This methodology ensures a thorough understanding of how stories are constructed at different levels.

Scene and narrative level identification represents the most developed areas of high-level narrative structure research, yet existing methods and annotations lack standardisation. This part investigates and addresses these standardisation gaps to enable consistent computational analysis across diverse literary texts.

Chapter 5: Narrative discourse annotation guidelines Annotation guidelines are researched and developed, and used to identify narrative discourse features like scenes, narrative levels and temporal aspects for Chapter 7 and Chapter 8.

Chapter 6: Problems during narrative discourse annotation Problems encountered while annotating the narrative discourse features in Chapter 5 are identified and discussed in this chapter.

Chapter 7: Annotated narrative discourse data This chapter presents the final data set that is used to train automatic narrative annotation systems in Chapter 8.

Chapter 8: Automatic narrative discourse annotation systems A system to automatically identify narrative discourse is developed and the output of the system is compared to human annotations.

Chapter 9: Discussion The key findings and contributions of the entire study are highlighted and discussed.

Chapter 10: Conclusion This chapter provides an overall conclusion of the research, including a short summary of each chapter. Limitations and future work are discussed.

Chapter 2

Literature Review

*This chapter presents a literature review on the computational analysis of narrative structure in literary texts. The work forms the basis of this thesis, establishing the theoretical and methodological context for the subsequent chapters, each of which builds on specific aspects of narrative analysis introduced here. This chapter aims to answer **RQ 1**: “What computational approaches have been developed for analysing high-level narrative structures in literary texts?”, **RQ 2**: “Which under-explored aspects of high-level narrative structure are both foundational to overall story organisation and amenable to automated detection?” and **RQ 3**: “What are the primary challenges in annotating high-level narrative elements, across diverse literary texts?”*

2.1 Introduction

Computational approaches to narrative structure often draw on narratological theory, but this work regularly seems to take place independently of established theoretical foundations. Theoretical concepts are also often defined in abstract ways and lack the precise definitions essential for computational research. Consequently, many computational researchers define those concepts independently, leading to multiple overlapping definitions of these concepts.

Currently, there is no systematic overview of how narrative structure is computationally modelled in literary texts, particularly with respect to which theoretical concepts are used, how narrative structure is annotated, and which computational methods are applied. This lack of standardisation may affect both annotation consistency and the comparability of results across studies.

To address this gap, this chapter provides a scoping review of computational approaches to narrative structure in literary texts, focusing on high-level narrative structures. This review focuses on work centred around the automatic identification of narrative structures,

which is strongly related to temporal attributes in terms of order, duration and frequency and the role of social agents.

Given the growing interest in computational literary analysis of narrative structure, a solid overview of terminology, techniques and annotation strategies is needed. Standardised terminology enables researchers to compare work. By collating research highlights, we can identify active areas and shortcomings in the current literature, which may lead to new opportunities.

To summarise, this literature review (i) examines terminology for automatic narrative structure analysis, (ii) groups overlapping terms, (iii) identifies concepts with existing methodologies, and (iv) the different methodologies being utilised. In particular, the focus falls on (v) work on narrative levels and time attributes that distinguish these levels, and the identification of social agent networks occurring in narratives. In addition, an investigation into the (vi) annotation methods and (vii) availability of data sets is done.

2.2 Background

Before looking at the relevant computationally oriented literature, this review first describes relevant terminology from a theoretical perspective. This helps to identify the scope of the review and to structure the description of the findings as well.

2.2.1 Structure

In general, “structure” refers to the organised arrangement of elements. Linguistically, this includes morphological (within words), syntactic (within sentences) and discourse structure (within texts or conversations). Structure in a literary context refers to the organisation and arrangement of elements within literary works, which gives shape to the narratives. This influences the way the story is presented and how its components relate to each other.

Within narratives, we can identify several types of elements and relations. For instance, character networks depict social interactions, while event structures map temporal relations. Events are placed on a timeline, illustrating time representation through techniques like foreshadowing and flashbacks.

“Character networks” or “social networks” are typically used to represent the relations between characters, whereas the terms “discourse structure”, “temporal discourse structure” and “temporal dependency tree structure” are associated with the study of how information is organised within the text, emphasising the sequencing of events or information over time. In this review, these terms are referred to as “structure”.

Narrative (temporal aspects)

Prince (1982) defines a narrative as the “representation of real or fictive situations and events in a time sequence”. Situations provide the setting and background as descriptions of the geographic location and time, whereas events in narratives are placed at particular moments in time. This definition highlights the important role of time in narratives.

To analyse a narrative according to time, three questions arise: “When?”, “How long?” and “How often?”. From these questions, Genette et al. (1980) discern three narrative aspects related to time.

Order refers to the sequence of events in relation to the order of narration. Genette et al. (1980) distinguish between *histoire* (story) and *récit* (narrative discourse). Where story represents the causal-chronological order of events, while narrative discourse refers to the order of the events as narrated. This is also called logical order, determined by the cause-and-effect relationships between events, which can be non-linear.

Discordance between “story time” and “discourse time”, known as *anachrony*, occurs when events are narrated out of their chronological sequence, as in *prolepsis* (flashforward: future event told earlier) or *analepsis* (flashback: past event told later) (Genette et al., 1980). Additionally, there might be a shift between one narrative level and another (*metalepsis*) due to a change in various narrative elements, such as plot lines, perspectives or layers of reality.

Duration refers to the time covered by the events narrated in the text. Genette et al. (1980) distinguish between *scene*, where the duration of the text is equal to the time it took to narrate the events, *summary*, which is a compressed presentation of events where a longer duration is conveyed in a shorter narrative space, and *ellipsis*, which contains a description, reflection or commentary.

Frequency denotes how often an event is narrated within the text. Genette et al. (1980) identify *singulative*, when an event is only narrated once, and *iterative*, where the same event is described multiple times within the same narrative.

2.2.2 Plot

Baroni and Revaz (2016) trace the concept of *plot* to Aristotle’s *muthos*, encompassing both sequence and a more causal relationship between events (Forster, 1927) as cited by Baroni and Revaz (2016). Plot also functions as a rhetorical device to build anticipation for the ending (Baroni, 2007; Brooks, 1992; Dannenberg, 2008) as cited by Baroni and Revaz (2016). However, Baroni and Revaz (2016) refer to the “slippery notion of plot” based on Abbott (2007), who criticises the term plot for being too vague, leading narratologists to avoid the term altogether. Barth and Dönicke (2021) argue that plot is in some way

connected with the sequential nature of narratives and therefore there is a need for clarification between the more precise concept of narrative sequence and the broader notion of plot.

In this thesis, *plot* refers to the *story-level causal arc*, detectable via thematic shifts (Part I) and composed of discourse-level units like scenes (Part II).

2.2.3 Narrative level

Along with *histoire* and *récit*, Genette et al. (1980) also introduce the narrator's role in telling and ordering the story through plot. This classification of the narrator is referred to as narrative levels that, according to Genette et al. (1980), can be *extradiegetic* (external to the story world), *intradiegetic* (internal to the story world), and *metadiegetic* (in other words, a story within a story). Thus Genette et al. (1980) describe narrative levels as the relations between embedded narrative and the diegesis.

2.2.4 Event and scene

Events are the occurrences that drive the story forward, forming the building blocks of plot. For example, a character unknowingly drinks poisoned coffee, followed by the discovery of a note revealing the danger. These individual events, linked by a common character, location and time frame, form a scene. The plot is the overarching story arc built from a sequence of such connected scenes, each driving the narrative forward and revealing the cause-and-effect relationships between events (Forster, 1927).

Not all events are created equal. Mani (2005) notes that some sequences influence the narrative's flow, forming the core plot, while others, though memorable, do not affect progression. Consequently, understanding a narrative extends beyond identifying the chronological arrangement of events; it necessitates the ability to discern familiar patterns within that sequence and recognise which scenes drive the story forward.

2.2.5 Social networks

As mentioned before, narratives describe sequences of events in time. Often, social agents, in other words, (potentially human) characters, are involved with these events. The characters may be related or interact in various ways. To describe the relationships between characters, they can be placed in a social network. Such social networks are often depicted using characters as the nodes in the network, while the edges (in other words, the links) describe the relationships.

Note that the social network structure is typically different from the temporal structure described earlier. However, there may also be interactions between time in the narrative

and the social network as the nature of relationships between characters may change over time.

2.3 Methodology

In order to provide a structured approach to this scoping review, this chapter follows the *Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR)* process (Tricco et al., 2018).

The review starts by describing how relevant publications are selected and how relevant information is extracted from these publications. This is followed by an analysis of the terminology used in the publications to describe specific concepts associated with automatic narrative structure analysis. This highlights key themes and their terminology. Then the different narrative annotation methods are investigated. Based on the annotation methods, data sets have been created which will also be described. Lastly, the focus falls on the (computational) methodologies used. Common approaches, techniques and tools used for computational narrative analysis are then identified.

2.3.1 Source selection

The review started with a structured search for relevant publications using Publish or Perish¹, which retrieves academic citations. Publish and Perish enables database-specific searches (for example, ISI/Web of Science, Scopus) and provides detailed publication lists. The search is done using Google Scholar² as a comprehensive digital database to ensure the search covers the relevant literature and journals. Although it is difficult to find all published sources, Google Scholar houses a large amount. To our knowledge, there is no more complete database of publications related to computational narrative.

Publications are collected using the queries *computational “narrative time”*, *computational “narrative scene”*, and *computational “narrative segment”*. Quotation marks ensure phrase-based search rather than individual word matches, preventing unrelated results, as the words “time”, “segment”, “scene”, and “narrative” also occur in many other non-narrative related contexts.

This review focuses on academic publications, excluding books (due to limited access), abstract-only proceedings (such as conference programmes or presentation lists without full papers), article preprints (as they are not peer-reviewed), and retracted publications (for obvious reasons). Due to personal limitations, only English publications are selected.

Since terminology used in the field of computational narratology is also used in other domains, this review excludes publications in non-related fields. Similarly, only publications

¹<https://harzing.com/resources/publish-or-perish>

²<https://scholar.google.com>

that focus on literary narratives are selected. This means that publications on related fields like film, games, news, government and social media texts are excluded. Although this type of work is related, the attributes of the narratives require different computational techniques to analyse. Finally, studies focusing on literary attributes like gender also fall outside of the scope of this study, as they do not deal with structure.

After searching and downloading the publications, the relevant publications are reviewed manually according to the inclusion and exclusion criteria. A spreadsheet tracks publications, noting inclusion/exclusion reasons.

2.3.2 Analysis

The content of all relevant publications is manually analysed by the author. This analysis is performed by taking a hypothesis-free approach to close reading. The publications are labelled in an open-ended way (without pre-defining the labels beforehand). The reason for this is that the labels cannot be elaborated on before having read the publications. While reading the publications, labels are added, and if needed, new labels are also added to publications already read. Two rounds of annotation have been performed.

2.4 Results

The initial search resulted in a total of 1 152 unique publications. After excluding the non-eligible publication types and non-English publications, 921 publications remained. Removing publications from non-relevant fields resulted in 446 remaining publications. After excluding publications on non-literary text, 285 publications remained. Out of these, 208 are about structure in literary text and from these, 90 are about computational narrative structure, which are used for further analysis. A more detailed overview can be found in Figure 2.1.

To investigate the content of the selected publications, this review focuses on a number of areas. First, the terminology used in the different publications identifies key themes, concepts and topics relevant to the field. Second, approaches to annotating the narrative structures in texts are investigated. There may be different ways to annotate texts, which may lead to different tasks or varying levels of granularity. Third, the available data sets that contain structured texts are examined. Fourth, the different computational approaches that people have used to automatically structure the narratives are described. Finally, correlations between the use of terminology in the different publications are provided.

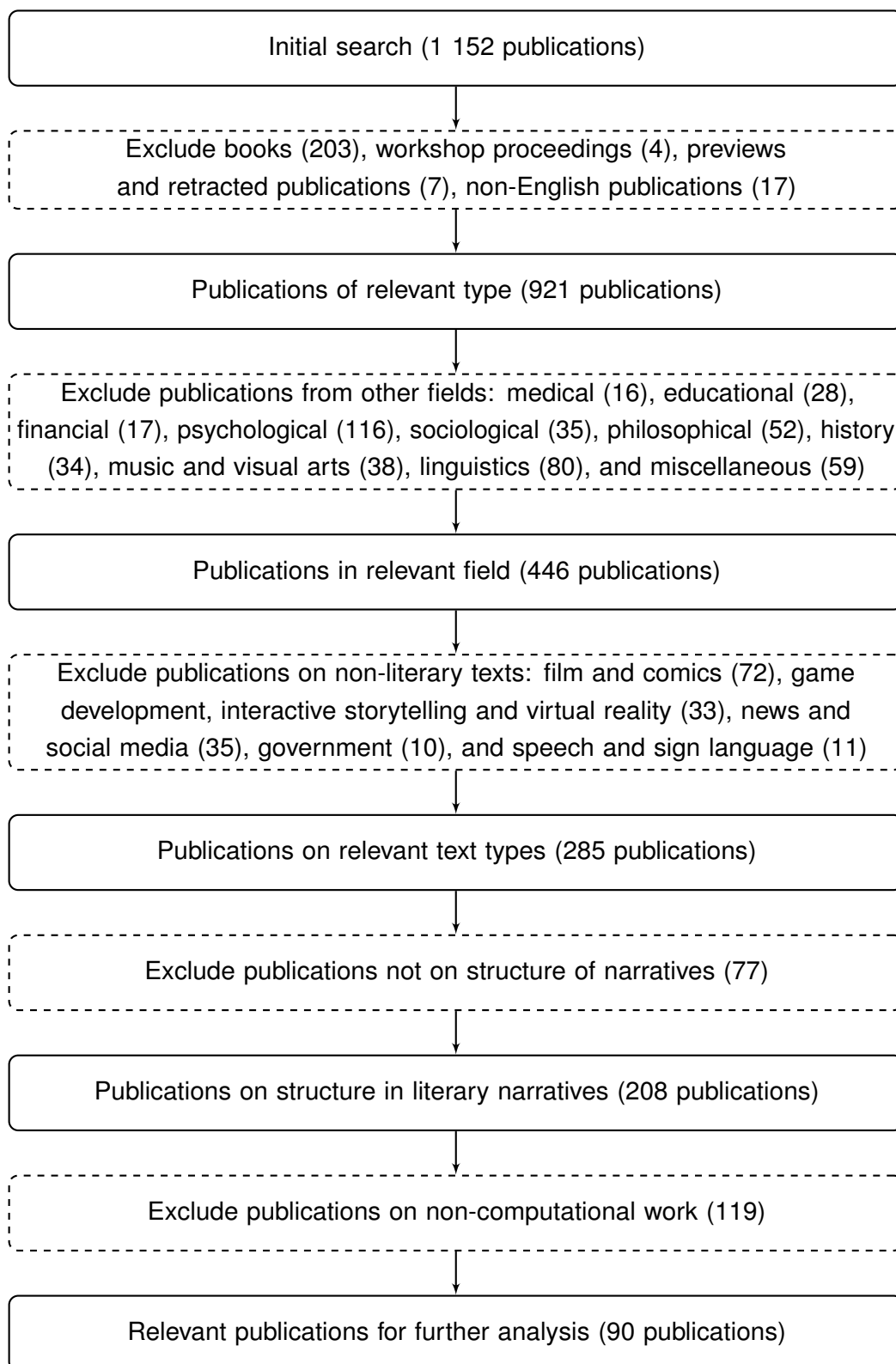


Figure 2.1: Overview of the publication selection process.

2.4.1 Terminology

To gain a deeper understanding of the main themes within the literature, similar or closely related terms are grouped together. The grouping is performed based on the manual labelling of the publications (see Section 2.3.2). This approach recognises the importance of building upon existing theoretical work on narrative understanding, allowing us to leverage established definitions in this review.

To link terminology found in the literature to the theoretical basis presented in Section 2.2, the terminology related to plot, narrative levels and the temporal aspect are looked at first before going into more fine-grained elements, such as segment, scene and event. Social networks, which represent a different type of narrative structure is then looked at.

Order

As mentioned in Section 2.2.1, order is defined as the sequence of events in relation to the order of narration. Other terms used to describe order are “temporal ordering” and “hierarchical structure,” which this study will collectively refer to as “order”.

Plot

Narratologists differentiate between the “what” and the “how” of narration. However, different theorists use varying terms. As defined in Section 2.2.1, Genette et al. refer to *histoire* and *récit*. Similarly, the Russian formalists introduced *fabula* (the chronological sequence of events in a narrative) and *syuzhet* (the representation of events through re-ordering of the temporal sequence). This review, will refer to the terms story and narrative discourse, but regardless of the terms being used, there seems to be a consensus that narrative discourse (also *syuzhet* and *récit*) focuses on the arrangement of events. This might also include time shifts (jumps in the story’s timeline), pacing (speed at which the story progresses), and perspective (point of view from which the story is narrated).

Story, on the other hand, is what this study in some cases might refer to as plot structure. The plot structure can also refer to narrative stages: the organisation of a story into distinct phases, with various theories offering different approaches. One well-known example is Freytag’s Pyramid, which outlines five stages: exposition (world introduction), rising action (builds tension), climax (peak moment), falling action (aftermath), and resolution (provides closure) (Freytag, 1894). Other theories provide alternative frameworks, such as the three-act structure credited to Aristotle, often extended to four parts, the five-stage model (Todorov, 1971), and 31 folk-tale functions (Propp, 1968). Research on plot usually aims to identify narrative stages, as defined by one of these approaches.

Narrative levels

For Genette et al., the narrative level changes when a new narrator enters, however, Ketschik et al. (2021) notes examples where the narrative level changes without the narrator changing. For example, any new story occurring in a narrative should be seen as a new narrative. This means the definition of boundaries of a narrative level needs (re)defining (Ketschik et al., 2021).

Eisenberg and Finlayson (2021) differentiate between *embedded narratives* and *interruptive narratives*. Embedded narratives exist within the original story, such as a character telling a story in a dialogue (Eisenberg & Finlayson, 2021). In interruptive narratives, the original narrator is interrupted, for instance, when the narrator changes (as suggested by Genette et al., 1980) or the point of view changes (for example, from first person to third person narrator). This often happens in the form of different narrators for different chapters.

While Ketschik et al. (2021) point out that anachronisms (flashbacks and flashforwards) are concerned with the chronological order of discourse and as such do not leave the present narrative level, Schwan et al. (2019) suggest that there is a correlation between narrative levels and temporal aspects.

The review found that some work refers to “subplots” or “plot levels”, especially in work on generative narratives. However, since these are not clearly defined, this review considers these terms to refer to narrative levels. Regardless of the specific definition of a narrative level, the detection of narrative levels is an important step for any text analysis with a focus on narrative discourse, character or temporal aspects of a narrative (Reiter et al., 2019).

Temporal aspect

Given that narratives rely heavily on temporal aspects like order, duration, and frequency, this review examines publications focused on time in narratives. Relevant terminology describing the temporal aspects in the literature include terms like “temporal”, “temporal information”, or “time”. Publications that focus on these aspects are clustered together in this review under the “temporal” keyword.

Time in narratives examines the relationship between story time (the passing of time within a story) and narrating time (the time it takes to narrate a story), focusing on the duration of narrative segments. Underwood (2018) and Yauney et al. (2019), building on Genette et al. (1980), distinguish between scene, summary and ellipses (see Section 2.2.1).

Anachronisms, such as analepsis and prolepsis, deviate from the main temporal progression but remain on the same narrative level (Kearns, 2021). Analepsis shifts to past events, while prolepsis appears as visions, prophesies and foreshadowing (Eisenberg & Finlayson, 2021). Both can be embedded, where the original narrator is telling a story in the present

and then shifts to telling a story about events in the past, or interruptive, where the original narrative is replaced or interrupted by a new narrative of events that occurred before the original narrative (Eisenberg & Finlayson, 2021). Here, the original and the interrupting narrator do not have to be the same narrator or have the same point of view. Flashforwards are typically interruptive (Eisenberg & Finlayson, 2021), as narrators usually pause the present story to predict future events.

Segment

The concept of *segments*, a key component of narrative structure, remains undefined in narratological theory, though implicitly present (Reiter, 2015). Evidence for their existence lies in Genette et al.'s use of categories like anachrony and pause, which function within "specific 'narrative segments'". Yet, Genette et al. leave them undefined, creating a challenge for computational work that requires clear descriptions of boundaries.

To address this gap, this review considers a segment as a logical unit within the text, representing shifts in topics, themes, or events. Reiter (2015, p. 34) defines a narrative segment as a "coherent and separable sub-sequence of events". Segments can include logical units of narrative stages (plot structure), scenes, narrative levels or time.

Later in this chapter different segmentation tasks are explored where terms like "scene detection", "scene segmentation" and "identification of scene changes" refer to identifying segments made up of scenes, which often overlap with identifying narrative levels and time shifts.

The term "segment" is rarely used in the context of plot, where a narrative stage is, according to the definition this review provides, also a segment. This is likely due to the lack of a formal definition in narratology. Instead, the term appears to have originated from computational work focused specifically on scene segmentation, as evidenced by Reiter (2015). "Boundaries" or "transitions" refer to the edges, thresholds or bounds between segments. This study will use "segmentation" to describe identifying segments or their boundaries.

Scene

Most research (see, for instance, Barth & Dönicke, 2021; Gius, Reiter, & Marcus, 2019; Piper & Bagga, 2022; Zehe, Konle, Dümpelmann, et al., 2021) base their definition of a scene on Genette et al. (1980, p. 86), who define a scene as: "scene, most often in dialogue, . . . , realises conventionally the equality of time between narrative and story", see also Section 2.2.1. However, as Zehe, Konle, Dümpelmann, et al. (2021) argue, defining a scene only by its time aspects does not allow us to define the boundaries of a scene, in other words, the points in the narrative where a scene begins and ends. At the same

time, however, this is essential to analyse the narrative discourse. Therefore, several more operational definitions have been proposed or extended.

Alrashid and Gaizauskas (2021), Gaizauskas and Alrashid (2019), Konle and Jannidis (2022), Kurfali and Wirén (2021), Mani (2005), Zehe, Konle, Dümpelmann, et al. (2021), and Zehe, Konle, Guhr, et al. (2021) agree that scenes also include character constellation, space time, and action. Scene is neatly defined by Gius, Jannidis, et al. (2019b, p. 3169) as “a segment of the *discours* of a narrative which presents a part of the *histoire* such that (i) time is equal in *discours* and *histoire*, (ii) place stays the same, (iii) it centres around a particular action, and (iv) the character constellation is equal”. Mani (2005) adds that scenes can be of various types ranging from “setting scenes” focused on perception events to those containing actions or thoughts.

Event

The definition of plot relies heavily on the concept of “event”. This means that it is necessary to look into how events are defined. Similar to plot, unfortunately, this term is hard to define. Most publications skim over the definition of event. Although event is typically described as a change of state (Gius & Vauth, 2023), abstract eventualities (Mani, 2005), or the smallest unit of narration (Gius & Vauth, 2023). Sandeep et al. (2023) and Zehe, Konle, Dümpelmann, et al. (2021) claim that the definition of event is unclear and therefore using events to try and establish the plot of a narrative is a vague task. Instead, most research refers to scenes (and not events) as Zehe, Konle, Dümpelmann, et al. (2021) suggest.

Character networks

In computational literary studies, character networks represent relationships between characters, visualised as graphs with nodes for characters and edges for relationships. Properties like size or thickness can indicate relationship strength or character frequency.

To create these networks, characters are identified through Named Entity Recognition (NER), which extracts entities such as names, locations, dates, and other pre-defined categories. In narrative analysis, NER helps identify character names, while recognising locations and dates aids scene identification. Relationships between characters are usually identified through co-occurrence within a sentence or paragraph.

While NER provides a useful starting point for identifying named characters, it does not fully capture the complexity of character recognition in literary texts. Characters are often introduced or referred to using descriptive phrases (for example “the baker”, “her father”), role-based labels, or pronouns, which may not be recognised by standard NER tools. As a result, additional methods such as co-reference resolution (CR), entity linking, or the use

of heuristics informed by literary annotation schemes (for example, BookNLP³) are often required to capture the full range of character mentions. BookNLP was originally designed to provide structured information from literary texts, including character identification, CR chains, and even event annotations. Although later versions have moved away from event annotation largely due to difficulties in defining events consistently, especially in the presence of negation, modality, or embedded clauses, its early inclusion of event detection highlights the need for broader entity representations in narrative processing. Sims et al. (2019) highlight these challenges in their work on literary event detection, showing how character recognition in literature must account for contextually implied and narratively significant entities beyond named mentions. For the purposes of character network construction, this broader notion of character identity must be taken into account.

Additionally, narratological frameworks such as Greimas' actantial model (Greimas, 1984), which builds on Propp's functional analysis of narratives, provide a theoretical lens for interpreting these relationships. Greimas (1984) categorises characters as actants with specific roles (for example, hero, opponent, helper) that define their narrative functions and interactions, a perspective widely applied in social sciences and political studies to analyse relational dynamics. Integrating such frameworks with computational methods can enrich the analysis of character networks by mapping actantial roles onto network structures, revealing how narrative functions shape social interactions in literary texts.

Analysing character networks provides insights into narrative structure and character roles, with main characters often appearing more frequently and having stronger relationships. For example, Greimas' actantial model can be used to identify whether a character with high centrality in a network corresponds to the "hero" or another key actant, offering a deeper understanding of narrative dynamics.

Various studies (Botelho, 2021; Chen & Bunescu, 2021; Deng et al., 2022; Droog-Hayes et al., 2018; Konle & Jannidis, 2022; Min & Park, 2016a, 2016b, 2016c; Porteous et al., 2016) explore this concept, using terms like "character believability", "social networks", "character networks", and "character types". For simplicity, this review groups them under "social networks."

2.4.2 Annotation strategies

The initial step in annotating narrative text involves establishing a comprehensive framework for narrative concepts. This framework should define these concepts with sufficient detail to facilitate the development of clear, computable annotation guidelines and boundaries.

Given that annotation poses a significant challenge in the field (among others due to vaguely described concepts as discussed earlier), this review aims to address this by examining proposed annotation schemes and exploring existing annotated data sets.

³<https://github.com/booknlp/booknlp>

Using these frameworks enhances efficiency and consistency in future research, which is discussed in this section.

Annotation of narrative structure involves a particular attention to temporal aspects. The social networks research requires annotation of the characters in the narrative, which falls under the task of NER (and once identified, they serve as the basis for the creation of the network). NER is a well-understood task and more information can be found in overview articles, like Mansouri et al. (2008), Nadeau and Sekine (2007), and Sharnagat (2014).

Of the publications in this review, 23 are identified that discuss annotation strategies. These annotations are mostly in the form of narrative segments. Out of these publications, ten are part of the *The systematic analysis of narrative texts through annotation* (SANTA) project and discuss the same data set or extend this task. This annotation task, along with other annotation strategies for narrative levels, are reviewed in this section. Annotation strategies that focus on plot, temporal, and lexical features are also looked at.

Plot

Li et al. (2018) established an annotation scheme for narrative structure where the fundamental concepts came from several prominent theories like Freytag (1894), Labov and Waletzky (1997), Labov and Waletzky (1967), Labov (2013), Ouyang and McKeown (2015), Prince (1973), and Todorov (1971). To annotate the narrative structure, Li et al. (2018) introduced ten functional labels: (i) abstract, (ii) orientation, (iii) complicating action, (iv) the most reportable event (MRE), (v) minor resolution, (vi) return of MRE, (vii) resolution, (viii) aftermath, (ix) evaluation, and (x) direct comment to audience. To show the practical applicability, Li et al. (2018) annotated 360 unique texts from online sources (Quora and Reddit) and stories previously annotated by Lukin et al. (2016) with these labels. These texts had strict constraints. The text should contain only one story, consist of between 90 and 700 words and have fewer than six lines of dialogue. The non-narrative elements must be less than 50% of the text. For inter-annotator agreement (IAA), they used Cohen's Kappa, which indicated fair agreement between annotators.

Reiter (2015) discussed the problem of annotating narrative segments. Their focus was on segmenting text into plot parts. What set this publication apart from previous work, is that they did not use information on topical shifts to segment the text. Instead, Reiter (2015) used crowd-sourcing and student annotators to identify segments and measure the IAA with Fleiss' k (Fleiss, 1971) where the annotators reached a k -agreement of 0.27 and 0.21.

Narrative scene and level

The most notable work on narrative structure annotation is the SANTA project introduced by Gius, Reiter, and Marcus (2019). This project centred around defining annotation

guidelines for narrative levels, applied to a shared task for consensual assessment. The first task produced several publications (Barth, 2019; Bauer & Lahrsow, 2020; Eisenberg & Finlayson, 2021; Hammond, 2021; Kearns, 2019; Ketschik et al., 2019; Wirén & Ek, 2021), with later extensions (Barth, 2021; Kearns, 2021; Ketschik et al., 2021). The data set is discussed in Section 2.4.3 and a follow-up task on automatic detection is discussed in Section 2.4.4. Several other publications focus on narrative levels, such as Zehe, Konle, Dümpelmann, et al. (2021), which introduced a German corpus annotated with information on scene boundaries and segment information. For example, they differentiated between scenes and non-scenes, and labelled the boundaries according to SCENE-SCENE, SCENE-NONSCENE, and NONSCENE-SCENE. Their annotation method considered narration speed and spatial granularity. They proposed a container principle, where spaces (for example, a hotel lobby and elevator) or actions are grouped based on context. Using these ideas, two annotators achieved high IAA, and this annotation scheme was later used in Konvens 2021 (Zehe, Konle, Guhr, et al., 2021). SceneML, introduced by Gaizauskas and Alrashid (2019) and refined by Alrashid and Gaizauskas (2021), explored scene segmentation guidelines, showing feasibility but suggesting improvements. Newberry and Bailey (2019) highlighted the importance of context in event segmentation, with context improving annotation quality. Several publications address scene and narrative level segmentation with a focus on temporal aspects. Kearns (2020) provided guidelines for annotating time in fiction, focusing on anachronisms, narrative levels, stream-of-consciousness, and free indirect speech.

NarrativeTime (Rogers et al., 2024) introduced a timeline-based framework, linking events by before/during/after relations, helping to structure the narrative based on time progression. Eisenberg and Finlayson (2021) annotated temporal narrative levels, distinguishing embedded and interruptive narratives and flashbacks (see Section 2.4.1 on Narrative levels and temporal aspect). While helpful, their data is drawn from TV scripts and court proceedings⁴, which differ from literary texts, limiting its applicability to training systems for literature. Joon et al. (2017) proposed crowd-sourcing to align narrative event's temporal order. Narratology researchers interpreted a summary of a text and assisted a crowd in reordering the text in chronological order.

Linguistic and temporal features

To identify interesting structural properties in narrative, information from the narrative needs to be extracted or identified to provide guidance to the (computational) annotation process. Often, linguistic and temporal features are used for this purpose. The publications mentioned here focus on how discourse, expressions, and tense affect the interpretation and presentation of narratives.

Aksan and Aksan (2015) examined how certain linguistic constructs, like multi-word expressions, play a role in defining and distinguishing genres. Their work explored how

⁴This work is part of the SANTA project and as such is kept during the source selection phase.

language patterns are used for different rhetorical effects in narratives. They annotated their data with properties relating to time, place, transition signals, and interactional markers. Also focusing on linguistic structures, Reese et al. (2007) examined the effect of linguistic structures on the overall narrative. They provided a framework for analysing rhetorical structure, focusing on how texts are organised to achieve communication goals.

Shifting the focus to the temporal aspect of narratives, E. Lee (2017) focused on how the pluperfect tense is used to manage shifts in time, which is essential in understanding a story's progression and structure.

2.4.3 Data sets

Several publications describe data sets and related resources, focusing on human-annotated or publicly available data sets. Only substantial contributions are included and referenced but unused data sets are excluded.

The SANTA project provides a shared data set consisting of annotated narrative text, focusing on narrative levels. The data set consists of seventeen training texts and eight evaluation texts, each under 2 000 words (Reiter et al., 2019). All texts can be found on a GitHub repository⁵. All participants of the SANTA project as well as a group of (paid) student assistants, helped annotate the text according to the final narrative guidelines that resulted from this task. An extended data set includes 22 novels, including the annotated texts that formed part of the SANTA project⁶.

Another data set for the task of scene or narrative level segmentation is described by Kearns (2020) who introduced a corpus of six modernist novels (dating between 1909 and 1935) and seven hypertext fiction parts (dating between 1986 and 2017). The author annotated narrative level and time disruptions (analepsis and prolepsis) to quantify how time functions in these genres. While the author annotated the text, IAA was calculated by comparing a small annotated corpus with the similar guidelines and annotations of the SANTA project, in which the authors also participated, resulting in a gamma of 0.24, the third highest of the projects partaking in the SANTA project. Note that gamma can compare annotation schemes with different levels of complexity.

Underwood (2018) estimated the time (in minutes) that is narrated in a text segment. They introduced a data set of 1 765 passages (~250 words each) from 108 best-selling and randomly chosen novels. Yauney et al. (2019) later extended this data set with 56 newly labelled passages that they promised to release publicly to the extent possible under copyright.

Shifting the focus to character annotation, the following resources offer valuable information. Digital Yoknapatawpha is a comprehensive data set of all the locations, characters, and

⁵<https://github.com/SharedTasksInTheDH>

⁶<https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/German-Scene-Dataset>

events in Faulkner's Yoknapatawpha fictions, aiming to visualise the data through a series of “deep atlases” and other displays (Burgers, 2021). The NOC (Non-Official Characterisation list) is a publicly available data set of characters used in story generation systems (Botelho, 2021). Chen and Bunescu (2021) created a data set altering narrative perspective (for example, first to third person).

The following work uses resources for text annotation and analysis. The HathiTrust Digital library⁷ provides page-level word frequencies from millions of volumes of text (Yauney et al., 2019). The AMR (Abstract Meaning Representation) bank is a manually constructed set of thousands of English sentences paired with their semantic representations. It is used by Droog-Hayes et al. (2018). Rogers et al. (2024) re-annotated existing TimeBank data with narrative time data⁸, and Reese et al. (2007) presented a reference manual for the annotation of MUC6 and ACE2 corpora for discourse structure. These texts are marked with rhetorical relations such as elaboration, contrast, and cause-effect, forming a resource for analysing the organisation of rhetorical elements in narrative texts.

2.4.4 Computational approaches

Computational approaches to structure identification in narrative text can be categorised into two main groups: *story (plot) modelling* and *narrative discourse (narrative scene and level) segmentation*. Most work on plot modelling uses either topic modelling or sentiment analysis, whereas scene or level segmentation generally relies heavily on human-annotated data to identify different narrative levels or scenes. Both of these categories often make use of character networks and entity extraction as well as lexical features.

Plot

Research on plot extraction often uses sentiment analysis to identify literary patterns (Deng et al., 2022; Konstantopoulos, 2023; Min & Park, 2016b; Sui et al., 2023; Woods, 2022), based on Vonnegut's idea of emotional arcs (Vonnegut, 2005). Several studies, including Burgers (2021), Fudolig et al. (2023), Kim (2022), Piper et al. (2021), and Zehe, Konle, Dämpelmann, et al. (2021), used the Syuzhet R package⁹, developed by Jockers (2015), which assigns sentiment scores to sentences and applies a Fourier transformation to the sentiment scores to reveal emotional arcs. These arcs represent the rise and fall of emotions throughout a story, following Vonnegut's concept of rising and falling fortune.

Woods (2022) combined sentiment analysis and lexical changes to quantify emotional arcs, while Deng et al. (2022) mapped character emotions over time and space, linking emotional shifts to specific events and locations, creating spatial-temporal emotional

⁷<https://www.hathitrust.org/>

⁸<https://github.com/text-machine-lab/nt>

⁹<https://doi.org/10.32614/CRAN.package.syuzhet>

paths for characters. Their framework provides a broad view of how shifts in emotional development over time influence the narrative plot.

Also, focusing on unsupervised plot extraction through sentiment analysis is the work of Sui et al. (2023), who applied sentiment analysis to nonlinear novels, handling fragmented emotional shifts in temporally disordered narratives like *Mrs. Dalloway*. While Deng et al. (2022) focused on linking emotional shifts to time, space, and character networks, this study accounts for the complexities of emotional shifts in texts where chronology is not straightforward. In contrast, Min and Park (2016b, 2019) used sentiment analysis and character networks to represent the narrative as interaction maps between characters with associated sentiment values and keywords to understand the plot dynamics. These studies focused less on the individual characters and more on how relationships evolve across the narrative represented by the character network.

O. Lee et al. (2021) proposed that plot is driven by characters' traits and behaviours, derived from network interactions, dialogue lengths, and degrees of influence in character networks, showing how these factors shape plot progression. Lastly, Sandeep et al. (2023) proposed a spatial relationship categorisation task to help understand plot structure. They assigned a spatial relationship category for every character and location co-mentioned within a window of text. They also annotated 2 500 book excerpts and trained a model using contextual embeddings as features to predict these relationships.

Character networks

Along with the work discussed in the previous section that uses character networks to identify plot (Deng et al., 2022; Konle & Jannidis, 2022; O. Lee et al., 2021; Min & Park, 2016b, 2019), other research used character networks with other goals in mind. Work such as Adanay and Sporleder (2015) clusters novels based on the relationships between characters, providing insight into genre and authorship based on how interactions are structured. Although genre and authorship identification are not the primary goals of Burgers (2021), their work comes in useful for this purpose as well. They based their work on the Digital Yoknapatawpha initiative, which aims to map out Faulkner's fictional county by visualising intricate and multi-layered narrative elements in particular geographical and structural representations of narrative elements.

Character networks are also being used to validate literary theories (Jayannavar et al., 2015; Labatut & Bost, 2019). The social networks identified by Jayannavar et al. (2015) are analysed for metrics like centrality, clustering, and degree distribution, which are then compared to established literary theories. For example, if a literary theory posits that a particular character is central to the plot, this means that the character should have a high centrality score. A more detailed survey of computational tools for extracting and analysing character networks is presented by Labatut and Bost (2019).

Bi et al. (2018) and de Vree (2016) used social networks to interpret narratives. Specifically, de Vree (2016) used co-occurrence analysis to identify patterns in biblical texts, while Bi et al. (2018) compared writing styles by analysing the centrality and connectivity of characters.

Some publications are specifically focused on analysing character interactions. For instance, Agarwal (2011) and Fields et al. (2022) focused on using network analysis and named entity recognition to map out social relationships from textual data, distinguishing between personal and social networks. Similarly, J. Lee (2017) and Min and Park (2016c) applied social networks to analyse character dynamics within specific literary works, exploring how social interactions shape the narrative. On the other hand, publications such as Macovei (2017) and Min and Park (2016a, 2016b) tracked character development over time, highlighting how their interactions evolve throughout a narrative. Finally, Chen and Bunescu (2021) and Porteous et al. (2016) focused on narrative generation and viewpoint analysis, differing slightly from the others by examining narrative structure from the perspective of plot generation and point of view shifts, rather than character interaction alone.

Narrative scene and level

A popular way of segmenting narrative text is scene segmentation, where character constellation, place, and time stay relatively constant. Some studies differentiate between scenes, non-scenes, flashbacks, and narrative levels (Zehe, Konle, Dümpelmann, et al., 2021). The SANTA project's second task, the Shared Task on Scene Segmentation (STSS) at KONVENS 2021, explored various approaches for automatic scene or narrative-level detection.

Schneider et al. (2021) based their approach on the idea that significant shifts in narrative content, such as new locations, characters, or events, lead to measurable changes in word embeddings over time. They created segments in their text and calculated the average word embedding vector for each segment. They then calculated the difference between consecutive segments to represent the semantic shift between neighbouring portions of text. Large shifts indicated scene boundaries.

Gombert (2021) trained twin BERT embeddings, separating scene-border sentences from in-scene ones. The sentence embeddings produced by this model serve as feature vectors for input into a gradient-boosted decision tree ensemble, which then makes the ultimate predictions. In the STSS leaderboard, the system achieved second place in track 1 and first place in track 2.

Barth and Dönicke (2021) proposed a statistical approach focusing on feature vectors covering time, space, and entities. They used a random forest classifier to find Scene-to-Scene boundaries but never non-scenes. They attribute this lack of non-scene identification to the small training set available.

Additional contributions were by Kurfali and Wirén (2021), who proposed a Bidirectional encoder representations from transformers (BERT)-based solution with a weighted cross-entropy loss, and Hatzel and Biemann (2021), who utilised a transformer model, enriched with features derived from CR to identify scene boundaries by analysing pairs of sentences. Zehe, Konle, Dümpelmann, et al. (2021) applied a text classification model to detect transitions and label them accordingly.

Beyond the STSS, Portelance and Piper (1997) examined the pacing of narrative scene changes and the distribution of plot lines. They segmented text based on shifts in entities, actions and objects recognised through POS tagging. In the context of narrative pacing, Underwood (2018) and Yauney et al. (2019) annotated time elapsed in scenes to analyse patterns and trends in narrative pacing. This work can also be applied to narrative genres or, as Yauney et al. (2019) mentioned, to analyse trends in the language usage associated with shorter and longer elapsed narrative time in the form of word co-occurrence patterns.

Visual techniques for scene segmentation include DeBuse and Warnick (2023)'s graph-based system that represents the order of scenes and *Narrelations* (Schwan et al., 2019), which visualises narrative levels and temporal aspects. Richardson and Hay (2022) introduced a dynamic topography to visualise temporal-spatial relationships of both linear and non-linear narrative structures that allow the users to explore how time and space evolve across the narrative, particularly in cases where the narrative unfolds across multiple locations or time periods. Bushell et al. (2022) also presented a topological approach for the mapping of literary texts and is centred around Bakhtin (1981)'s concept of the “chronotope” or “time-space”.

Other work on scene segmentation and narrative levels that also focuses specifically on temporal aspects of the task, includes Bae and Young (2008), who looked at anachronisms in story generation to create surprise in the reader. Their system determines the point in a linear story where a flashback or flashforward can be inserted.

Linguistic and temporal features

In the identified publications, narrative analysis examines both lexical semantics (word meaning) and discourse structure (text organisation). Several studies explore how word choice impacts narrative understanding. Aksan and Aksan (2015) analysed frequent tri-gram patterns in fiction versus non-fiction, while Fudolig et al. (2022, 2023) studied lexical chain evolution and power/danger-related words across books. Mahmoodi and Savab (2023) introduced a “semiotic ladder” to model how meaning builds in narrative structure.

Temporal aspects benefit from semantic analysis. Becker and Egetenmeyer (2018) and E. Lee (2017) explored temporal discourse structure, with Becker and Egetenmeyer (2018) integrating tense, aspect, adverbials, and connectors to understand relationships between elements. Zhang (2020) and Zhang and Xue (2018) designed a neural temporal

dependency parser using lexical features to parse temporal expressions and events into a dependency tree structure, achieving promising results. Similarly, Reese et al. (2007) segmented text on a sentence level to identify events, while Dery and Koenig (2014) used discourse structure to investigate what temporal factors determine the movement of narrative time and how much narrative time moves between events.

Droog-Hayes et al. (2018) applied Propp's folktale functions and used Abstract Meaning Representation (AMR) for CR to identify character importance. CR aids sentence disambiguation and machine readability. CR is also explored by Krug et al. (2015), who presented a rule-based system for CR in novels, while Chen and Bunescu (2021) explored CR annotations for training data in "mentioned selection tasks" with the ultimate goal of the system being to change narrative perspectives. Other research (Abrusán, 2021; Ren et al., 2023) also focused on perspective shifts in narratives, which are related to dialogue. Muzny et al. (2017) measured text dialogism using grammatical features, drawing on Bakhtin (1981)'s theory of dialogism.

2.4.5 Applications

Many of the identified articles on generating narratives relies on neural network language models trained on large data sets to produce coherent, contextually relevant text. These models require plot structure, character behaviour, and narrative time knowledge, with features like flashbacks enhancing human-like storytelling. Publications on temporal ordering in narrative generations include Bae and Young (2008), Botelho (2021), Lukin and Walker (2019), Porteous et al. (2016), Porteous and Cavazza (2009), and Porteous et al. (2010). Rosenbaum (2019) proposed a foundational model for the "narreme" (a term mostly used in game studies or generative narratives), outlining its potential values, permissible combinations, and codified conventional patterns within the construction of narrative objects that can inherently be recognised as a "story".

Narrative summarisation condenses longer narrative texts into shorter versions. For simplicity, narrative abstraction is also considered as part of this process. Both tasks require understanding the structure and flow of time within a narrative to accurately capture the key information. Droog-Hayes et al. (2018) suggested using Propp's folktale theory for better summaries, while Mani (2005) emphasised the importance of temporal information. Crampes et al. (1998) explored narrative abstraction, addressing event identification and narrative order.

Taivalkoski-Shilov (2019) argued that literary machine translation benefits from narrative structure awareness. Additionally, other work includes Miller and Park (2020), who provided a description of projects attempting to create analytical models of narratives and added some implications these models might have for critical inquiry of stories and their structure. Building on this foundation of narrative analysis, Rodosthenous and Micheal

(2019) presented an online platform for story understanding. Novakova and Siepmann (2020) focused on identifying motifs in narrative text.

2.4.6 Term correlations

To illustrate how concepts across publications relate, assigned labels are analysed (Section 2.4.1) by computing pairwise correlations. Using the hypothesis-free labelling approach, the following labels are identified: “Theory”, “Annotation”, “Method”, “Structure”, “Narrative level”, “Order”, “Segmentation”, “Temporal”, “Scene”, “Plot”, “Event”, “Sentiment”, “Characters”, “Lexical”, “Generative”, “Summarising”, “Longer text”, and “Shorter text”.

Figure 2.2 visualises these correlations. The cells indicate the correlation’s r values, which indicate the strength of the linear relationship between the two concepts. Absolute values $|r| \geq .50$ are typically considered strong correlations, $.30 \leq |r| < .50$ moderate, and $|r| < .30$ weak.

Strong correlations between the labels “Segmentation” and “Narrative level”, and “Scene” and “Narrative level” are observed. This may be due to the large number of publications in this review forming part of the shared task on scene segmentation, discussed in Section 2.4.2, which had a strong emphasis on connecting the annotation guidelines of scenes and narrative levels to theory. The first part of the shared task focused on the annotation of narrative levels and scenes for segmentation, and this is also reflected in a high correlation between annotation, narrative levels, segmentation, scene, event and time.

Furthermore, the “Structure” label is correlated with “Order” and “Plot” as expected. The “Order” label is further correlated with “Temporal”, “Event”, and “Plot”, but also with “Generative”. The “Plot” label is also correlated with “Sentiment” and “Characters” and, interestingly, negatively correlated with “Annotation”. Most work that focuses on plot will resort to sentiment analysis or character networks that do not require human annotations.

The “Theory” label is lightly correlated with labels “Structure”, “Scene”, “Plot”, and “Event” and negatively correlated with “Sentiment”, “Characters”, “Lexical”, and “Generative”. This indicates that most theory publications focus on annotation, narrative level, scene, event, and temporal aspects of a narrative, which illustrates that these publications tend to turn to literary theories. In contrast, publications focusing on lexical features, sentiment, and character networks generally do not use theoretical work.

The “Summarising” label is strongly correlated with “Shorter text” and negatively correlated with “Longer text”, which is interesting. Additional correlations are with “Lexical” and “Event” (which relates to features used). “Longer text” labels are correlated with more structural labels, such as “Annotation”, “Method”, “Structure”, “Narrative level”, “Order”, “Segmentation”, and “Scene”.

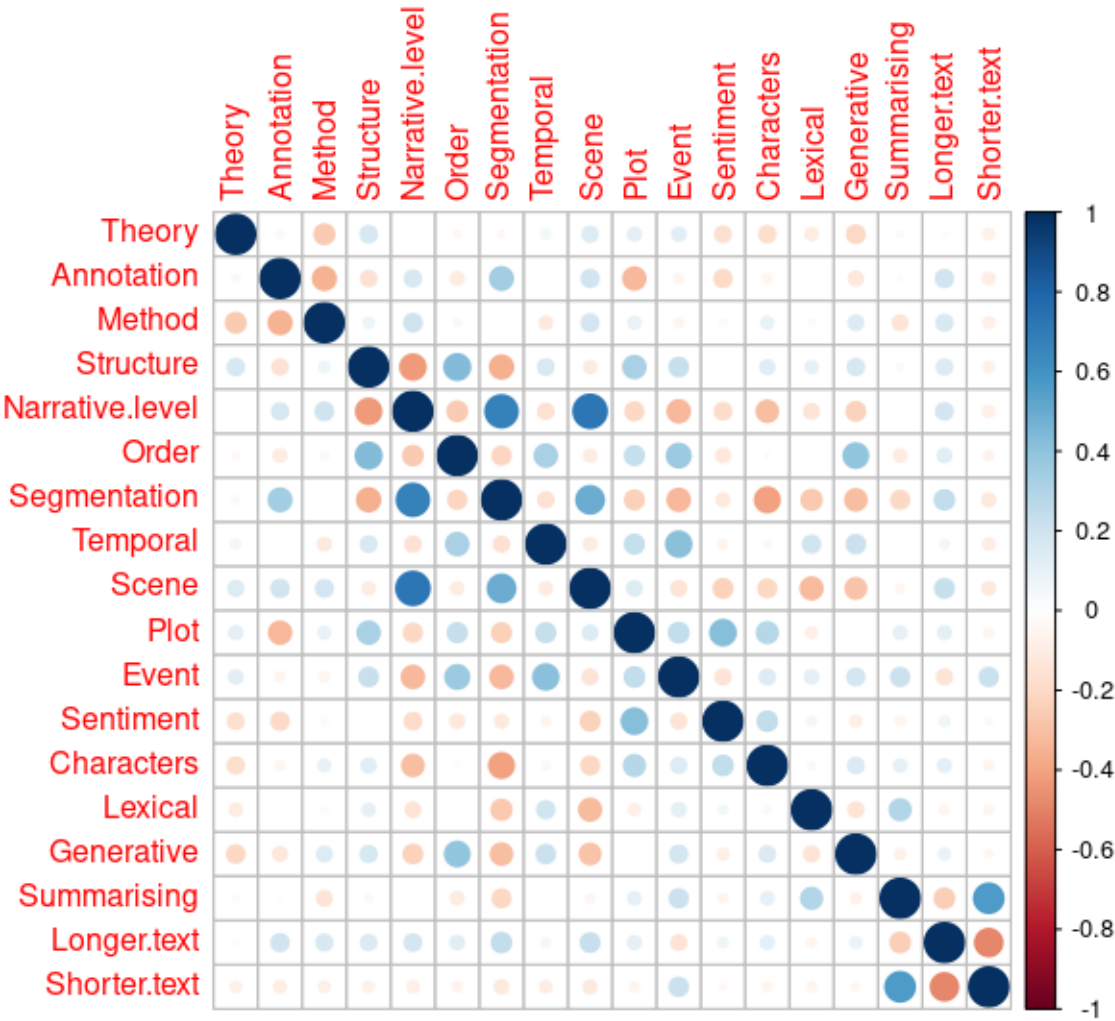


Figure 2.2: Correlogram showing how labels correlate to each other. Blue represents positive correlations and red represents negative correlations between labels.

2.5 Discussion

The review in this chapter explored existing work on identifying narrative structure, including social networks which analyse character relationships through co-occurrence, and temporal aspects, such as scene segmentation and narrative levels. Various approaches rely on different language properties and techniques.

This chapter structured the review using Genette et al. (1980)'s temporal framework (order, duration, frequency), with most work focusing on order. Among the existing frameworks, the SANTA shared task contributed annotation standards for narrative levels and scene segmentation, making annotated data sets publicly available. This has led to SANTA becoming the standard in this area. No other general-purpose frameworks for annotating high-level narrative structure have gained similar traction. This dominance may be due to the difficulty of defining generalisable narrative categories across diverse literary genres, as well as the high cost of human annotation for such abstract features. Furthermore, many computational studies either avoid manual annotation or focus on lexical or sentiment features, which limits the incentive to develop new annotation schemes. As a result, while SANTA has significantly contributed to standardising work in this area, further development of alternative or complementary frameworks would be valuable.

A larger variety of annotated data sets or extensions of the existing data sets would also be valuable for future research. As narratives show a wide variety, large amounts of annotated data are needed to cover different situations.

While the SANTA project provided annotation guidelines, additional annotation guidelines may be useful, such as guidelines that deal with tasks other than the identification of narrative levels or scene segments. This may, for instance, be in the form of specialised guidelines, such as those that deal with narratives in particular genres, such as whodunits.

Note that even though there have been several publications on the identification of scenes, etc., limited research on the identification of the ordering of events on a timeline has been done. For the placement of events on a timeline, the identification of structure in the narrative can be a first step. The different scenes or plot segments should then be ordered chronologically. To order scenes on a timeline, it is important to have access to information on the duration of the scenes or segments. Notable work on duration includes Underwood (2018), who identifies the time elapsed in a scene.

Comparing the duration of different scenes can reveal themes and narrative development. Related to this is the task to identify the speed of narration using factors like sentence length, or to identify dialogue and descriptive passages (summaries). More work can be done in this area, for example, to differentiate between scene descriptions and summaries. Such information can be beneficial, among others, for narrative generation.

The identification of the temporal flow of a narrative provides the potential for deeper text interpretation and analysis. Narrative is inherently temporal; it unfolds across a

continuum of past, present, and future. Accurately identifying temporal attributes allows us to understand the sequence of events and themes embedded in a narrative on a higher level.

Comparing the computational work against theoretical research, it seems that order and duration have been investigated in varying degrees. However, frequency has received less attention. The task of detecting and analysing multiple narrations of the same event (typically from different viewpoints) would open new avenues for understanding narrative structure. It can provide another dimension of the contents of a narrative, which is more related to the intention of how a narrative is presented. It sheds light on recurring motifs, character habits, and thematic emphasis.

Note that most of the computational work uses some form of machine-learning. This means that either enough annotated data is needed, or unsupervised techniques should be applied. Additionally, one needs to decide which features to use. It is currently unclear which features perform best in the context of narrative structure identification. Because narrative structure is a high-level annotation, the relationship with the low-level features that can easily be extracted from the text is not trivial.

In addition to theoretical research, this review has highlighted various applications that can benefit from advances in this field. Automatic narrative analysis opens up potential in literary analysis, enabling large-scale studies of narrative style and thematic trends. Understanding the temporal flow of a narrative can benefit dialogue systems, generative text, automatic summarisation, and machine translation, enabling them to create more natural and coherent output.

2.6 Conclusion

This review examined what computational approaches exist for the structural analysis of narrative in literary texts, drawing on narratological theory to frame key terms such as scene, plot, and time¹⁰. This review described narrative structure related to characters in the narrative, which is typically presented as a social network. Three aspects of temporal structure as presented by Genette et al. (1980) are also described: order, duration, and frequency.

Based on terminology from the theory, existing academic publications on computational narrative structure were identified. These publications were described and grouped according to the theoretical concepts. This review first considered different annotation strategies. Much of this work (guideline development and annotation of data) stems from the SANTA project. Next, computational approaches to identify narrative structure at the

¹⁰Due to the rapid pace of NLP advancements, some emerging studies were not included as they were published after the selection of the publications in this review.

various levels were investigated, which was followed by an overview of applications.¹¹ The co-occurrence of terms used in the publications was then looked at.

Most of the existing research on temporal aspects of narrative structure deals with order and duration, but hardly any work on frequency was found. Again, most work comes from the SANTA project, which consisted of shared tasks. This led to a set of annotation guidelines and annotated data sets. Both can be reused.

While the overview shows the current dominant research direction, it uncovers promising avenues that remain mostly unexplored. First, in response to **RQ 2**: “Which under-explored aspects of high-level narrative structure are both foundational to overall story organisation and amenable to automated detection?”, the effectiveness of features for automatic detection of temporal changes in the narrative remains an under-explored area.

Second, the theoretical concept of frequency in narratology has not seen much attention in the computational area. Frequency provides information on a very high level of narrative structure, distinct from story versus discourse, by revealing how events are repeated or refracted through multiple viewpoints. While this represents a largely unexplored area with potential for future research, annotating frequency requires fine-grained tracking of event identity and iteration across the text. This aspect will not be explored in greater depth within the current thesis. It is therefore deferred for future research.

Finally, the knowledge gained from the different narrative structures has seen limited integration in practical applications. In particular in the context of automatic story generation, but also in student writing evaluation systems, among others, this can be a valuable addition. This thesis will also not pursue such applications, focusing instead on foundational advancements in structure detection.

The main takeaway from this review chapter is that existing annotation guidelines remain fragmented and unconsolidated, with varying definitions and approaches across studies. Consolidating these guidelines into a unified framework is essential before advancing to more complex computational models, as it would enhance comparability, reduce redundancy, and facilitate broader adoption in the field.

To answer **RQ 1**: “What computational approaches have been developed for analysing high-level narrative structures in literary texts?”, two dominant paradigms were identified, mostly focusing on order as defined by Genette et al. (1980): *plot modelling* via unsupervised methods (sentiment arcs, topic modelling, character networks) and *discourse segmentation* via supervised scene/level detection (mainly the SANTA task). Key techniques include BERT-based classifiers, word embeddings, and lexical/temporal features. However, *frequency* (Genette et al., 1980’s third temporal aspect) remains almost entirely

¹¹Of the surveyed frameworks, Genette et al. (1980)’s story-discourse distinction and LDA-based segmentation directly inform this thesis’ systems in Chapters 3 to 8, while others (e.g., network-based character analysis) provide contextual background but are not implemented.

unaddressed computationally, with no data sets or models for detecting iterative narration. Duration is limited to scene-level pacing (for example, Underwood, 2018).

There is also a clear answer to **RQ 3**: “What are the primary challenges in annotating high-level narrative elements, across diverse literary texts?”. Annotation is hindered by terminological inconsistency (for example, scene versus segment versus narrative level), genre sensitivity (SANTA guidelines struggle beyond short modernist texts), and boundary ambiguity (for example, container principle for space, focalisation shifts). IAA is low outside controlled tasks (for example, $\gamma = 0.24$ in Kearns, 2020), and no unified, modular scheme exists for combining plot, scene, level, and time. The high cost and subjectivity of human annotation further limit data set scale and diversity.

Building on these insights, this thesis will focus on two complementary levels of narrative structure, each addressed in a dedicated part. Part I targets thematic arcs, the macro-level causal progression of the story (*histoire*). Part II targets discourse-level annotation like scenes, narrative levels and temporal aspects (anachronisms). By prioritising these elements that are central to Genette et al. (1980)’s *order* and *duration*, yet computable within the thesis scope, foundational resources for scalable computational narratology is laid, while deferring frequency and social-network integration to modular future extensions.

Part I

Story

Chapter 2 established that plot, the causal macro-structure of the story (*histoire*), remains a “slippery” concept in computational narratology (Abbott, 2007; Baroni & Revaz, 2016). While narratologists debate its boundaries, computational studies often reduce plot to sentiment arcs or character networks, bypassing the need for explicit story-level segmentation. This gap is particularly acute in the absence of standardised, high-level annotated data sets for plot structure.

This part addresses that gap by realising plot as *thematic narrative arcs*: macro-level shifts in story topics that trace the causal progression of events. Arcs are extracted via both computational topic modelling (Chapter 3) and human interpretation (Chapter 4). These arcs serve as a computable proxy for plot, enabling large-scale analysis of causal architecture while remaining grounded in narratological theory.

The focus on thematic arcs is theoretically motivated and practically necessary. Unlike sentiment-based approaches, which conflate emotional tone with causal structure, topic transitions directly reflect shifts in *what happens*, the core of *histoire*.

This part follows a structured progression:

1. Chapter 3 introduces two systems (as well as an extension of the first system) for detecting single and multiple arc boundaries.
2. Chapter 4 develops annotation guidelines for human-identified thematic arcs, with inter-annotator analysis.
3. Chapter 4 also compares the system annotated data with human annotated data to evaluate the feasibility of this computational approach.

By producing a data set and computational method for story-level arcs, Part I lays the foundation for Part II, where discourse-level units (scenes, levels, anachrony) are annotated and automatically detected. Together, these parts construct a two-tiered, modular framework for computational narratology: from *thematic causal arcs* to *structural discourse segments*.

Chapter 3

Automatic Narrative Arc Identification

Chapter 2 found that plot is mostly modelled computationally in the form of emotional arcs. Yet Chapter 2 also established that plot fundamentally concerns the arranged sequence of events, often organised into discernible stages or segments (Genette et al., 1980; Reiter, 2015). This chapter argues that whereas emotional arcs capture the affective journey, they are secondary to the structural arrangement of events and less reliable for detecting plot boundaries. Thematic shifts, by contrast, are inherent to the text and directly signal changes in key story elements such as characters, places, actions or themes that advance the narrative and represent its segments (Reiter, 2015).

*This chapter focuses on identifying the high-level structure of a text in the form of thematic narrative arcs. The work is adapted from two articles, namely, Heyns, N., & van Zaanen, M. (2021). *Finding topic boundaries in literary text*. Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021. https://dh2021.digitalhumanities.org.za/wp-content/uploads/2021/11/DHASA_book_of_abstracts.pdf and Heyns, N., & van Zaanen, M. (2022). *Detecting multiple transitions in literary text*. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 3375–3381. <https://aclanthology.org/2022.lrec-1.360.pdf>*

*This chapter aims to partially answer **RQ 6**: “Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?”. Chapter 8 will also address this question by focussing on a different purpose-built system.*

To detect thematic narrative arcs computationally, a method is required that:

- *Captures dominant themes in text segments.*
- *Identifies discontinuities between segments.*
- *Operates on unannotated text.*

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) satisfies all three. It treats each text segment as a mixture of latent topics and assigns a dominant topic per segment. A thematic shift then

appears as a change in dominant topic (a boundary signal). This approach transforms a high-level narratological concept into a manageable NLP task: topic transition detection.

To evaluate whether the approach proposed in this chapter can serve as a strategic entry point into under-researched high-level structure, three interconnected sub-questions are proposed:

- 1. Can a baseline LDA-based system reliably detect a single thematic transition in concatenated literary texts, outperforming a random baseline? (Section 3.3.1; Heyns & van Zaanen, 2021)*
- 2. Can this single-transition system be recursively extended to identify multiple thematic boundaries, and what are the limitations of such an approach? (Section 3.3.2; Heyns & van Zaanen, 2022)*
- 3. Can a purpose-built, non-recursive system designed from scratch for multiple transitions outperform the recursive extension in both accuracy and robustness across diverse texts? (Section 3.3.2; Heyns & van Zaanen, 2022)*

3.1 Introduction

A narrative arc represents the trajectory of a story's development, marked by transitions that segment the text into distinct parts, each characterised by changes in focus, perspective, or theme. These transitions serve as boundaries, separating segments of the text with differing properties, such as a shift from one setting to another or the introduction of a new storyline. By detecting these boundaries, we can map the high-level structure of a narrative, providing a framework to compare narrative storylines across authors, genres, or styles.

To identify narrative arcs, the focus is on the transitions that occur in the storylines throughout a literary text. Transitions in the storyline can be seen as boundaries, separating the text into parts of the text that have different properties. How these parts differ depends on the type of transition. However, because the text before the transition and that after will be different in some respect(s), we may assume that such transitions can automatically be identified on the basis of differences between properties of the part of the text before and after the transition. This chapter explores the development of automated systems to detect such transitions, focusing specifically on thematic narrative arcs, which this chapter defines as sequences of thematic shifts that reflect changes in the direction or focus of the narrative.

The literature review in Chapter 2 showed that prior research has predominantly identified narrative arcs as emotional arcs, tracking the rise and fall of sentiment or character emotions across a text, inspired by Vonnegut's concept of storytelling "shapes" (Vonnegut, 2005). Studies such as Deng et al. (2022), Jockers (2015), Sui et al. (2023), and Woods (2022) have used sentiment analysis to map these emotional trajectories, often tying them

to plot progression or character development. While emotional arcs effectively capture the affective journey of a story, they are less suited to detecting structural shifts in the narrative. Emotional fluctuations may occur without a corresponding change in the plot or thematic focus. For example, a character's mood may shift subtly while the storyline remains constant. Conversely, a major plot twist or change in setting may introduce a new theme without altering the emotional tone. Relying solely on sentiment risks conflating affective changes with structural boundaries, limits the ability to map the broader architecture of the narrative.

To address these gaps, this chapter proposes a topic-based system, where latent dirichlet allocation (LDA) (Blei et al., 2003) is used to identify narrative boundaries through thematic shifts. Thematic arcs, defined by coherent changes in topics such as events, settings or character interactions, align more closely with the structural properties of a narrative. For instance, a shift from a theme of "investigation" to "confrontation" may signal a new narrative arc, even if the emotional tone remains consistent. By focusing on thematic arcs, our approach captures the transitions that define the progression of a story, offering a clearer lens for analysing narrative structure.

This chapter chooses to focus exclusively on thematic arcs for two reasons. First, emotional arcs have been extensively studied, with robust methods already established (Deng et al., 2022; Jockers, 2015; Sui et al., 2023). Second, this thesis argues that thematic arcs are more fundamental to the concept of a narrative arc, as they directly reflect shifts in the core ideas and direction of the story, which are central to its structural identity. That said, thematic and emotional arcs are complementary. Emotional arcs add depth by revealing how characters and readers respond to thematic changes, often amplifying the impact of structural shifts.

This chapter develops two topic-based systems for detecting thematic narrative arcs. Section 3.3.1 introduces the first system (based on Heyns & van Zaanen, 2021) to detect a single thematic transition in a text using LDA.

Section 3.3.2 builds on Heyns and van Zaanen (2021) by first refining the single-transition identification (STI) system from Heyns and van Zaanen (2021) to detect multiple transitions and introduce a multiple transition identification (MTI) system using a collation-based clustering of topic sequences. This progression addresses the core limitation of Section 3.3.1, namely that real literary texts contain multiple thematic shifts, necessitating a more robust approach. The evaluation expands to nine text pairs, including pairs that are semantically similar, making boundary detection more challenging. Using the same evaluation methodology, the extended STI and MTI systems are compared against the random baseline. This broader evaluation tests the generalisability of results across diverse texts, addressing whether the system can reliably identify thematic arcs in varied narrative contexts.

Together, the systems form a cohesive progression from identifying single thematic transitions to handling multiple transitions in complex literary texts.

3.2 Background

Section 3.2.1 begins by looking at methods with similar goals in mind. Since the systems proposed depend heavily on the performance of LDA, Section 3.2.2 focuses on previous methods to determine the length of the documents given to LDA. Section 3.2.3 looks at different ways to evaluate the performance of LDA, mostly focusing on the limitations of evaluating LDA directly.

3.2.1 Related work

To our knowledge, only a limited amount of research has been conducted on the automatic detection of transitions in literature. Previous research by Grosz and Sidner (1986) and Hirschberg and Litman (1993) focused on specific properties in the linguistic signal. This is illustrated by the identification of transitions in the area of spoken dialogue. In particular, characteristic features of transitions that can be found in the language signal are used, for example, phrases used to signal a topic change, significant pauses in the speech signal, changes in intonation or domain-specific cue phrases. A similar approach has been applied to textual data, where the text structure, for example, headings, chapter divisions and paragraphs, is used to detect transitions.

However, these techniques that rely on specific aspects of the linguistic signal, are specialised and difficult to use for longer, unstructured texts, and can be expected to lead to low-quality results in situations in which the structure of the text does not directly follow the layout (for example, when a topic crosses multiple paragraphs). In literary text in which, for instance, we may be interested in the structure that shows a non-linear storyline, the relationship between properties such as paragraph transitions and storyline transitions becomes unclear.

Reynar (1994) proposed a method based on lexical cohesion and a graphical technique called dot-plotting to identify transitions in text. Dot-plotting was first proposed by Church (1993) as a simple visualisation technique for aligning parallel bilingual corpora. In this approach, two texts are represented as sequences of characters or words, and a 2D grid is created in which the x -axis corresponds to positions in the first text and the y -axis to the second. Matches between elements at positions i and j (for example, identical words) are marked as points (i, j) , revealing diagonal bands of density that indicate aligned segments. Reynar (1994) adjusted the dot-plotting method for monolingual texts to automatically detect topic boundaries on the basis of lexical cohesion (the tendency of related words to cluster in the same discourse segment). Reynar (1994) treated a single text as a concatenated sequence of word positions. For each repeated lemma (base word form) appearing at distinct positions x and y , he plotted four symmetric points on a square grid with both axes representing positions in the same text, in other words, the area indicated by (x, x) , (x, y) , (y, x) , and (y, y) is plotted. This Cartesian product captures self-matches along the main diagonal (indicating within-segment repetition) and cross-

matches off-diagonal (linking repeated words across positions). The repetition of lexical items occurs more frequently in regions of a text discussing the same topic. The density of the areas outside a region is calculated and the boundary is identified at the lowest density point. Choi et al. (2001) extended and improved upon this method by introducing LDA as a classification method in the popular C99 algorithm. While Reynar (1994) relied on sparse, high-dimensional dot-plots of lexical repetitions, prone to noise in long texts and sensitive to preprocessing choices (for example, stopword removal, lemmatisation), C99 transforms the text into a sequence of fixed-size sentence blocks and computes a cosine similarity matrix between all pairs of blocks on the basis of term frequency vectors. This similarity matrix acts as a denser, smoother version of the dot-plot method proposed by Reynar (1994), capturing topical cohesion using word frequencies instead of exact word repetitions.

Aurnhammer et al. (2019) performed a comparison between a close reading approach, which relied on manually annotated Reddit posts and a distant reading approach which relied on the identification of topics using LDA. Here the texts were already separated (as they were individual posts), but this work showed that there is a relationship between manually annotated texts and LDA texts. Similarly, for instance, Huang and Huan (2013) and Zhou et al. (2016) generated storylines given a collection of news articles. Gupta et al. (2009) aimed to visualise storylines (from video data), but relied on weakly labelled data.

3.2.2 Length of LDA documents

Sbalchiero and Eder (2020) systematically investigated how document length affects LDA model fitting in literary texts. In their study, they examined the performance of LDA on literature texts by splitting the text using six different sample sizes (500, 1 000, 5 000, 10 000, 20 000, and 50 000 words). They found that the optimal number of topics is inversely related to chunk length: shorter chunks support more topics, while longer chunks require fewer. Intuitively, an extremely short text chunk and a many topics will provide many very specific topics, which causes the model to overfit. Reversely, an extremely large text chunk combined with too few topics will result in very broad, general topics, causing the model to underfit. The optimal number of topics and the size of the text chunk are required to be in equilibrium, based on this statement. While fewer topics are needed for larger chunks to avoid overfitting, reducing them excessively risks underfitting. Sbalchiero and Eder (2020) concluded that the best number of topics for different sizes of the samples should be evaluated using, for example, the elbow method suggested by Kodinariya and Makwana (2013). According to Sbalchiero and Eder (2020), previous studies have already demonstrated that LDA performs well when applied to short texts, but there is a lack of empirical evidence to show that LDA also performs well on longer texts. In contrast, Syed and Spruit (2017) argued that longer texts are less affected by noise in the topic-word distributions, often producing more coherent and stable topics due to richer co-occurrence statistics. In practice, several studies recommend moderate document sizes for interpretable results. Jockers and Mimno (2013) indicated that the

ideal size of the sample texts should be large enough to allow for the proper measurement of word co-occurrences, but small enough that they can reasonably be assumed to contain a small number of themes. They found that applying LDA to full texts typically resulted in vague topics. However, splitting texts into approximately 1 000-word samples, breaking at the nearest sentence boundary, resulted in more interpretable topics. Similarly, studies like Syed and Spruit (2017), Blei et al. (2003) and others have used abstracts or short excerpts as proxies for coherent documents, reinforcing the consensus that controlled document length is critical for high-quality topic modelling in narrative texts.

3.2.3 Evaluation of LDA

LDA models can be evaluated using either extrinsic or intrinsic methods. Extrinsic evaluation methods measure the performance of LDA models on a secondary task, such as document classification or information retrieval (Wallach et al., 2009). Intrinsic methods include measurements that help to distinguish between topics that are semantically interpretable and topics that are artefacts of statistical inference. Usually, an intrinsic method relies on the estimation of the probability of an unseen held-out data set given the trained model (Wallach et al., 2009). Popular intrinsic methods are log-likelihood and perplexity measures, as well as topic coherence.

The log-likelihood approach measures how well an LDA model fits the data. The probability of a held-out data set, not used during training can be estimated in several ways, such as with importance-sampling methods, harmonic mean, annealed importance sampling, a Chib-style estimator, or a left-to-right evaluation algorithm (Wallach et al., 2009). Perplexity is a normalised form of log-likelihood, computed as the exponentiated average negative log-likelihood of a held-out test set. Perplexity thus describes how well an LDA model predicts a topic for a sample by computing the normalised log-likelihood of a held-out test set. A model will be considered good when it has a high log-likelihood and, hence, a low perplexity score. Chang et al. (2009) have, however, shown that the log-likelihood and perplexity scores have poor correlation to human judgement and are sometimes even slightly anti-correlated.

Another approach deals with the top words found in a topic. For each word, a vector representation of its contextual distribution can be created on the basis of occurrences in large amounts of text. Based on these vectors, the LDA topics can be evaluated by measuring the cosine distance between the words that describe the topic. If words from the same topic are closer together, the coherence is considered high. The underlining idea behind topic coherence is the distributional hypothesis of linguistics. The distributional hypothesis states that words with similar meaning tend to occur in similar contexts (Harris, 1954).

Note that, ideally, human topic rankings should be available, to compare the LDA topic coherence scores to the human topic rankings. Unfortunately, in most cases topics

identified by humans are not available and researchers have to rely on some automatically computed coherence score alone.

3.3 Methodology

This section start by explaining the Single Transition Identification (STI) system, which detects a single thematic boundary. The data used for this system is detailed in Section 3.3.1, along with the experimental settings. The evaluation methodology is also presented in Section 3.3.1, followed by the results.

The analysis is then extended to multiple thematic transitions using two complementary systems in Section 3.3.2. First, an extension of the STI system is introduced under *Improvements over STI*, which applies the single-transition logic iteratively to detect multiple boundaries. Second, a novel MTI (Multiple Transition Identification) system is proposed, using collation-based clustering of topic sequences to handle complex, multi-segment narratives. The evaluation data sets for both systems covering single-transition (ST) and multi-transition (MT) configurations are described under *Data sets*. Experimental settings, evaluation methodology, and results for both the recursive STI and MTI systems are also provided in Section 3.3.2.

The chapter concludes with a comparative discussion of the STI and MTI systems, highlighting their strengths, limitations, and implications for computational detection of narrative arcs.

3.3.1 System 1: Single boundaries

The single transition identification system (this system will be referred to as the STI system from now on) proposed in this chapter consists of four steps. First, it takes the input text T and subdivides the text into n snippets: $S = \langle s_1, \dots, s_n \rangle$, where $T = s_1 \oplus s_2 \oplus \dots \oplus s_n$ with $\oplus$ the concatenation operator. Second, this sequence of snippets (S) is given to the LDA system, which essentially provides a mapping LDA , which results in a sequence of LDA topics: $LDA(S) = \langle LDA(s_1), LDA(s_2), \dots, LDA(s_n) \rangle$. Third, potential transitions are identified. Each position between two snippets, (s_x, s_{x+1}) in the sequence (with $x = 1 \dots n - 1$), is considered. For each of these positions, the size of the intersection of the set of LDA topics before this position and the set of LDA topics after the position is computed. The minimum value of all these intersections indicates the best potential transition. There may be several positions that have the same minimum intersection sizes: $\arg \min_{x=1}^n |\bigcup_{i=1}^x LDA(s_i) \cap \bigcup_{j=x+1}^n LDA(s_j)|$. Finally, the system selects one of the potential transitions. If there are multiple potential transitions, it selects one at random.

The STI system is compared to two other systems: a baseline and an oracle system. The baseline system does not use any LDA information, but selects a transition at random

from all possible positions between the snippets. This system serves as a lower limit. In contrast, the oracle system follows the regular transition identification system with one difference: when multiple potential transitions are identified, this system selects the best of these potential transitions. In other words, it makes use of information on where the real transition can be found. This method serves as an upper limit.

Data

In order to properly evaluate the performance of the STI system, the system needs to be applied to a text in which the transition is known. The output of the system can then be compared against the true transitions in the text. However, manually identifying text transitions is a subjective task. Texts often have a hierarchical structure (Grosz & Sidner, 1986), in which text parts can consist of multiple (sub-)text parts, as opposed to a simple linear structure (Skorochod'ko, 1971). This means that textual transitions can take place on multiple levels in the hierarchy.

A synthetic evaluation approach is suggested, wherein a text is created by concatenating two source texts known to discuss different topics. For this purpose, two books were used as source texts: *Utilitarianism* (Mill, 1861) and *Hide and Seek* (Collins, 1854). Straightforward pre-processing was applied to these texts: stopwords were removed using NLTK¹, as these words occur so frequently that they do not help in identifying LDA topics of the snippets (but they do have an impact on the size of the snippets). In addition, the text was lower cased and lemmatised, and punctuation was removed using spaCy². All snippets were extracted from the pre-processed text, ensuring that snippet boundaries and content reflected the cleaned representation used by the LDA model.

From these two books, the first 25 snippets of 500 words each was selected, resulting in a list of 50 snippets in total with the known transition after 25 snippets. Table 3.1 shows a sample from both of the source texts.

Experimental settings

As was mentioned before, the STI system relies on LDA to identify topics for each of the snippets. LDA has a parameter that sets the number of topics that LDA is allowed to assign to the snippets. As this is a manually assigned parameter, it can be varied in the experiments. In the experiments described in this chapter the number of LDA topics was varied from two to 30 in steps of two. For each of the number of LDA topics, the system was run 100 times (as LDA may lead to slightly different results due to a random factor.) The median, average, and standard deviation results for each of these settings are provided.

¹<https://www.nltk.org/>

²<https://spacy.io/>

Table 3.1: Sample from each of the two source texts.

Source text	Sentence:
Mill (1861)	desire different thing desire happiness love music desire health They included happiness They elements desire happiness made Happiness abstract idea concrete whole parts And utilitarian standard sanctions approves Life would poor thing ill provided sources happiness provision nature things originally indifferent conducive otherwise associated satisfaction primitive desires. . .
Collins (1854)	ruddy face suddenly turned pale left circus determined find really going behind red curtain He walked round outside building wasting time found door apply admission At last came sort passage tattered horse-cloths hanging outer entrance You can't come said shabby lad suddenly appearing inside shirt sleeves Mr. Blyth took half-a-crown I want see deaf dumb child directly Oh right go muttered lad pocketing money greedily Valentine hastily entered passage As soon inside sound reached ears heart sickened turned faint No words describe horror helplessness moan pain dumb human creature. . .

Evaluation

To measure how well the different systems performed, an evaluation metric needs to be decided on. Standard evaluation metrics used in classification tasks, for example, precision, recall, and F-score, are not directly suitable when trying to evaluate this particular problem, even though the metrics could be used. True positives can be defined as when the system identified a transition that was the same as the real transition. In the same line, if the identified transition is not the same as the true transition, it is a false positive. False negatives occur when a true transition is not identified by the system. True negatives are cases in which the system correctly identified a non-transition. From these, the standard metrics can be calculated. However, the problem with these evaluation metrics is that they do not take distance into account. It is better for the system to propose a transition close to the real transition than to propose one that is far away.

Reynar (1994) proposed a metric in which a window of three sentences was considered, which allowed for a bit of leniency with respect to the transition location. However, this window length was arbitrary and still ranked a terrible system on par with a better system that proposes transitions just outside the window. Alternative measures have been proposed by Beeferman et al. (1999), Pevzner and Hearst (2002) and Georgescu et al. (2008), each building upon one another. The advantages and shortcomings of each are discussed in detail in Purver et al. (2011).

The objective is to find a transition that is as close as possible to the real transition (which is known due to the concatenation of two different texts). In other words, an evaluation

metric is required that takes into account the distance between the proposed and real transition. For this, the root mean squared error is used, which is defined as follows:

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

where n is the number of runs, p_i is the position of the proposed transition position (which can range from 1 to 49) in run i (which ranges from 1 to 100, as the system is run 100 times due to the random factor of LDA and the random selection in case of multiple possible transitions) and r is the position of the real transition (at position 25). Lower RMSE values indicate better performance, as they reflect smaller average deviations from the true boundary. The scikit-learn Python package³ was used to calculate the RMSE. Note that this approach does not directly evaluate the performance of LDA, but instead focuses on how well the overall system identifies the boundaries. In other words, an extrinsic evaluation is performed.

Results

To investigate the performance of the transition identification system, the RMSE results of the system as well as the random baseline and oracle system are provided in Table 3.2. This table also shows this information for each of the settings for the number of LDA topics.

These results show that the RMSE of the baseline is around 15. Note that the baseline always selects a random position for the transition, which may range from 1 to 49, with the real transition at position 25.

The STI system performs perfectly with two LDA topics as can be seen by the RMSE of 0.0 and a standard deviation of 0.0 (remember, lower values of RMSE are better as they relate to the distance of the position of the proposed transition compared to the position of the real transition). In each run, exactly the correct position for the transition is identified. Effectively, LDA identifies that there are two main topics that can be identified in the complete text and these correspond to the two original texts that were concatenated.

The performance of the STI system gradually deteriorates when more LDA topics are made available. When four LDA topics are available, the performance is still quite good, with an RMSE of 0.957 but the standard deviation is already 2.941, which indicates that if a wrong transition is identified, it may be relatively far away from the real position.

Increasing the number of LDA topics generally decreases the performance. Overall, the mean RMSE becomes larger, indicating that more often incorrect positions for the transition are proposed. The standard deviation also becomes relatively large, which again indicates the spread of proposed transitions. Note that the median also becomes larger which emphasises the larger spread. The slight improvement of the system at 22 LDA topics is

³<https://scikit-learn.org>

Table 3.2: RMSE results of each system for the range of LDA topics. Note that the baseline does not rely on LDA and hence has no # LDA topics provided.

system	# topics	median	mean	sd
Baseline		15.0	14.603	1.067
STI	2	0.0	0.000	0.000
STI	4	0.0	0.957	2.941
STI	6	0.0	2.943	5.357
STI	8	0.0	5.814	7.437
STI	10	6.0	7.229	7.390
STI	12	6.5	9.543	8.576
STI	14	6.0	8.371	9.305
STI	16	9.0	10.057	8.485
STI	18	9.0	10.671	8.759
STI	20	12.0	12.914	8.165
STI	22	9.5	12.114	9.454
STI	24	10.5	12.086	7.910
STI	26	10.5	11.014	8.893
STI	28	10.0	12.886	8.596
STI	30	10.0	12.129	9.119
Oracle	2	0.0	0.000	0.000
Oracle	4	0.0	0.471	2.263
Oracle	6	0.0	0.843	3.242
Oracle	8	0.0	1.843	4.652
Oracle	10	0.0	3.429	6.788
Oracle	12	0.0	3.000	6.347
Oracle	14	0.0	2.514	6.611
Oracle	16	0.0	1.243	4.206
Oracle	18	0.0	1.543	4.989
Oracle	20	0.0	2.429	6.135
Oracle	22	0.0	1.700	5.176
Oracle	24	0.0	0.714	2.649
Oracle	26	0.0	1.714	5.491
Oracle	28	0.0	1.229	4.304
Oracle	30	0.0	0.914	3.202

probably due to the random factors of LDA and the selection of the proper transition. The standard deviation is relatively large, so it is unlikely to be a real improvement.

The performance of the oracle system demonstrates that it, like the STI system, performs well with a low number of available LDA topics. The fact that the STI system already performed perfectly with two LDA topics means that the oracle system cannot improve as the STI system already always selects the best position for the transition. However, with four available LDA topics, sometimes the oracle system leads to runs that do not contain the correct transition. Here the impact of the random factor of LDA can be seen as the oracle system always selects the best transition position. Incorrect possible transitions are also found, which leads to a lower score for the STI system with four LDA topics. The performance of the oracle system also deteriorates with larger number of LDA topics, which means that the correct transition cannot be found in any of the proposed transitions according to the sequence of LDA topics. However, the median remains at 0.0, indicating that often the correct transition is proposed. There are runs in which the random factor of LDA leads to sets of possible transitions where the correct transition is not proposed. In other words, the lower performance of the oracle system with larger number of LDA topics is the result of the random factor in the LDA system, whereas the difference between the oracle system and the STI system can be attributed to the random selection of transitions when multiple possible transitions are identified.

To support the idea that the number of potential transitions increases with the number of LDA topics, the correlation between the number of LDA topics and the number of possible boundaries can be looked at. Computing Pearson's product-moment correlation results in a moderate, significant ($p < .0001$) correlation between the number of LDA topics and the number of possible boundaries with $r = .698$ (an $r > .70$ is considered a strong correlation). Figure 3.1 shows the relationship between the number of topics and the number of possible boundaries the system identifies. The x -axis of the graph shows the number of LDA topics and the y -axis the number of potential transitions identified by the system. Note that the transparency of the points in the graph indicates how frequently that situation occurs, with darker points having a higher frequency. Increasing the number of LDA topics indeed increases the number of possible positions for the transitions. This results in situations in which the STI system has a harder time as there are more possible transitions to choose from. The line is computed using the local polynomial regression fitting and the shaded area indicates the 95% confidence interval.

3.3.2 System 2: Multiple boundaries

Improvements over STI

In Section 3.3.1, the original STI system proposed in Heyns and van Zaanen (2021) identifies a boundary describing a topic transition in the text.

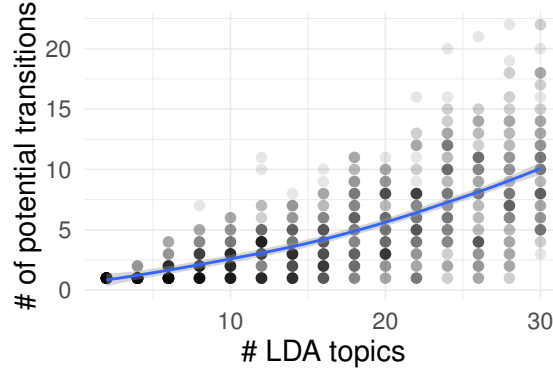


Figure 3.1: The relationship between the number of LDA topics and the number of possible transitions proposed by the system. Darker points indicate higher frequency of that situation. The line indicates the local polynomial regression fitting and the shaded area around the line represents the 95% confidence interval.

A proposed improvement to the STI system enables the identification of multiple transitions in a text. This improvement is based on Heyns and van Zaanen (2022) and will be referred to as the recursive single transition identification (RSTI) system. It starts the same way: the text is divided into snippets, LDA assigns a topic to each, and a sequence of topics is produced. But now, instead of using intersection, the system computes entropy at each possible boundary position. The position with the lowest entropy is chosen as the first transition. Once found, the text is split into two parts at this point. This process is then repeated independently on each half to find the next transition. This recursion continues, allowing the system to identify more than one boundary. If multiple positions tie for minimum entropy, one is selected at random. This approach considers topics on both sides of the boundary and therefore better captures narrative shifts compared to the original STI system described in Section 3.3.1.

Although the RSTI system successfully extends the original STI approach to detect multiple boundaries through recursive splitting, it inherits limitations from the STI system’s design. The original STI system was intended for a single transition, and recursive application may not optimally capture complex, non-linear narrative structures in real literary texts.

An alternative system is therefore proposed, the multiple transition identification (MTI) system, which was developed with the intent of identifying multiple transitions in a text. The MTI system start with the same input as the RSTI system, namely a text consisting of a sequence of snippets (S), which is handed to the LDA system, again resulting in a sequence of LDA topics: $LDA(S) = \langle LDA(s_1), LDA(s_2), \dots, LDA(s_n) \rangle$. A collation algorithm is then used to search for the boundary using the following steps:

1. Number snippets according to their position in the text.
2. For each of the LDA topics, identify all snippets (represented by their number) that have that LDA topic assigned to them.

3. For each LDA topic, identify all sequences of consecutive snippet numbers. Calculate the length of all these consecutive snippet numbers and add the lengths, which forms the value for that LDA class. Note that a snippet with number x forms its own group (which then has value zero) if no snippet with number $x - 1$ and $x + 1$ can be found with the same LDA topic.
4. Identify the two LDA topics with the highest value and call these list A and list B . If two LDA topics have the same value, pick the topic with the most snippets. If more than one topic also have the same number of elements, pick the first topic.

Steps 1–4 are demonstrated in Table 3.3.

Table 3.3: For each LDA topic, corresponding snippet numbers are ordered. Groups of consecutive snippet numbers are identified, indicated by square brackets. The value belonging to an LDA topic is the sum of the lengths of all groups of consecutive snippets in a topic. The LDA topics with the largest values (which will be list A and B) are indicated in bold in the value column.

LDA	Snippet numbers	Value
1	[13, 14, 15], [17, 18], [22, 23, 24, 25], [33, 34, 35, 36, 37]	$A = 14$
2	[1, 2, 3], 39, [44, 45, 46, 47, 48]	8
3	27, 50	0
4	49	0
5	6, 19, 21, 26, [28, 29, 30, 31, 32]	5
6	[4, 5, 6, 7, 8, 9, 10, 11, 12], 38, [40, 41, 42, 43]	$B = 13$
7	16, 20	0

Table 3.4: Calculate the minimum distance for the start and end snippets (indicated using open and close square brackets) of each group to the closest snippet value in list A and list B . Here this is illustrated for LDA topic two.

Group	[1	3]	39	[44	48]	Sum
A	12	10	2	7	11	42
B	3	1	1	1	5	11

Table 3.5: Lists A and B after the snippets of LDA topic two are added to the list with the minimum total values, i.e., list B .

List	Snippet numbers
A	13, 14, 15, 17, 18, 22, 23, 24, 25, 33, 34, 35, 36, 37
B	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48

5. Identify the LDA topic X with the next highest group value.
6. For the snippets that can be found at the start or end of a group in topic X or are not present in a group, identify the snippet that is closest to the snippet under

Table 3.6: The final versions of list *A* and list *B*. All possible boundary positions are printed in bold.

List	Snippet numbers
<i>A</i>	13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37
<i>B</i>	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

consideration in both lists *A* and *B*. Store the minimum distances for each of these snippets. Add the distances. Table 3.4 illustrates this process for LDA topic 2. For example, one is the beginning of a group from LDA topic 2. The closet snippet for this in list *A* is snippet 13, so the distance is 12. For list *B*, this is snippet 4, so the distance for list *B* is 3.

7. Add the snippet numbers of the LDA topic under consideration to the list with the smallest sum of distances. Table 3.5 shows the altered list *A* and *B*.
8. Steps 5 to 7 are repeated for all the remaining LDA topics. If an LDA topic has the same sum of minimum distances to list *A* and list *B*, the length of the group in list *A* and list *B* that follows the start and end of each group in the LDA topic is used to calculate the minimum distance instead. If the minimum distance to list *A* and *B* is still the same, the segment is added to list *A*.
9. After all the LDA topics are added to either list *A* or *B*, the outliers in both lists are identified and removed. An outlier is defined as a single snippet number x for which $x - 1$ and $x + 1$ are not found in the same list. The outliers are added to an outlier list.
10. For every outlier in the outlier list, it is checked whether it can be added to a group in either list *A* or *B* and the outlier is added to the appropriate list. Table 3.6 shows the final division of the snippets in lists *A* and *B*.
11. The last step is to identify the start and end of each group in list *A* and *B*. One group will always start with snippet number 1 and one group will always end with snippet number 50. These values can be deleted from the possible boundaries as a boundary cannot occur in these positions. The remaining list of possible boundaries will then contain groups of 2 consecutive numbers. The boundaries will be between the consecutive numbers found in the different lists. In this example (see Table 3.6) the possible boundaries are between snippets number 12, 13 and between snippets number 37, 38.

Data sets

For the RSTI and MTI systems evaluation data is also created by concatenating text snippets from different texts following the same steps as in Section 3.3.1. This way, control is maintained over the position of the transitions. The data from Section 3.3.1 cannot

be used as there are now multiple transitions in a text instead of a single transition. Our first data set (called *ST* for single transition) contains concatenated texts with only one boundary (which means that the text is created by concatenating snippets from *A* and *B* such that it follows the sequence *AB*). The second data set (which is called *MT* for multiple transitions) contains texts with multiple boundaries and these are created by concatenating snippets in the format *ABA*.

In contrast to the evaluation in Section 3.3.1 of Heyns and van Zaanen (2021), which only used one pair of texts, here nine pairs of texts are used (extracted from a total of 18 books). In particular, the following text pairs from Project Gutenberg⁴ is used:

1. Utilitarianism (Mill, 1861), and Hide and Seek (Collins, 1854),
2. Crime and Punishment (Dostoevsky, 1866), and Great Expectations (Dickens, 1861),
3. Eureka (Poe, 1848), and A Study in Scarlet (Doyle, 1887),
4. The Count of Monte Cristo (Dumas, 1846), and Our Mutual Friend (Dickens, 1865)
5. Middlemarch (Eliot, 1871), and Jude the Obscure (Hardy, 1895)
6. Through the Looking Glass (Carroll, 1871), and Anne of Green Gables (Montgomery, 1908)
7. Jane Eyre (Brontë, 1847) and Little Dorrit (Dickens, 1855)
8. The Moonstone (Collins, 1868), and Frankenstein (Shelley, 1818)
9. Barchester Towers (Trollope, 1857), and Cranford (Gaskell, 1853)

For the *ST* data set, 25 snippets of 500 words each are selected from both texts *A* and *B*, and concatenated. This is the same as in Section 3.3.1, but now it is done for each of the nine pairs of texts. For the *MT* data set (which contains *ABA* created texts), 12 snippets from text *A* are used for the first part and 13 snippets from text *A* for the last part. The middle part, *B*, still consists of 25 snippets from text *B*. This means that all concatenated texts in both data sets consist of 50 snippets of 500 words each, where for the *ST* texts the true transition occurs after 25 snippets and for the *MT* texts, the transitions occur after snippet 12 and after snippet 37. All concatenated texts contain 25 000 words in total.

Before creating the snippets, the same straightforward pre-processing is applied to the texts as was done in Section 3.3.1.

Experimental Settings

To evaluate the performance of the MTI system against the RSTI system, an experiment is first performed that requires the identification of one transition in texts. Note that the

⁴<https://www.gutenberg.org/>

extended RSTI system behaves like the STI system proposed by Heyns and van Zaanen (2021) in Section 3.3.1, when identifying only one transition, although this decides on a transition using entropy. Section 3.3.1 showed that the STI system performs well on the data. However, the STI system so far has only been tested on one text. Here the evaluation is extended to multiple texts (as described in Section 3.3.2). The results are shown under the results subsection of Section 3.3.2. Additionally, the performance of the MTI system is compared against that of the RSTI system when assigning one transition.

The second experiment focuses on the identification of multiple transitions. Both RSTI and MTI systems will be applied to the MT data set, which illustrates how well both systems can identify multiple boundaries. The MTI system is expected to clearly outperform the RSTI system as this system has been specifically designed to deal with multiple boundaries.

As seen in Section 3.3.1, the number of LDA topics assigned to the text can have a drastic influence on the performance of the system. Results from Section 3.3.1 show that in cases where only one transition is identified, the (STI) system performs well. To provide a good overview, both the RSTI and the MTI system will be evaluated with two to 30 LDA topics (in steps of two). For each number of LDA topics, each system is run 100 times as, due to the random factor inherent in LDA, the LDA topics might be slightly different in each run. The median, mean and standard deviation results for each of these settings is provided.

Evaluation

To take into account the distance between the proposed transition and the true transition, the approach of Heyns and van Zaanen (2021) is followed (in other words, the same as in Section 3.3.1), which uses the root mean square error (RMSE).

The RMSE is calculated in the same way as in Section 3.3.1. However, to calculate the RMSE for multiple transitions, the RMSE is computed for each transition and these values are combined using the average. The scikit-learn Python package⁵ is used to calculate the RMSE.

Results

The results of all the experiments are provided in Table 3.7. First consider the performance of the RSTI system on the ST data set that requires identification of a single transition. These results are comparable to the previous results as provided in Section 3.3.1 of the STI system as proposed by Heyns and van Zaanen (2021), although the use of entropy to identify boundaries has improved performance. The differences between the texts are marginal, indicating that the RSTI system is robust.

⁵<https://scikit-learn.org>

Table 3.7: RMSE results (MED: median, M: mean, SD: standard deviation) of the RSTI and MTI systems on both single (ST) and multiple (MT) transitions data sets, for the range of LDA topics.

System	# topics	ST data set			MT data set		
		MED	M	SD	MED	M	SD
RSTI	2	0.000	0.000	0.000	1.000	2.250	3.284
	4	0.000	0.000	0.000	2.000	2.875	2.900
	6	0.000	0.000	0.000	4.500	5.125	4.291
	8	0.000	0.000	0.000	5.000	5.250	2.435
	10	0.000	0.000	0.000	1.000	3.375	4.779
	12	0.000	0.000	0.000	5.000	5.000	4.408
	14	0.000	0.000	0.000	4.500	5.125	2.100
	16	0.000	0.000	0.000	6.000	6.375	3.583
	18	0.000	0.000	0.000	6.500	6.875	4.324
	20	0.500	0.000	0.000	4.000	6.125	4.643
	22	0.000	0.000	0.000	8.500	8.000	4.660
	24	0.000	0.250	0.000	8.000	8.500	3.464
	26	0.000	0.125	0.463	12.000	10.625	4.534
	28	0.000	0.375	0.354	12.000	11.875	1.853
	30	0.050	0.050	1.061	12.000	11.750	1.581
MTI	2	0.000	0.000	0.000	0.000	0.000	0.000
	4	0.000	0.000	0.000	0.000	0.000	0.691
	6	0.000	0.000	0.000	0.000	0.000	0.000
	8	0.000	0.000	0.000	0.000	0.000	0.000
	10	0.000	0.230	0.332	1.000	0.333	0.873
	12	0.000	0.000	0.000	0.000	0.000	0.000
	14	0.000	0.641	0.732	1.000	0.610	1.002
	16	0.000	0.000	0.000	0.000	0.000	0.000
	18	0.000	1.031	1.732	0.500	0.833	1.102
	20	0.000	0.750	1.500	1.200	1.050	1.321
	22	0.000	1.000	1.732	2.000	1.267	1.845
	24	0.000	1.000	1.732	2.300	1.714	1.500
	26	3.000	2.333	2.082	2.000	1.667	1.800
	28	0.500	1.000	1.414	1.650	1.200	2.004
	30	0.000	0.667	1.155	2.000	1.004	2.679

Next, the performance of both RSTI and MTI systems is compared on the ST data set. The MTI system consistently outperforms the RSTI system (with the exception when only two LDA topics are used where both systems have the same perfect performance).

When increasing the number of LDA topics the performance starts to decrease as can be seen by the increase of the mean and median of RMSE values. This holds for both RSTI and MTI systems. However, the results of the MTI system degrade more slowly.

Next, the performance of the RSTI and MTI systems on the MT data set is compared, which requires the identification of multiple boundaries. Here, the performance of the RSTI system is worse than that of the MTI system. Similarly to the results on the ST data set, the MTI system shows a perfect performance when identifying the intersections using two LDA topics. The performance starts to decrease as the number of topics increase, although the mean for each number of LDA topics is not much higher than the mean of the MTI system when a single boundary is identified.

3.4 Discussion

This chapter proposed three systems that aim to identify topic transitions in the storyline in a text by first subdividing the text into smaller snippets. In the first system (STI), this sequence of snippets is used as the input to LDA, which assigns topics to each of the snippets. Next, based on the size of the intersection of the LDA classes of the snippets “to the left” and “to the right” of each of the positions between snippets, the best potential transitions are identified. If multiple potential transitions are found, one is selected at random.

The reasoning behind using the size of the intersection of the LDA topics on both sides of the potential transition is that transitions will show a change of topics. The current system assumes that the topics at one side of the transition will not occur at the other side of the transition (or at least less frequently). This means that the approach described here relies on the performance of LDA in assigning the correct topics.

This chapter evaluated the STI system on texts that are created by concatenating snippets from two texts. One may assume that it would be easiest to identify the transition between the texts if only two LDA topics are requested. The results show that this is indeed true. However, it may be the case that the snippets from one source text already contain two or more topics. In that case, LDA may have problems assigning the right topics to the snippets. Essentially, in that case, underfitting will occur. This corresponds to the idea described by Sbalchiero and Eder (2020).

It is interesting, however, that the performance goes down if the number of LDA topics goes up. The STI system does not directly evaluate the performance of the LDA system; it only relies on the intersection of the topics. From this, the conclusion can be drawn that increasing the number of LDA topics results in the creation of topics that occur frequently

on both sides of the potential transitions, which essentially introduces noise when trying to decide on the best transition. This again shows that the system is overfitting the data, again following the results from Sbalchiero and Eder (2020).

The STI system proposed in Section 3.3.1 is extended, so that it can identify multiple transitions in Section 3.3.2, referred to as the RSTI system. A new system is also introduced specifically designed to identify multiple transitions in a text in Section 3.3.2, referred to as the MTI system.

Both systems are compared on data sets requiring single transitions as well as multiple transitions. This showed that the MTI (Multiple Transition Identification) system proposed in Section 3.3.2 consistently outperformed the STI (Single Transition Identification) system proposed in Section 3.3.1 and the RSTI system proposed in Section 3.3.2 when multiple transitions occur in the text. Both systems rely on LDA topic information assigned to snippets of the text under consideration. However, the MTI system performs a more complex, and as a result, more robust analysis of the behaviour of the LDA topics over the snippets. The variation of the assigned LDA topics to the snippets in the text has a larger impact on the performance of the RSTI system as a result. What could be seen as noise in the assignment of LDA topics has a larger effect on the RSTI system compared to the MTI system. This is emphasised by the fact that this influence becomes larger when a larger number of LDA topics is assigned.

In contrast to the evaluation of the STI system, the RSTI and MTI systems are evaluated on texts that are created by concatenating snippets from two source texts. The fact that the results are highly comparable means that (even) the RSTI system is robust with respect to different texts.

Finally, as could be expected, the task of identification of multiple boundaries in a text is more complex than that of the identification of a single transition. However, the performance of the MTI system shows that multiple transitions can be identified with a reasonable performance in particular if a low number of LDA topics is assigned.

Based on the results, the proposed MTI system can be used to identify transitions in a text. The system is relatively stable even if the system tries to assign more LDA topics than are represented in the text, the system still has reasonable performance. Currently, however, several aspects have not been evaluated yet, such as the influence of the actual texts, and the length of the snippets. It is already known (again, based on Sbalchiero & Eder, 2020) that there is a relationship between these variables.

Regarding topic selection in real use, experiments showed optimal performance at low numbers (e.g., 2), but this may vary with text length. In practice, it is recommended to do a sensitivity analysis varying topics from 2 to 10 and making a selection based on coherence scores, especially for longer texts where more topics may capture subplots without noise. The experiments also pose a limitation in that the benchmark remains small (three short whodunit stories), limiting any claims about robustness across genres or lengths. This

genre-specific focus aids initial validation but requires future expansions to diverse corpora for generalisability.

3.5 Conclusion

This chapter demonstrated that thematic narrative arcs- structural shifts in dominant textual themes- can be automatically detected using LDA-based systems, providing a content-driven alternative to the affective focus of emotional arcs. This work partially answers **RQ 6**: “Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?”.

This chapter answers yes, and provides empirical proof across three progressive systems developed to answer our sub-questions:

Can a baseline LDA-based system reliably detect a single thematic transition in concatenated literary texts, outperforming a random baseline? This question was addressed in Section 3.3.1. The original system subdivides a longer text into smaller snippets, which are the input to LDA. The system then tries to identify possible transitions by considering the size of the intersection of the LDA topics on either side of the possible transition, which may occur between each pair of snippets. The positions that show the smallest intersection are considered possible transitions and if more than one is found, the system selects one at random. This system consistently outperforms the random baseline, indicating that the information that comes from LDA is indeed useful. When more LDA topics are requested, the performance goes down, but perfect results were found when LDA was run with only two topics.

Second, can this single-transition system be recursively extended to identify multiple thematic boundaries, and what are the limitations of such an approach? Section 3.3.2 extends the single-transition system to handle multiple transitions (in other words, the RSTI system). The approach recursively splits the text at the lowest-entropy boundary and repeats the process on each half. While functional, performance drops when using more LDA topics. An oracle system, which always selects the best transition, leads to somewhat better results, but even with the oracle system, the performance drops when using more LDA topics. Sometimes the oracle system does lead to perfect results and sometimes it does not, which highlights the influence of the random factor in the LDA system and the architectural constraints of recursive patching.

Third, can a purpose-built, non-recursive system designed from scratch for multiple transitions outperform the recursive extension in both accuracy and robustness across diverse texts? Section 3.3.2 also proposed a system that is specifically designed to identify multiple transitions (MTI system). Both systems build on information provided by LDA. The systems are evaluated on multiple texts which are created by concatenating snippets from

two source texts, showing that they lead to robust results. The MTI system consistently outperforms the RSTI system for the identification of two transitions in the text. Also, the performance of the MTI system is comparable in the experiments requiring the assignment of one or two transitions. This superiority stems from MTI's collation algorithm, which clusters topic continuity in a single pass, avoiding the fragmentation and noise sensitivity of recursive splitting.

Across all systems, it was observed that increasing the number of LDA topics leads to lower results. This means that with higher numbers of LDA topics, additional topics that do not really seem to describe proper topics are assigned to snippets in the text. This can be concluded as they influence the performance of the system because more topics can be found on both sides of the potential transitions. Essentially, this introduces more noise in the LDA topics, due to overfitting. Optimal performance occurs at low dimensionality (for example, 2 topics), where broad thematic blocks dominate.

This chapter has shown that even a simple intersection-based system outperforms random selection, therefore the approach is feasible. However, purpose-built design is essential as seen by the superior performance of the MTI system. Thus, the MTI system serves as a strategic entry point. It transforms an abstract narratological concept (thematic shift) into a computable signal, validates the approach on diverse texts, and establishes a foundation for future high-level structure detection.

This chapter lays the groundwork for Chapter 4, where the MTI system is applied to annotated text data to evaluate the systems' performance against human judgement.

3.6 Future work

With respect to future work, further investigation is needed regarding the influence of the length of the snippets used to assign the LDA topics. This particular question has not been addressed in this chapter and remains an open question. From Section 3.2.2, it is known that LDA needs texts of a particular length to get reasonable probabilities to learn the topic model, but very short snippets (for example, sentences) allow us to better identify the transitions in the text. Alternatively, paragraphs may be used as snippets, if it is assumed that no transition will occur within a paragraph. Given the robust results that are based on the LDA topics, it is feasible that the snippets may be made shorter, allowing for a more fine-grained assignment of transitions.

Concatenating snippets into a text to evaluate the performance of the systems is, of course, artificial. In particular, when the source texts are very different semantically, LDA might provide exaggerated differences. Transitions in texts identified by human judgement should lead to a more natural means to evaluate the performance. Chapter 4 will investigate this hypothesis. However, one has to keep in mind that previous work indicated low agreement between annotators on such a task.

Chapter 4

Comparing Arc Systems with Human Judgement

This chapter has been adapted from Heyns, N., & van Zaanen, M. (2024). Automatic narrative structure identification in literary texts. Digital Humanities Benelux Journal, 6, 71–90. https://journal.dhbenelux.org/wp-content/uploads/2024/11/5_Heyns_individual.pdf and Heyns, N. (2024). Post-digital narrative analysis. In The Routledge Handbook of AI and Literature. Routledge.

Chapter 3 developed three LDA-based systems to detect thematic narrative arcs. The best system, MTI, worked well on artificially concatenated literary texts. However, the conclusion was that such evaluation is inherently artificial. When the two stories are very different in semantic content, LDA may identify big topic changes that may not show true story breaks. Real texts are likely to have smoother changes. A more natural benchmark requires boundaries identified through human judgement on literary text. As noted in Chapter 3, prior research has shown low IAA on structural segmentation tasks, a challenge that must be addressed before computational systems can be meaningfully evaluated against human perception.

This chapter introduces a human-centered approach to annotating thematic narrative arcs in short stories. A set of structured annotation guidelines is designed to capture these high-level shifts in a consistent, replicable manner, bridging the gap between narratological interpretation and computational detection. These guidelines are then applied to construct a data set of annotated short stories, enabling direct comparison between human-annotated boundaries and MTI predictions. Finally, IAA is assessed across multiple annotators to validate the reliability and robustness of the guidelines.

4.1 Introduction

The ability to understand thematic narrative arcs 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 research into 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, summarisation tools, automatic essay grading and narrative popularity predictions.

Ouyang and McKeown (2014) refer to a couple of different uses of automatic thematic narrative arc 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 may 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 may present all orientating information at the start of a narrative, whereas human narrators may 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 other things.

The development of computational 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 chapter contributes by creating guidelines for annotating literature texts based on their thematic structure. The guidelines are based on existing structure theories. A data set is constructed comprising three short stories annotated following these guidelines, thus establishing a valuable resource for future scholars. In Chapter 7 we annotate a dataset consisting of 10 text where amongst other annotations, we include thematic structure annotations.

Furthermore, the importance of human annotator agreement is recognised when structured annotations are applied to literary texts. Inter-rater annotator agreement is assessed by involving multiple annotators in the annotation process, ensuring the reliability and robustness of these guidelines.

RQ 4: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?” and **RQ 5:** “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so

that unique narrative features can be captured?" are investigated in this chapter. These questions are addressed further in Part II.

The objectives also extend beyond the theoretical domain into the practical application of these guidelines. **RQ 6:** "Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?" is partially answered in this discussion. Ultimately, this research bridges the gap between human analysis and computational capabilities. The annotated data set is used to train the LDA model used to identify topics in the MTI system, introduced in Chapter 3. The performance of the MTI system is compared against human judgement. This development holds potential for literary analysis and applications in various fields. The system could also provide scholars with empirical evidence for a narrative structure theory.

4.2 Background

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 the linear order of events into account (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 for example, flashbacks or flashforwards. 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 storyline.

A long-standing 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. Section 4.2.1 looks at several well-known structure theories to demonstrate the overlap between theories. Ultimately two of these theories will be focused on: the three-act and the four-part structure theories. These theories are applied to the text to evaluate which theory suits our suggested genre, short mystery stories, better. Section 4.3.1 motivates the use of mystery stories specifically whodunits.

4.2.1 Structure theories

Structure theories aim to describe the combination 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 in the field. The most common theories are listed here:

The three-act structure: The three-act structure divides the story into three parts. This theory is often credited to Aristotle, but, the three-act structure known 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, in which the conflict is resolved. Steele (1981) notes that many commentators have noticed the Aristotelian properties of classic detective stories. This is the main reason the three-act structure theory is selected for this work, as detective stories will be annotated in this experiment.

The six-part structure: It is worth mentioning that Labov (2013) further extended the three-act structure theory to include the abstract, resolution, and coda. This creates a six-part structure consisting of the *abstract*, which 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. This structure theory will not be used in this research as short stories do not generally include an abstract and coda.

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 the 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 in their goal.

Similarly, 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 with or adjust to daily life.

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 derived from the three-act structure that is 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.

The hero's journey: Campbell (2008) bases 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. This results 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.

The seven-point structure: The seven-point story structure is a relatively new structure compared to the ones discussed above. 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 for the protagonist to *solve the problem*. A *resolution* is reached when the major conflict is resolved and the antagonist is defeated.

This chapter will work with the three-act and four-part structure as short stories are believed to follow a three-act structure (Bal & van Boheemen, 2009; Bell, 2004), whereas the mystery genre tends to follow a four-part structure (Babelyuk, 2019; Milutinovi, 2020; Stolarek, 2009).

4.2.2 Previous and related work

Research on the computational analysis of narrative arcs includes Jockers (2015), who aimed to identify the change in sentiment over a text, whereas Finlayson (2016) and Droog-Hayes et al. (2018) presented a case for inferring Propp's functions using machine-learning methods. Finlayson (2016) mentions that even though Propp described the functions in detail, there was still some ambiguity when text was annotated according to functions. They argued that the description of the function groups was 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 was to extract a timeline from the annotations. They then associated each event with an agent and patient using the semantic roles. This approach is extremely resource-intensive and requires significant human judgement.

Ouyang and McKeown (2014) introduced a system to automatically detect one of the narrative structure elements on the basis of William Labov's theory of narrative analysis

(Labov & Waletzky, 1967), as was mentioned in Section 4.2.1 (the three-act structure). They presented experiments in which they detected the complicating action element in Labov’s theory with a 71.55 f-score. They made use of a corpus consisting of 20 oral narratives collected by Labov (2013), as was mentioned in Section 4.2.1 (the six-part structure). Although this work had the same goal as ours, they relied on properties of speech data to identify boundaries. This technique would not work with text data.

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) found that summarising a text before attempting to annotate it, helped to create greater agreement between manual annotations.

A more comprehensive annotation scheme for narrative structure was proposed by Li et al. (2018), which was discussed in Section 2.4.4. This annotation scheme draws on classic theories (Freytag, 1894; Labov & Waletzky, 1997; Labov & Waletzky, 1967; Labov, 2013; Ouyang & McKeown, 2015; Prince, 1973; Todorov, 1971), and introduced ten functional labels (abstract, orientation, complicating action, MRE, minor resolution, return of MRE, resolution, aftermath, evaluation, direct comment to audience). The authors annotated 360 short personal narratives from online sources (Quora, Reddit) and re-annotated stories from Lukin et al. (2016). Texts were strictly constrained: a single story, 90 to 700 words, fewer than six dialogue lines, and non-narrative material <50%.

4.3 Methodology

Chapter 3 described the MTI method (Heyns & van Zaanen, 2022) to identify high-level textual transitions in text using LDA (Blei et al., 2003). The results of this system will be compared with human annotated texts from this chapter.

Section 4.3.1 motivates the choice of using short stories from the mystery genre, presents the three texts annotated in this chapter and provides information on the preprocessing steps. The data annotation subsection discuss the development of a data set where five human annotators annotated the texts. The agreement between human annotations will be computed. Section 4.3.2 describes the system used to analyse the text. Lastly the human annotations are compared with the system annotations in Section 4.4.

4.3.1 Data

Gius, Jannidis, et al. (2019a) argue that manually annotated data sets are imperative for developing automatic scene or narrative detection systems. This work will take inspiration from Gius, Jannidis, et al. (2019a) by establishing rules to identify sections in a text. This chapter will create a human-annotated data set that will also benefit future studies in auto-

matic scene and narrative detection. This data set will serve as a baseline for comparing the automatically identified sections using the system described in Section 3.3.2.

A case for whodunits

This subsection (based on Heyns, 2024) is arguing that whodunits are an ideal choice for computational structure analysis due to their highly structured narrative framework and the intriguing structural shifts observed in the genre over time.

The focus in CLA often leans towards asking how NLP tools and techniques can enhance the study of literature. However, Frermann et al. (2017) propose a different perspective, arguing that crime dramas can offer a valuable training ground for NLP advancement. This stems from the inherent structure of crime dramas, which often adhere to a defined narrative framework. By analysing these narratives, NLP models can be exposed to a controlled environment that approximates real-world language use and the complexities of human reasoning. Instead of treating literature merely as something to be analysed by NLP, we should also recognise the characteristics of literature that can guide the development of NLP models.

Let us distinguish between crime dramas, mystery novels, and whodunits. A crime drama is a story about the world of crime and its moral consequences (Cawelti, 2014; Malmgren, 2004). The focus is not necessarily about solving anything. On the other hand, a mystery novel is primarily driven by uncovering a hidden truth (usually a crime). A whodunit is a specific type of mystery novel where the goal is to figure out which of a closed group of suspects committed a crime where the clues are (supposedly) fairly presented to the reader (Cawelti, 2014; Malmgren, 2004). For a whodunit the main focus is on who committed the crime. Other types of mystery novels include howdunits or whydunits (Innes & Brookman, 2023), where the goal is clearly shifted.

Classic whodunit mysteries, even more so than crime dramas, exemplify the linear structure of traditional text beautifully: a crime is committed, a detective investigates, unveils the culprit, and order is restored. The whodunit, despite its focus on the unsettling crime of murder, presents a structured and predictable format that separates it from other forms of crime writing. Cawelti (2014) goes into detail describing the six main phases¹ (introduction of the detective, crime and clues, investigation, announcement of the solution, explanation of the solution and denouement) and four main roles (the victim, the criminal, the detective and those threatened by the crime but incapable of solving it). This formulaic nature

¹These six phases can be understood as a genre-specific realisation of Labov and Waletzky (1997)'s general schema (see Section 4.2.1). The primary difference is that whodunits frequently omit the initial abstract and sometimes merge the final coda into a brief epilogue. Because the short stories in the present data set rarely contain a distinct abstract or coda a simplified three-act and four-part scheme is used in this chapter. The three-act and four-part structures collapses Cawelti (2014)'s six phases into the primary units while still preserving the logical and dramatic architecture of the whodunits. In Chapter 5, once reliable macro-boundaries have been established, the annotation scheme will be refined to Cawelti (2014)'s six-phase model for finer-grained analysis.

offers a sense of security for both the writer and the reader, allowing them to explore the disturbing concept of murder within a safe and controlled environment (James, 1997).

Beyond their inherent structure, whodunits are a compelling genre for computational literary analysis (CLA) due to observed structural shifts over time. Early Golden Age whodunits (1920–1950) often resembled puzzles (James, 1997) rather than in-depth investigations, with authors crafting elaborate murder scenarios solved by detectives with near-supernatural deductive abilities (Koliassa, 2019). This emphasis on the detective's deduction skills is changing according to Cawelti (2014) who mentioned that the evolution of the classical detective genre has been to give “increasing importance to the intricacy of the puzzle surrounding the crime and less prominence to the detective's initiative in the investigation”. Now this is an interesting observation that will certainly have an impact on how the narrative is structured. Due to this shift, and the de-emphasis on the detective's extraordinary deduction skills, it can be expected to see a rise in the number and complexity of clues embedded within the narrative, allowing readers to actively participate in the investigative process. It may be that this shift is due to the addition of advanced technologies that add to the intricacies of the case, but also relieve the detective of his role as master of deduction.

Another motivational factor to use murder mysteries is the prevalence of technology in the post-digital era which has fundamentally reshaped the landscape of whodunit narratives. Traditionally, whodunit narratives relied on limitations in forensic science and communication to create puzzles for detectives to solve. However, with advancements in DNA analysis, surveillance technology, and instant communication, the element of surprise and the investigative process itself are challenged. To compensate, contemporary writers have adopted several strategies. In crime fiction, settings frequently employ tropes that limit technological capabilities, such as unreliable cellphone reception. As Gonzalez (2023) notes, these limitations serve a narrative purpose, often playing a crucial role in the development of the plot and the unveiling of evidence. Some writers depict technology as unreliable to create loopholes for the crime to occur. Others introduce criminals who themselves utilise technology to commit or conceal their crimes. Ultimately, technology's influence necessitates a re-evaluation of the classic whodunit narrative structure, demanding new approaches from writers to maintain the core elements of suspense and intrigue of the genre.

These structural shifts present computationally tractable questions for CLA, for example, to study whether modern whodunit exhibit a reduced emphasis on the detective's role, an increase in the frequency and complexity of clues, and a prominent role for technology in shaping narrative dynamics.

At this point, it is worth noting that these are very genre-specific questions. Four more general questions come to mind, given the post-digital impact on temporal and spatial structure as well as focalisation should also be mentioned. With regard to temporal and spatial structure, one can investigate if modern stories utilise flashbacks or time jumps

more frequently compared to traditional narratives. Related work (Underwood, 2018) has also focused on the difference in how much time is spent on specific scenes. Another possibility is to evaluate if the focus has shifted from closed spaces to open environments given greater connectivity. With regard to focalisation, it can be identified if there is a shift towards internal focalisation in modern stories, allowing access to the character's thoughts as they utilise technology or as was already mentioned, compare the role of technology to deduction in the detective's methods.

These questions demonstrate how CLA can quantify structural changes in a controlled narrative environment, offering insights into broader literary trends influenced by digitisation. But this technological re-structuring of the classic whodunit is not just an interesting side-note it creates exactly the kind of large-scale structural shifts that will challenge CLA to develop genre-aware computational tools. Because of these rigid patterns yet evolving structures, the personal choice of whodunit is grounded in their computational advantages rather than their literary significance alone. Their structured nature facilitates the development of algorithms to detect plot boundaries and narrative levels, as explored in this thesis's projects. While this study focuses on whodunit as an illustrative example, the methodologies developed, such as thematic arc detection in Chapter 3 and scene annotation that will be investigated in Part II, can be applied to other genres.

Data collection

Stories from the golden age (1920–1950) of whodunit are used. This era witnessed the rise of iconic detectives like Hercule Poirot and Miss Marple, providing an abundance of material for our analysis. For now, the focus is on mystery short stories, as the text needs to be annotated in its entirety. Short stories make this objective more feasible. Because manual annotation will be done, the data set will be kept relatively small. Three short stories of similar length from Agatha Christie will be used, listed in Table 4.1.

Data processing

We want to compare the manual annotations with the MTI system from Chapter 3, therefore we need to prepare the text to be used by the system.

Pre-processing is done on the text. Stopwords are removed using NLTK². The texts are lowercased, lemmatised, and the punctuation is removed using spaCy³. 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. These section breaks were removed during pre-processing, although human annotators would have had access to that visual clue.

²<https://www.nltk.org/>

³<https://spacy.io/>

To apply LDA to the text, the text is split 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 4.1 shows the short stories chosen to use as well as their length. The average segment length is 35.75 words.

Table 4.1: 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

Data annotation

Five human annotators annotated 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 to create greater agreement between manual annotations. For this study only events will be included, simplifying the nature of a literary narrative by leaving out the nuanced information that a non-event can provide in a literary analysis.

There will be two rounds of annotation. For the first round annotators divide the story into three parts by stipulating the line number where a boundary is identified. The annotator identifies the line number where a phase starts and ends. Note that each phase (except the first) starts on the next line to where the previous phase ended. For example, if phase 1 ends on line 24, phase 2 starts on line 25. They 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 are given, but the annotator divides the story into four parts.

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. 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.

Second round (four-part structure): The four-part narrative structure is another framework that many storytellers use to construct a mystery novel. The four parts 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 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 4.2 gives an example of a completed annotation form.

4.3.2 System

Chapter 3, Section 3.3.2 proposed the MTI 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 LDA topics. 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. This chapter considers the segments created by these transitions to function as a story act. The number of segments to divide the text into is specified.

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 4.2.1. For example, if a text is divided into three sections, the three-act story structure of that text is tested. For each text sequence (with $x = 1 \dots n - 1$) every position between two segments (s_x, s_{x+1}) is considered to be a potential transition point. The information gain at each potential position is computed and the position with the highest information gain is considered to be 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.

Table 4.2: 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 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.

Number of LDA topics

Section 3.3.1 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. Previously, Section 3.3.1 established that the system performs best with two LDA topics and gradually deteriorates as more topics are introduced. 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, the results of two LDA topics will be used to make comparisons.

Uglanova and Gius (2020) argue 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 the structure of a text. 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 support this idea by asserting that the topics generated through topic modelling, are more aligned with genre signals than with 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, this may not be measuring narrative structure, but rather attributes that occur around a narrative change.

As an example, words like [evening, long, night, sky, dark] that indicate time, and words like [door, house, window, world, side] indicating location can co-occur in one topic, let us call this topic one. Meanwhile, words like [hear, find, look, answered, understand, ask, remark] indicating actions and words like [inspector, police, friend, woman, doctor, father] indicating people can co-occur in another topic, let us call this topic two. These topics can be intertwined in the same section of the text. For example the text is broken into multiple segments. For the first segment topic one is the most dominant topic. For the next segment topic two is the most dominant. For the third segment topic one might, again, be the most dominant topic. However, for the next couple of segments topic one and two might no longer be the most dominant topics. They might have been succeeded by another topic, let us call this topic three. The first three segments can be grouped together as a section. So by no means can it be said a specific section is only focusing on one of these topics. The system proposed here simply finds the location where the order of these topics give the most information gain. For example for the three-act structure of *The Jewel Robbery at the Grand Metropolitan*, we identify three topics. We found that 69% of segments in the first section is topic 1, 65% of segments in section 2 are tagged as topic 2 and 57% of segments in section 3 are tagged as topic 3.

4.4 Evaluation

The transitions identified by the MTI system are compared with human-annotated transitions. This shows in how much the system is in agreement with human judgement. The automatically identified sections are compared with human-annotated sections using a distance metric. The start and end points between the automatically annotated sections and the human-annotated sections are compared.

To take into account the distance between the two transitions the RMSE is used, similar to Chapter 3. RMSE allows us to measure the distance between a predicted value and an observed value. In this case, the computer-identified transition point will be used as the predicted value and 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 1 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 package⁴ was used to calculate the RMSE.

4.4.1 Inter-rater annotator agreement

Different human annotations of the text are compared to assess the agreement between their annotations. Table 4.3 shows that human annotators reach a greater agreement with the three-act structure theory than with 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 whether the difference in agreement is significant, a random baseline was added where the text is divided into three and four equal parts. The RMSE is then compared between the baseline of each theory and the annotator's transition placement. Table 4.4 shows the average RMSE of all annotators compared to the baseline. Here there is no real difference in the average annotation standards when compared to a baseline. Although if each annotator is looked at individually in Table 4.5, a bigger difference in agreement between annotators is seen 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*.

⁴<https://scikit-learn.org>

Table 4.3: 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

Results from Table 4.3 also show that for both structure theories, the annotators struggled most with *The Jewel Robbery at the Grand Metropolitan*. 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 a 0.19 RMSE value for the four-part structure.

Table 4.4: 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

4.4.2 Human annotator versus computer annotator

Table 4.6 shows the average RMSE between the human annotators and the computer annotator. Here a similar trend is seen 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 may 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, the first text, shows that the RMSE value is the lowest for both structure theories, which also agrees with the inter-rater annotator agreement results.

Table 4.5: 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 indicate how close their annotation compares to a random baseline.

Text	Annotator				
	1	2	3	4	5
<i>Three-act structure</i>					
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
<i>Four-part structure</i>					
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

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 4.6: The average RMSE per structure theory in which 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

4.4.3 Comparing structure theories

Comparing 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.

There are specific sections that are easier to identify than others. It was expected that the first section will 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, the variation in later sections may be expected to be more. However, this is only the case in two of the texts. For the three-act structure, all sections can be identified with the same ease, 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 (see Table 4.7 where the top half of the table shows the RMSE values per section for the three-act structure).

For the four-part structure (see the bottom half of Table 4.7), the first section is easier to identify for two of the text. Here again *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.

Table 4.7: The average RMSE per section with a lower RMSE showing a higher agreement between annotators and the system per section.

Text	Section		
	1	2	3
<i>Three-act structure</i>			
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	–
<i>Four-part structure</i>			
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

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 suboptimal initial split for a three-act structure. To assess potential bias towards either structural theory, the system is compared against a random baseline. Table 4.8 shows that the system does place three-act structure transitions slightly further away from the baseline transitions than for the four-part structure.

Table 4.8: The RMSE between the system and a random baseline to evaluate the bias of the system 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

4.5 Discussion

Computational literary analysis relies heavily on the ability to identify the high-level structure of a text. To do this, first transitions within the text needs to be identified. 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 the transitions on the sentiment in the text. Heyns and van Zaanen (2022) and Ouyang and McKeown (2014) were followed in using structure theories to identify transitions in a text.

This chapter reviews 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, guidelines were created for annotating a text according to its structural elements. A small data set was also created, using these guidelines and comparing the inter-rater annotator agreement. In contrast to previous work using this system, human-annotated data is now used to test the system.

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

4.6 Future work

Overall, the system reflects very similar results to that of human annotators when identifying the overarching structure. However, the structure may be more complicated than identified here, 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 storyline. Although the same method can be used to identify more sub-plot sections (narrative levels) in a story, confirming that a section is a sub-plot proves more challenging.

Furthermore, structure theories often overlap in sections that share a common purpose, as shown in Section 4.2.1. Multiple theories with a similar number of sections exist, but the description/purpose of each section may be different. To that end, it may be beneficial to not only divide the text into sections but also identify the purpose of each section.

Section 4.3.1 refers to the fact that only the events that take place in a narrative are taken into account 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. These narrative elements fall under narrative discourse and will be investigated in Part II.

In the current work, segments are created on a sentence level to avoid the unlikely event of a transition mid-sentence, identified by the system. However, this may not be the optimal length for the LDA model to identify a topic. Multiple sentences can be concatenated 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.

4.7 Conclusion

This chapter directly addressed the three main research questions posed in the introduction to this chapter.

RQ 4: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?” is motivated by the need for systematic and standardised methods of literary analysis in the digital age. A set of clear guidelines was created based on well-known structure theories (three-act and four-part). Five human annotators used these guidelines to mark up three short mystery stories. The results show high agreement between annotators (average RMSE of 0.061 for three-act and 0.193 for four-part). This is much better than past work that reported low agreement on similar tasks. 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.

To address **RQ 5:** “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?” two theories were tested: the three-act structure (common for short stories) and the four-part structure (common for mysteries). The three-act structure showed higher agreement between annotators (RMSE 0.061 vs. 0.193) and between humans and the MTI system (RMSE 0.092 vs. 0.144). This means the three-act fit is better for short whodunits. A small data set of three annotated short stories was built, which will be replaced by a larger dataset in Chapter 7 and be made publicly available. This is a useful resource for future work.

RQ 6: “Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?” was also partially addressed. The MTI system, originally developed on artificial concatenated text (Chapter 3), was fine-tuned on the human-annotated corpus and evaluated against human judgement. It achieved RMSE values of 0.092 (three-act) and 0.144 (four-part) when compared with human judgement- performance comparable to IAA.

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. The human and system annotations were also compared to a baseline. The human annotators compared to the baseline did not show a big difference, but again the three-act structure shows slightly higher and more consistent agreement. Only when comparing the system with the baseline the four-part structure shows slightly better agreement. From this, it can be concluded that generally, the three-act structure is 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.

By bridging the artificial-evaluation limitation identified earlier, MTI provides a practical entry point for large-scale computational literary analysis, supplies empirical evidence for narrative structure theories, and establishes a reusable benchmark for subsequent refinements, for example, multi-level sub-plots, dialogue-aware segmentation.

Part II

Narrative Discourse

Part I explored ways in which thematic narrative arcs, high-level shifts in story topics, can be extracted from text, both through human interpretation and with computational methods using the MTI system. Part I of the thesis focused on story as described in Chapter 2.

However, to fully understand a narrative's structure, it is not enough to consider arcs alone. Stories are not simply sequences of emotional shifts or thematic evolutions; they are also spatially and temporally situated, embedded in layers of narrative perspective, and shaped by the alternation between scenic and non-scenic passages. Part II of this thesis, shifts the focus from story to narrative discourse, structural aspect that has increasingly gained attention in CLA. This move is motivated not only by theoretical considerations, but also by the need to refine and expand our annotation framework to accommodate the complexity of narrative organisation beyond arcs.

A range of recent work has engaged with the segmentation of narratives into scenes and the classification of narrative levels, often using overlapping yet inconsistently defined terminology. Concepts such as *scene*, *narrative level*, *embedded narrative*, and *focalisation* recur across publications, though seldom with identical definitions. As was shown in the literature review in Chapter 2, annotation efforts in this space have largely been informed by Genette et al. (1980), with expansions that account for character constellations, spatial continuity and shifts in discourse time. Yet, despite these developments, a unified terminology remains elusive. What follows is an examination of the terminology and definitions proposed in the literature, with the aim of consolidating existing approaches into a single guideline for scene and narrative level annotation. This lays the groundwork for an updated annotation framework that reflects both narratological theory and computational feasibility.

This second part of the thesis focuses on automatically detecting forms of narrative discourse, for example, narrative levels, scenes and anachronisms using a fine-tuned transformer model. Transformers (like BERT) are powerful because they understand context, word order and meaning in long texts, better than LDA for fine details. Chapter 8 discusses the advantages of this approach. However, a model cannot be fine-tuned without annotated data. Before data can be annotated, clear definitions and guidelines are needed. So, the path is clear:

1. Chapter 5 defines and creates guidelines for marking narrative levels, scene breaks, and time shifts. This is based on theories from Genette et al. (1980) and Propp (1968), and others.
2. Chapter 6 annotates texts using the guidelines from Chapter 5 and discusses problems met along the way.
3. Chapter 7 presents the final data set that will be used as a training set.
4. In Chapter 8 a transformer model is fine-tuned that can predict labels from raw text, then tests how well it matches human annotation.

Chapter 5

Narrative Discourse Annotation Guidelines

This chapter reviews and compares existing annotation guidelines for narrative discourse (for example, scene, level, focalisation, temporal aspects) annotation. The content of this chapter is based on Heyns, N., & Van Zaanen, M. (2024, November). Annotating mystery novels: Guidelines and adaptations. In Y. K. Lal, E. Clark, M. Iyyer, S. Chaturvedi, A. Brei, F. Brahman, & K. R. Chandu (Eds.), Proceedings of the the 6th workshop on narrative understanding (pp. 55–66). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.wnu-1.9> where existing guidelines are consolidated as far as possible and a new framework, on the basis of the consolidated guidelines, is proposed.

This chapter also shows how these guidelines can be effectively extended from general-purpose to specialised contexts, such as mystery novels that feature unique narrative elements like clues and detective thoughts.

The resulting annotated corpus of whodunits serves as a reusable resource for training and evaluating computational models of narrative structure.

5.1 Introduction

The task of narrative discourse analysis is often reduced to the process of narrative scene segmentation, which involves the identification of distinct scenes within a narrative. It is a crucial task in the field of computational literary analysis. For instance, it allows researchers to better understand the structure and pacing of literary works, which can reveal insights about the author's stylistic choices and the overall narrative arc. In addition, the ability to identify scenes can improve other tasks, such as summarisation (Droog-Hayes et al., 2018), literary machine translation (Taivalkoski-Shilov, 2019), generation of narratives (Botelho, 2021; Lukin & Walker, 2019; Porteous et al., 2016), character interaction analysis

(Agarwal, 2011; Chen & Bunescu, 2021; Fields et al., 2022; J. Lee, 2017; Macovei, 2017; Min & Park, 2016a, 2016b, 2016c; Porteous et al., 2016), and topic modelling (Schmidt, 2015).

Annotating literary texts presents challenges due to the often ambiguous and multifaceted nature of literary terms, which resist straightforward, computer-friendly definitions. To tackle this, standardised definitions and annotation schemes for key narratology elements such as narrative level, scene, focalisation, and anachronisms (including flashbacks and flashforwards) are essential.

This chapter therefore answers two research questions posed in Chapter 1: **RQ 4**: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?” and **RQ 5**: “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?”

Standardised guidelines ensure that these analyses are conducted in a consistent and comparable manner. They save time and effort in annotating texts, and enable the creation of reusable annotated data sets. Well-defined annotation guidelines are crucial to obtain high quality IAA (Alrashid & Gaizauskas, 2021). Despite previous attempts to standardise the guidelines for narrative scene segmentation (Alrashid & Gaizauskas, 2021; Gaizauskas & Alrashid, 2019; Kearns, 2020; Zehe, Konle, Dümpelmann, et al., 2021; Zehe, Konle, Guhr, et al., 2021), there remains a need for a comprehensive and widely adopted set of best practices.

First, essential concepts are identified, such as narrative, narrative levels, anachronisms, focalisation, scene, non-scene and ellipsis in Section 5.2. Next existing data sets are identified in Section 5.3. Section 5.4 then compares existing annotation guidelines, noting that similar concepts are often defined differently or annotated using various techniques. The aim is to consolidate concepts and streamline the annotation process. Section 5.4 also combines and integrates guidelines from previous work to create a cohesive set of annotation guidelines that can be found in Appendix A. Section 5.5 discuss the annotation scheme. Finally, to investigate practical applicability, the annotation guidelines are applied to a new data set in Section 5.6.

As mystery novels have specialised phases and the characters have specific roles, this chapter proposes to extend the annotation scheme in a modular fashion. Using this annotation scheme, a genre-specific data set is annotated, and how it complements the existing publicly available data sets is discussed. Note that other modular extensions may be proposed as well, which can be added and taken away as needed.

The contribution of this chapter is threefold. First, existing narrative annotation schemes are consolidated. Second, a modular extension of the annotation scheme is proposed. Modular extensions allow for annotation of specialised narrative genres, such as whodunits. Third,

the new annotation scheme with its modular extension is applied to a set of narratives, showing its practical applicability. This results in an annotated data set of whodunits.

5.2 Background

To form a solid basis for investigation into the annotation guidelines, the foundational concepts of narratology as defined in literary theory, particularly drawing from the works of Genette (for example, Genette et al., 1980) are presented. In addition, it is examined how previous work on computational narrative understanding has translated these literary concepts into computer usable definitions.

5.2.1 Narrative

According to Eisenberg and Finlayson (2021), a narrative is a linguistic representation that presents a coherent sequence of events involving specific characters and times, organised into a structured plot. It goes beyond common sense coherence by employing elements such as climaxes and other plot structures. This aligns with the concept of narrative discourse of Genette et al. (1980) with the complexity of storytelling lying in the strategic choices of detail revelation, plot order, and narrative interruptions.

While Part I of this thesis focused on story arcs and the high-level thematic and structural progression of the *histoire*, this chapter shifts attention to narrative discourse (*récit*): the “how” of storytelling, encompassing temporal manipulation, embedding, focalisation, and the segmentation of discourse into scenes and levels.

Narrative discourse is shaped by the interaction of characters and events driving the plot forward. Narratives can appear contiguously as a single, solid text, or they can be embedded in another narrative, or can even interrupt the preceding narrative (Eisenberg & Finlayson, 2021).

Segmenting narrative discourse means identifying thresholds in the narrative that relate information about the internal structure and pacing rather than thresholds related to chronological plot (as discussed in Part I). The goal of these narrative discourse segmentation tasks is typically to identify changes in time, place, or focalisation that signal the beginning and end of scenes, shifts between narrative levels (embedded stories, flashbacks), or deliberate manipulations of narrative speed (summary versus scene versus ellipsis).

5.2.2 Narrative levels

Narrative levels refer to the hierarchical structure of a narrative, in which the overall story can be composed of multiple nested levels of narrative. Genette et al. (1980)

identify narrative levels in terms of the role the narrator plays in telling and ordering the story. According to Genette et al. (1980), there are three primary narrative levels: the *extradiegetic* level, which is the level of the narrator or implied author; the *(intra)diegetic* level, which is the level of the characters and the events they experience in the story; and the *metadiegetic* level, which is a secondary narrative embedded within the primary diegetic level.

On a practical level, annotating the narrative levels requires identification of clear thresholds between diegesis. From the framework by Genette et al. (1980), it can be inferred that the threshold of a level is where the narrator changes. This leads to narrative levels in the form of embedded or interruption narratives (Eisenberg & Finlayson, 2021). *Embedded narrative* occurs when a plot event in the original narrative triggers the telling (in other words, embedding) of another narrative. This occurs, for example, when a character narrates a story in a dialogue of the main narrative. Embedded narratives typically occur on the metadiegetic level. *Interruptive narratives* interrupt the original narrator's narration. This is common in narratives where, for instance, each chapter has a different narrator.

5.2.3 Anachronisms

Anachronisms are deviations from the main temporal progression of the story (Kearns, 2021). A number of types of anachronisms can be identified (Eisenberg & Finlayson, 2021).

A *flashback* (analepsis) occurs when the time in which the events are told, shifts from the present to the past, whereas a *flashforward* (prolepsis) occurs in the form of visions, prophecies, or foreshadowing. Both can be either embedded or interrupted. *Embedded flashbacks* occur when the narrator is telling a story about the past, from the present time. In contrast, *interruptive flashbacks* replace the original narrative. Here, the original narrative ends and a new narrative starts (with events taking place before the original narrative). The narrator also moves in time, whereas the narrator is still in the present time with embedded flashbacks.

Some research places anachronisms at a narrative level, but Ketschik et al. (2021) mentioned that anachronisms deal with the logical order of the discourse and do not leave the present narrative level. Because the narrator does not change in anachronistic narrations, narrative levels should be distinct from anachronisms.

5.2.4 Scene, non-scene and ellipsis

Gius, Jannidis, et al. (2019a) introduced the concept of a scene as a segment of narrative discourse that presents the *histoire*, so that time, place, and character constellation stay more or less the same. They did not explicitly relate scene to narrative levels, although a relationship must exist since both serve different (yet similar) aspects of narratives.

According to Gius, Jannidis, et al. (2019a), there are four main aspects of a scene to define the boundaries: time, space, events, and characters. A scene often changes with a significant shift in *time*, such as when the narrative pace shifts from minutes to days. Similarly, a change in *space* or location triggers a scene change, though smaller locations may be grouped using the container principle (Gius, Jannidis, et al., 2019a), which groups smaller rooms or locations together, allowing for small changes of location (within the same container) without scene changes. A scene also changes when a new action or *event* starts. Again, a container principle can be used here. Finally, a shift in *character* constellation (for example, when characters join or leave) changes the scene. However, the scene does not change if the action remains the same. Also, a change of narrator does not necessarily cause a scene change if the main aspects remain the same (Gius, Jannidis, et al., 2019a).

Gius, Jannidis, et al. (2019a) also recognise *non-scenes*, which do not contain any acting characters. These mostly occurs as summaries, descriptions, or scenic passages. Non-scene information that briefly interrupts a scene or that occurs at the start or end of a scene, and that is too short to be considered a separate segment, is typically recognised as part of the scene.

Genette et al. (1980) defined *ellipsis* as a form of narrative duration that omits certain events or periods of time in a narrative. This creates gaps that the reader must fill in, and it is often used to accelerate the pacing or to focus on significant moments without detailing every occurrence.

Although Genette et al. (1980) do not specifically use the terms scene and non-scene, the terms (along with ellipsis) can be placed as a form of narrative duration on the intradiegetic level of a text. Therefore, there can be several stories on an intradiegetic level consisting of scene, non-scene and, indirectly, ellipsis (Ketschik et al., 2021).

5.2.5 Narrative perspective

Focalisation is the perspective from which the narrative is seen (Wirén & Ek, 2021), or how much information the narrator has access to. Different levels of focalisation can be identified (Todorov, 1971). *Zero or unrestricted* focalisation provides a fully omniscient perspective. The narrator knows more than any of the characters. *Internal* focalisation is narrated from the perspective of a character in the story, with the narrator knowing as much as the character. With *external* focalisation, the perspective is outside the character in the story and the narrator knows less than any of the characters.

Narrative voice indicates the narrator's relationship with the text, and whether they are present in the text or not (Ketschik et al., 2021; Wirén & Ek, 2021). Narrative voice can be either *homodiegetic*, when the narrator appears in the story. They usually refer to themselves in the first person. Narrative voice can also be *heterodiegetic*, when the narrator does not appear in the story. The narration is mostly in the third person.

5.3 Existing data sets

To our knowledge, there are only three publicly available data sets annotated with narrative segmentation. Two of these were created in the SANTA project (Gius, Reiter, & Marcus, 2019), which was a significant effort in developing annotation guidelines and annotating narrative structure. Several researchers (Barth, 2019; Bauer & Lahrsow, 2020; Eisenberg & Finlayson, 2021; Hammond, 2021; Kearns, 2019; Ketschik et al., 2019; Wirén & Ek, 2021) took part in this task by creating annotation guidelines. Barth (2021), Kearns (2021), and Ketschik et al. (2021) later extended their guidelines.

On the basis of these results, the project established annotation guidelines for narrative levels, which were also applied to a corpus¹ in a shared task. Note that the data sets of Chung et al. (2018), Kearns (2020), Newberry and Bailey (2019), and Rogers et al. (2024) are not publicly available.

Additionally, Gaizauskas and Alrashid (2019) proposed SceneML to annotate scenes, locations, characters, and time in narratives. Unfortunately, the annotation scheme is vague on how to treat narrative description and levels of narratives.

5.4 Annotation guidelines

A number of standardised guidelines for annotating the key narratological elements have been proposed, many stemming from the SANTA shared task (Barth, 2019; Bauer & Lahrsow, 2020; Eisenberg & Finlayson, 2021; Hammond, 2021; Kearns, 2019; Ketschik et al., 2019; Wirén & Ek, 2021). Note that some of these guidelines have been updated. This study only refer to the most recent version.

Gius et al. (2021) compare the SANTA annotation guidelines, highlighting the strengths of each set. In this chapter, the guidelines are analysed in detail and the best annotations are selected to ensure a consistent and coherent annotation scheme. This study focuses on narrative levels, anachronisms, scenes, and focalisations as defined in Section 5.2. Additionally, it is explored how guidelines propose handling metatext, paratext (in other words, the text that surrounds the narrative), and punctuation, while highlighting how guidelines use different terms to refer to similar concepts. An additional set of annotations is then provided, specifically for the annotation of mystery novels, that can be used in a modular fashion.

¹<https://github.com/SharedTasksInTheDH>

5.4.1 Narrative levels

As was mentioned in Section 5.2.2, three levels of narrative are recognised: extradiegetic, intradiegetic, and metadiegetic. Wirén and Ek (2021) introduced a guideline to annotate the extradiegetic level by using the tag `NARRATOR`, combined with the numerical value 0. For the intradiegetic level, they used the same tag, but combined it with the numerical value 1.

Metatextuality occurs on the extradiegetic level as moments in which the text comments on itself or the act of storytelling (Genette et al., 1980). Barth (2021) classified these sections as “metanarrative” or “metafiction”. However, Ketschik et al. (2021) argued that the exegesis (the narrator’s direct commentary or explanation addressed to the reader) and diegesis (the primary narrative presentation of the story events) become intertwined and suggested not annotating any levels here, although they added a “non-narrative” tag. Wirén and Ek (2021) annotate a form of metatextuality simply as narrator’s discourse on an extradiegetic level, regardless of the degree of overtness (and hence used the value 0).

The *metadiegetic level* refers to embedded stories, often told by characters in the intradiegetic level. Wirén and Ek (2021) used the term “narrator discourse”, which can be embedded. This embedded narrator discourse is the equivalent of what this study refer to as narrative level (as their criterium for a threshold is a switch in the narrator).

As a critique on using the narrator as a threshold for level changes, Barth (2021) and Ketschik et al. (2021) argue that not all narrator changes cause a change in level. New narrative levels can be introduced without a prototypical change of narrator. Also, the introduction of a new speaker does not necessarily signal a level change as the speaker would have to narrate a separate story (Ketschik et al., 2021). Similarly, Ketschik et al. (2021) argue that homodiegetic narrators can tell embedded stories, they are not part of. In this case, narrators remain the same, but the level changes, as there is a change in the narrator’s position in relation to the story they tell.

Barth (2021) collectively referred to embeddings and framed narratives as acts (which are placed on a horizontal level) and separated them from narrative levels (which are placed on a vertical level). They used Genette’s requirement of a narrator change to induce an act change. The main distinction was that the different narrators of acts were on the same narrative level. Barth (2021) argued that the opening of a new act requires a change in time, setting, or characters (or a combination thereof), which is similar to scenes as proposed by Gius, Reiter, and Marcus (2019). Ketschik et al. (2021) made a similar distinction between vertical and horizontal thresholds. However, they use the terms story (horizontal) and level (vertical). Here, story is defined as a self-contained action whose events and happenings are causally linked and cause a change of state. Hammond (2021) also made the distinction between vertical and horizontal levels but referred to “frames”.

Eisenberg and Finlayson (2021), Hammond (2021), and Ketschik et al. (2021) suggested using numbers to indicate the vertical degree of the narrative level and letters to indicate

the horizontal, sequential arrangement of acts. Similarly, Kearns (2019) used a level tag to indicate an embedded narrative, but also used a numerical value to indicate the sequential acts.

Barth (2021) and Ketschik et al. (2021) distinguished between illocutionary (for example, speaker change) and ontological (narrator change) boundaries as introduced by Ryan (1992). They added that boundaries can also be crossed actually or virtually. However, they only mentioned this and did not include these concepts in their annotation guidelines.

This study suggests keeping with the style of Wirén and Ek (2021) to annotate all three levels with the tag `NARRATOR`. Like Hammond (2021) and Ketschik et al. (2021), numbers are used to indicate the degree, with 0 for the extradiegetic level, 1 for the intradiegetic level, and 2 for the metadiegetic level. Letters can be used to indicate the sequential arrangement on the `INTRADIEGETIC` property and `METADIEGETIC` property. Furthermore, the value `meta` can be used (in addition to the 0 value) to indicate metatextuality in the `EXTRADIEGETIC` property.

5.4.2 Anachronisms

As discussed in Section 5.2, flashbacks and flashforwards can be embedded or interruptive. Eisenberg and Finlayson (2021) differentiated between these and included tags accordingly. Similarly, Kearns (2021) proposed using the tags `ANALEPSIS` and `PROLEPSIS`. These should be used when a new narrative or point in time starts.

Ketschik et al. (2019) emphasise that the narrative level does not change with prolepsis or analepsis and similarly with character thoughts, dreams, and visions. However, they did not seem to include tags for any of these cases in their guidelines.

Following Eisenberg and Finlayson (2021) and Kearns (2021), the tags `ANALEPSIS` and `PROLEPSIS` are used and the properties `EMBEDDED` or `INTERRUPTIVE` are assigned as needed.

5.4.3 Scene, non-scene, and ellipsis

Existing guidelines do not explicitly position scenes, non-scenes, and ellipses in relation to the narrative levels. Our understanding is that scenes, non-scenes, and ellipses can occur in either the intradiegetic or metadiegetic levels, but this does not imply a strict hierarchy between scenes and diegetic levels. In fact, there is not necessarily a strict hierarchical relationship between diegetic levels themselves. The intradiegetic level and the metadiegetic level may have a hierarchical connection since the metadiegetic level can only occur in the intradiegetic level. However, the extradiegetic level operates independently of this hierarchy, as it can exist outside or in the intradiegetic or metadiegetic levels. As a result, scenes can appear on both the intradiegetic and metadiegetic levels, with multiple

scenes potentially existing on the same level. In addition, the extradiegetic narrator may comment in a scene that is otherwise situated on an intradiegetic level.

The most widely accepted definition of a scene is when time, location, and main characters are constant and focus on one action. Alrashid and Gaizauskas (2021) suggested that a scene can contain multiple actions by grouping “scene description segments” (SDS), or continuous spans of text. This resembles the idea of multiple events in a scene, as long as place, time and characters remain the same. A scene can reference past or future events, similar to embedded narratives, with a character telling another story. Alrashid and Gaizauskas (2023) also proposed scene transition segments (STSs) to refer to text segments where the action shifts between locations as the narrative transitions from one scene to another.

Alrashid and Gaizauskas (2023) and Gius, Jannidis, et al. (2019a) distinguished between scene and non-scene. Alrashid and Gaizauskas (2023), however, only included `SCENE` and `NON-SCENE` tags and did not differentiate between different types of non-scene. The only other work that annotated a form of non-scene was Kearns (2019). They annotated extended (when time is extended relative to story time) or compressed time (when narrative time moves faster than story time) using the tags by the same names. Similar to how Gius, Jannidis, et al. (2019a) used the weight that an aspect carries to determine a threshold and time should in this context also be evaluated with respect to the overall text.

This study uses the definition of scene as provided by Gius, Jannidis, et al. (2019a) and annotates this using the `SCENE` tag. The `SCENE` tag also allows for the properties `TIME`, `PLACE`, and `CHARACTER_CONSTELLATION`. Events or SDSs were not annotated, but if needed, these annotations can easily be added as values to properties of the `SCENE` tag.

This study proposes adding a `TRANSITION` tag to mark STSs. In addition, the container principle is used, in which several (smaller) locations can be contained in a larger one. Furthermore, this container principle is extended to STSs: if the transition text carries significant weight in the overall narrative, it is marked as a transition. However, if the transition occurs between places that are contained in the same scene, then the transition is not annotated.

For simplicity, instead of using a separate tag for compressed time and extended time, to indicate the accelerated speed of narration (or summaries as defined in Section 5.2), or descriptive passages, the tag `NON-SCENE` is used and a property of `SUMMARY`, `DESCRIPTIVE_PASSAGE`, or `SCENIC_PASSAGE` is assigned.

To our knowledge, no work includes ellipsis in their annotation guidelines. This study advocates the annotation of omitted time in narratives as it plays an important role in the pacing of a narrative. This can be annotated with the tag `ELLIPSIS`.

5.4.4 Narrative perspective

Focalisation provides important information of the narrator's perspective and the extent of their knowledge in a narrative. In addition to information on the perspective of the narrator, dreams, visions, fantasies, and thoughts are forms of focalisation. Even though they do not represent a change in narrative level (Ketschik et al., 2021), they are essential for understanding the narrator's role and the structure of the narrative. Eisenberg and Finlayson (2021) similarly categorise dreams and visions in the same way as flashbacks and flashforwards, labelling them as either embedded or interruptive on the basis of their function within the narrative.

Wirén and Ek (2021) offer a detailed framework for annotating focalisation. In this framework, character discourse is broken down into "turns", which include a single speaker addressing multiple addressees, and "lines", which correspond to a single addressee or set of addressees. To enhance the precision of dialogue annotation, characters are assigned numeric values, and narrative construction (NC) tags are used to mark speech-framing constructions within lines.

Although these detailed annotations help capture the nuances of character interactions and narrator shifts, it was decided not to use them, as they seem to be overly detailed for the scope of most studies. Moreover, as Ketschik et al. (2021) advises, it is important not to overemphasise focalisation, so a more balanced approach will be adopted.

This study propose the FOCALISATION tag with the possible properties EMBEDDED and INTERRUPTIVE. As mentioned in Section 5.2, narrative voice distinguishes whether a speaker is present in the narrative or not. Following the suggestions of Barth (2021), the tag VOICE was introduced with the properties HOMODIEGETIC and HETERODIEGETIC.

5.4.5 Punctuation and paratext

When considering how to treat punctuation marks and paratext in the annotation process, the aim is to maintain a clear distinction between the narrative elements central to the story and the textual features that serve a more structural or contextual role.

Gius, Reiter, and Marcus (2019) suggest adding punctuation marks inside the annotated segment, but punctuation marks that structure the text, such as asterisks, should be placed outside of the annotated segment. Furthermore, Ketschik et al. (2021) suggests not annotating paratexts, such as titles, forewords, chapter headings, and genre indications. While important for understanding the broader context of the work, these elements are typically considered external to the narrative itself and therefore are excluded from the core annotation process.

This chapter proposes not to annotate punctuation and paratext. The annotation focuses specifically on the narrative itself.

5.4.6 Genre-specific annotations

The predictable structure of classic whodunit mysteries makes them ideal for analysing how narrative elements unfold and for employing digital tools to annotate texts. This section, introduces tags to capture the essential elements that are specific to whodunit mystery novels.

According to Cawelti (2014), whodunits typically include six phases: introduction of the detective, crime and clues, investigation, announcement of the solution, explanation of the solution, and denouement. The following structural tags will be used: `INTRODUCTION` for the detective's arrival at the crime scene, and `INVESTIGATION` for scenes in which the detective gathers clues or interrogates suspects. This combines the crime and clues, and investigation phrases. `CONFRONTATION` is used for the solution announcement. `CONFESSION` deals with the explanation of the solution and the confession. `REVEAL` describes explaining the solution, with `AFTERMATH` annotating the denouement.

Additionally, whodunits make use of specific concepts that can be identified. The crime scenes are annotated by specifying the value of the `PLACE` property of the `SCENE` tag, and clues are tagged with `CLUE`, with optional `IDENTIFIED` and `REFERRED` properties to distinguish when a clue is first found and when it is referenced later respectively.

Given the genre's reliance on the detective's thought process, a `DETECTIVE_THOUGHT` property is added to the focalisation tag. This includes not just direct thoughts, but also gestures reflecting the detective's thinking, especially when the narrator is another character.

5.5 Annotation scheme

The teams that took part in the SANTA project used different annotation schemes ranging from XML to Excel documents. Eventually, all the annotations were translated to CATMA annotations². CATMA offers several advantages that make it a suitable tool for annotating texts with narrative elements. Its flexibility and customisation allows researchers to create and adjust annotation categories and schemes to fit the specific needs of their analysis, making it particularly useful when dealing with genre-specific texts. The platform supports multiple annotation levels, enabling the tagging of narrative elements without a set hierarchical structure. This is especially useful in this context in which it is often unclear what the order of tags should be. For example, is the diegetic level the outer layer that can include scenes or are scenes the outer layer and can contain a diegetic level? Although it is possible to add attributes to an XML file to deal with these situations, CATMA provides a seamless approach to the structure. In addition, CATMA provides textual analysis integration to examine the data. The tool also supports XML export, ensuring

²<https://catma.de/>

that the annotated data can be easily shared, reused and integrated with other tools or systems, which is crucial for collaborative research and future studies. CATMA will also be used to annotate the data set in this and future chapters. Appendix A provides the complete list of tags, properties, and values used during the annotation process following the CATMA annotation scheme.

5.6 Data set

This study uses short mystery stories, allowing us to observe the annotation process across complete texts. Chosen for this study are *The Jewel Robbery at the Grand Metropolitan* and *The Adventures of the Italian Nobleman* both Agatha Christie stories and *A Case of Identity* and *The Red-headed League* which are Sir Arthur Conan Doyle stories. To ensure that the annotated data set can be made publicly available, texts in the public domain from Project Gutenberg³ are used. Information on the texts can be found in Table 5.1.

An overview of the number of tags annotated for each of the selected stories will be provided in Appendix B (due to space restrictions).

Table 5.1: Properties of the annotated texts. AWS represents the average number of words per sentence.

Text	# words	AWS	# sentences	# tags
The jewel robbery at the grand metropolitan	5029	9.5	526	41
The adventures of the Italian nobleman	3783	10.3	364	59
A case of identity	6990	17.5	399	97
The red-headed league	9115	15.9	574	82

Here only four texts are annotated in order to identify and understand potential problems that arise during annotation. Chapter 6 refines the guidelines based on these issues. The complete data set is annotated in Chapter 7. Table 5.2 shows a size comparison between our current subset of the complete data set presented in this chapter and the SANTA⁴ data set. While our data set is smaller at 26 825 words across four texts, its strength lies in containing full narratives, allowing for a thorough analysis of story flow and development, which shorter text extracts may miss. This provides a better overview of narrative techniques from start to finish.

Table 5.2: Comparison with the SANTA data set ranked by total word length.

Data set	# texts	Max words	# words
SANTA	25	2000	50000
Our data set	4	9115	26825

³<https://www.gutenberg.org/>

⁴<https://github.com/SharedTasksInTheDH>

5.7 Annotation process

All annotations were manually done by the author of this thesis using CATMA. No compensation was received for the annotations. The manual annotations provided a foundation for the development of a method to automatically annotate similar texts, which is explored in Chapter 8. First we discuss the problems that arose during the annotation process in Chapter 6 and present the revised, publicly available data set in Chapter 7

5.8 Discussion

This chapter has consolidated a diverse set of existing annotation guidelines into a unified, coherent framework for capturing key elements of narrative discourse. By systematically comparing definitions and practices from the SANTA project and related work (Barth, 2021; Eisenberg & Finlayson, 2021; Gius, Jannidis, et al., 2019a; Ketschik et al., 2021; Wirén & Ek, 2021), terminological inconsistencies and overlapping concepts particularly around narrative levels, scene boundaries, anachronisms, and focalisation were resolved, while preserving the theoretical grounding in Genette et al. (1980). The resulting scheme introduces clear, computable criteria for segmentation thresholds and supports flexible, non-hierarchical tagging via the CATMA platform.

A core contribution is the modular design of the framework. The base layer provides general annotations applicable to any literary text, while optional extensions, such as the mystery-specific module for clues and detective thought, or arc information as investigated in Part I, demonstrate how task-specific features can be integrated without compromising consistency or re-usability. This modularity addresses a major limitation in prior schemes: their rigidity when applied beyond the original domain.

The consolidated guidelines now offer a standardised and extendible foundation for computational literary analysis, enabling reproducible annotation, data set collaboration, and scalable model development. Practical challenges and edge cases encountered during initial application of the scheme to a subset of texts are analysed in detail in Chapter 6.

5.9 Conclusion

This work addressed the challenges of narrative scene segmentation by consolidating existing narratological annotation schemes and proposing a modular extension for a genre-specific data set.

RQ 4: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?” was addressed. This chapter reviewed and combined existing annotation schemes for scene boundaries and narrative levels (Barth, 2021; Gius et al.,

2021; Ketschik et al., 2019). From this, a unified framework was built that defined clear rules for scene changes, narrative levels, and event types (actions, thoughts, dialogue). These rules reduce confusion and made annotation repeatable. The framework is general enough to work on any literary text.

In addition, **RQ 5**: “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?” is answered. This chapter showed how to extend the core framework with a genre module for mystery novels. This module adds *clue events* (when a clue is found or hidden) and *detective thoughts* (internal reasoning). These fit modularly into the main system. The mystery module creates a controlled test space: whodunits have strict event order. This lets us study how genre shapes structure in a clean way.

Chapter 7, will expand the data set and encourage the development of additional genre-specific annotations and data sets. This study aims to contribute to a more comprehensive and widely adopted set of best practices for narrative annotation. Additionally, in the future experiments will be conducted with this data set to identify types of features that help with automatic annotation. It is currently unclear what kind of features will be useful, for example, lexical, syntactic, semantic and pragmatic features or properties from, for instance, character and location networks.

Chapter 6

Problems During Narrative Discourse Annotation

The annotation process from Chapter 5, despite meticulous planning and guideline formulation, revealed several practical challenges. These issues predominantly arose around the application of the container principle, deciding scene boundaries during transitions, and distinguishing embedded focalisations from simple dialogue. These challenges reflect the inherent complexity and subjective interpretation involved in annotating literary texts (Heyns & van Zaanen, 2022). Identifying these problems was instrumental in refining our annotation methodology. This chapter is based on work done in Heyns, N., & van Zaanen, M. Refining Narrative Annotation for Segmentation: Challenges and Debates. In: In Comparative Literature Goes Digital DH2025 Mini-Conference. 2025.

6.1 Introduction

Chapter 5 proposed a consolidated annotation scheme for narrative segmentation, extending it modularly to accommodate genre-specific elements in whodunit mystery novels. While the scheme aimed to standardise the annotation of narratological elements such as scenes, narrative levels, and focalisation, its application revealed practical complexities that warrant further exploration. These challenges stemmed from the inherent ambiguity of literary concepts, which resisted rigid computational definitions, and the unique narrative dynamics of the genre, such as embedded clues and shifting perspectives.

This chapter revisits the annotation process to address these difficulties, offering a reflection on the implementation of the guidelines. The chapter presents specific examples from the annotated data set from Chapter 5 to illustrate problematic cases. In addition, this chapter theorises on the conceptual boundaries of scenes, non-scenes, narrative levels, and character presence, examining how their definitions influence narrative interpreta-

tion. This process serves to refine annotation practices, enhance IAA, and deepen our understanding of narrative architecture in both theoretical and computational contexts.

The contributions of this chapter are threefold. First, it documents and analyses challenges in applying the annotation guidelines; second, clarifications are proposed to improve their usability; and third, the theoretical implications of narratological concepts are explored, emphasising their role in shaping narrative comprehension. This work seeks to advance standardised annotation practices and supports the development of robust computational models for narrative analysis.

This chapter continues to explore **RQ 4**: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?”.

6.2 Annotation challenges

Throughout the annotation process, several challenges arose that required careful consideration. These challenges were often linked to the application of the container principle, the tagging of transition segments, and the differentiation between dialogue and embedded sections. Each of these issues will be discussed in some detail below.

6.2.1 Container principle

The container principle proved valuable, but it introduced some inconsistencies in the annotation process. Deciding when to apply this principle was challenging, particularly when characters left and then returned (for example, *The man returned shortly; with him came the manager.*), or when the narrative moved between locations in the same building. The significance of a place or character constellation was not always straightforward to determine. For instance, in *The adventures of the Italian nobleman*, the story shifts between rooms in a hotel, such as the hallway, lobby, and lift. This chapter concluded that spaces like the hotel room and kitchen were significant enough to justify scene changes, while the hallway, lobby and lift were often transitions between these key spaces.

However, this raised the question of how to group these transitional spaces. When a character moves from the lobby to the elevator, then into the hallway, and finally enters a room, we want to split the scenes between the hallway and the room. This requires determining the precise moment the character crosses the threshold from one space to another. For example, consider the sentence: *The manager produced the key without more ado, and we all entered the flat.* Here, the first half of the sentence places the characters in the hallway, but by the end, they are inside the room. In such cases, we might annotate the sentence as a transition segment, although it could be seen as either part of the previous scene or the beginning of a new one. It can also be seen as a separate scene with the place value set to `transition`. This study took the latter approach where

a scene where characters move between two places, such as travelling in an elevator or a car between locations is annotated as a separate scene with the place value `transition`.

6.2.2 Scene and non-scene

Another challenge was understanding the interplay between scenes and summaries. Distinguishing between scenes and summaries proved challenging when passages exhibit continuity in time, place, and character constellation yet differ markedly in narrative weight or event significance. For instance, a rapid sequence of minor actions within a single location, such as a character dining, playing music, and retiring to bed, may superficially resemble a scene but functions primarily to bridge chronological gaps without advancing plot or character goals; such passages are better tagged as `NON-SCENE` with property `SUMMARY`. Conversely, a passage describing a search party forming, tracking footprints through snow, and reaching a wall contains multiple events but constitutes a single `SCENE` because the actions share a unified investigative purpose and yield a meaningful narrative outcome. To resolve this, the guidelines require that a `SCENE` contains at least one functionally significant event, an action, interaction, or revelation that drives the story forward. Multiple minor actions (for example, smoking, waiting, walking between rooms) within stable time, place, and character boundaries are grouped into one `SCENE` only if they serve a coherent dramatic or informational goal; otherwise, they are annotated as `NON-SCENE` with `SUMMARY`, regardless of surface-level continuity.

Often, a sub-scene (an event or moment that is part of a larger scene) can be narrated as a summary. In such cases, the boundary between scene and summary is not clear-cut. For instance, a narrative may describe the actions within a scene in detail, then briefly summarise the events that followed within the same scene. This overlap suggests that scenes and summaries are not mutually exclusive; rather, they can coexist, with a sub-scene being narrated through summary within the broader scene.

Furthermore, our guidelines assume that direct speech occurs exclusively within scenes, as it typically involves characters interacting in a specific spatio-temporal context. However, annotating scenes proved less straightforward than anticipated due to subtle shifts in these elements and the interplay with dialogue. Consider the opening of *A Case of Identity*: “*To Sherlock Holmes she is always THE woman. [...] And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.*” This passage lacks immediate action or dialogue, focusing instead on the narrator’s (Watson’s) summary of Holmes’s enduring perception. Though anchored in the intradiegetic level and tied to a specific character and implied past time, it contains no enacted event. this is annotated as `NON-SCENE` with property `SUMMARY`, but this decision underscores the ambiguity: should character-focused reflection, even without interaction, count as a scene, or must a `SCENE` require active presence and progression?

Non-scenes, encompassing summaries, descriptions, and scenic passages, posed further challenges when multiple non-scenes followed one another with distinct topics or purposes. Our guidelines permit tagging sequential non-scenes separately, but determining their boundaries required careful judgement. In the opening passage of *A Case of Identity*, two distinct summaries emerge: “*To Sherlock Holmes she is always THE woman. I have seldom heard him mention her under any other name. [...] And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.*” Which addresses Holmes’s feelings towards Irene Adler, followed by: “*I had seen little of Holmes lately. My marriage had drifted us away from each other. [...] Beyond these signs of his activity, however, which I merely shared with all the readers of the daily press, I knew little of my former friend and companion.*” Which summarises Holmes’s public activities over an extended period. These are annotated as separate NON-SCENE segments with SUMMARY to reflect their topical and functional shift. Yet their adjacency, both serving to advance the narrative timeline without introducing a new scene, raised the question of whether they should merge into a single non-scene. To maintain precision, it is required that adjacent non-scenes be segmented only when they shift in narrative purpose (for example, personal versus public summary); otherwise, they are combined under one NON-SCENE span to avoid over-fragmentation.

6.2.3 Focalisation

Distinguishing between embedded focalisation and simple dialogue posed significant challenges. Embedded focalisation is defined as a shift in perspective to a different character, or a different temporal or spatial point within the story, while still remaining within the broader narrative framework. For example, the main narrator might describe a scene, and within that scene, a character recalls a past event from their own perspective. However, the focalisation can sometimes shift between the character’s recollection and the narrator’s interruptions, making it difficult to determine where the embedded focalisation begins and ends.

Similarly, it can be challenging to differentiate between an embedded flashback and dialogue that briefly references a past event. For instance, if a character mentions something that happened the previous day in just one sentence, this might not seem significant enough to be tagged as an embedded flashback. An example from *The adventures of the Italian nobleman* illustrates this: “*... You—in company with a friend—visited the late Count Foscatini on the morning of Tuesday the 9th—*” *The Italian made an angry gesture.* In this case, the brief mention of a past event feels more like a part of the dialogue in the present moment rather than a true flashback. However, if that past event is described in more detail, expanding into a paragraph or more, it begins to take on the characteristics of an embedded flashback.

Similarly, annotating interruptions within metadiegetic narratives, where a character pauses an embedded story (level 2a) to return to intradiegetic dialogue (level 1a), presented

challenges in maintaining narrative continuity without fragmenting the overall embedded structure. For instance, a character recounting a past event may be interrupted by another asking a question, causing the text to alternate between levels without concluding the embedded tale. Existing guidelines, such as those in Wirén and Ek (2021) and Ketschik et al. (2021), tie level tagging to strict boundaries, risking the creation of disjointed segments that obscure the unified metadiegetic story. To preserve coherence, interrupting intradiegetic passages (for example, 1a) are annotated within the same span as the ongoing metadiegetic text (for example, 2a), using overlapping tags in CATMA to indicate that the interruption is temporary and part of the same embedded narrative. This differs from separate metadiegetic stories separated by unrelated intradiegetic text, which receive distinct sequential labels (for example, 2a, 1b, 2b). Embedded metadiegetic interruptions occur when the storyteller remains in the present (intradiegetic frame), allowing seamless resumption; interruptive metadiegetic levels, fully replace the intradiegetic narrative and prohibit interruptions, resuming only after completion.

Annotating letters and other forms of mediated communication posed a significant challenge in determining both narrative level and character constellation. When a letter is read aloud in a scene, as in the example “Read it aloud. . . . There will call upon you to-night, at a quarter to eight o’clock. . . .”, the voice originates from an absent character, often in a different time and place, yet is delivered through the current narrator (for example, Watson quoting the note). Existing guidelines, such as Wirén and Ek (2021), tie level changes to a shift in narrator, but here the primary narrator retains control, merely relaying content. Treating the letter as a metadiegetic narrative would incorrectly imply a fully embedded story told by a new narrator. Simultaneously, traditional terms like “offstage character”, “implied character”, or “absent presence” typically denote characters whose influence is merely suggested, perhaps through memory, or mentioned without direct engagement in the scene’s events. For instance, Abrams and Harpham (2011) note that some characters never appear on stage but exert a decisive influence on the action. Of course, this is a concept borrowed from drama and adapted to prose, yet it leans toward implied rather than active presence. However, this under-represents the absent character’s active role: the letter-writer directly shapes the scene’s dialogue, decisions, and plot momentum.

Altman (1982) introduces “epistolary character” often associated with letter-writing, which marks a step toward recognising virtual presence. Altman (1982) states, “the writer [becomes] a character whose absence from the scene of action is compensated by the presence of his or her written voice.” However, this term is often confined to epistolary fiction, limiting its applicability to broader or modern narrative forms like text messages. As a result, annotators face ambiguity: should the letter trigger a metadiegetic level (1a to 2a)? Should the sender be excluded from CHARACTER_CONSTELLATION as physically absent? To resolve this, the guidelines clarify that a letter, message, or recorded speech read in a scene does not change narrative level unless a character begins narrating a separate story in their own voice; the passage remains on the current diegetic level (here, intradiegetic 1a). However, this study propose the term “virtually present character” to encompass characters who actively contribute to a scene via mediated communication, whether

through asynchronous forms like letters or synchronous interactions like text messages. For example, a letter-writer becomes virtually present when their words are read in the scene, though their interaction originates from a prior time and place, distinct from the current setting. Conversely, a character texting engages in real-time, sharing the scene's temporal frame despite their physical absence. Therefore, the sender is designated a virtually present character and must be included in the `CHARACTER_CONSTELLATION` property of the `SCENE` whenever their mediated words actively influence interaction or plot, ensuring boundary detection without introducing false level shifts.

6.3 Misleading character description

Often times characters are described as having one role at the start of the story but then their role changes throughout the story. For example in *The story of the young man and the holy orders*, the main character is in fact the criminal, but at the start of the text he is described as “The detective that there is in all of us awoke and became clamant in the bosom of Mr. Rolles;”. Similar role disparities are seen with characters being described as victims, just to have them turn out to be criminals at the end. Reversed, characters are also strongly suspected as the criminal but later on (as is very often the case in a mystery novel) find them unfortunate bystanders (although often these characters might have been guilty of some other social crime, but not the lawful crime the detective is investigating). This naming convention makes it difficult to annotate a character's role. For example should the character be annotated according to their current role in the scene or according to the role they are revealed to have at the end of the text? The cleanest solution is to annotate the character with the role they are revealed to have at the end. This will of course require the annotator to read the text through at least once before annotating.

6.4 Annotation method

Our annotation process involved multiple passes through the text to ensure accuracy and consistency. The initial round concentrated on identifying and tagging scenes, non-scenes, and ellipses. This foundational layer allowed us to establish the basic structure of the narrative. The second pass, focused on annotating diegetic levels, focalisation, narrative voice, and anachronisms. A final pass, tagged clues, detective thought and the different acts in the narrative.

6.5 Discussion

The challenges encountered during annotation highlight a fundamental tension between the emphasis of narratological theory on interpretive flexibility and the rigid, repeatable

boundaries required for computational modelling. Concepts such as scene, non-scene, narrative level, and character presence are not neutral structural markers but active mechanisms of reader orientation. Scenes ground the “here and now,” non-scenes expand temporal scope, levels nest realities, and virtually present characters extend agency across physical absence. Our refinements, requiring functional significance for scenes, allowing overlapping tags for metadiegetic interruptions, and including mediated voices in character constellations, preserve this interpretive depth while enforcing consistency. Crucially, these adjustments align manual annotation with automated preprocessing pipelines, such as LDA-based topic modelling described in Part I of this thesis. By prioritising narratological continuity (time/place/characters) over discourse-level topic shifts, it is ensured that training data for transformer models reflects human-like structural perception rather than superficial lexical variation.

Theoretically, these issues underscore narrative as a dynamic system; boundaries are not merely segmentational but semantic, signaling shifts in reader attention, suspense, or comprehension. For instance, treating a letter as virtually present rather than metadiegetic avoids false level changes while acknowledging its plot, a decision that mirrors Genette’s distinction between voice and focalisation (Genette et al., 1980). Similarly, requiring a scene to contain a “functionally significant event” aligns with the emphasis of Gius, Jannidis, et al. (2019a) on narrative weight, preventing over-fragmentation of transitional or reflective passages. These refinements therefore bridge computational feasibility with literary insight, demonstrating that robust annotation must encode not just what happens, but how it matters.

6.6 Conclusion

This chapter has provided an answer to **RQ 4**: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?”.

Our analysis demonstrates that yes, such guidelines are possible, but only through iterative refinement grounded in real-world application. The consolidated scheme proposed in Chapter 5 offered a theoretically sound and modular foundation, but its practical implementation revealed significant ambiguities in boundary detection, conceptual overlap, and edge cases, particularly in the application of the container principle, the distinction between scene and non-scene, the handling of mediated communication, and the treatment of embedded or interrupted focalisation.

By clarifying the role of functional significance in scene definition, allowing overlapping tags for interrupted metadiegetic stories, and introducing “virtually present characters” for mediated communication, this study produced a refined, consistent, and computationally viable scheme. By refining the guidelines, their applicability is improved and consistent annotation practices are fostered. Theoretically, this study highlights how narratological concepts shape narrative comprehension, offering insights into their role in literary analysis.

Chapter 7 applies this improved methodology to a full data set of ten public-domain mystery short stories. Future work can further expand the annotated data set, incorporating additional genres to test the guidelines' versatility.

Chapter 7

Annotated Narrative Discourse Data

*This chapter is based on Heyns, N., & Van Zaanen, M. (2025). Annotated Mystery Narratives Data Set. Journal of Open Humanities Data that describes a data set that contains ten annotated short mystery stories. The annotation schema, including all tags and their application rules, is fully detailed in Chapter 5 and further refined in Chapter 6. By implementing the annotation guidelines developed earlier, this chapter relates to **RQ 4**: “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?” and **RQ 5**: “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?”. By applying these guidelines to create a unified dataset that combines general narratological annotations with mystery-genre extensions, this work demonstrates their practical feasibility and modularity as an implementation outcome.*

This genre-specific data set facilitates computational narrative analysis by offering structured narrative information that can be reused for tasks in story segmentation, plot structure analysis, and genre studies. The resulting data set also serves as a tangible realisation of the guideline framework, enabling consistent, reproducible annotations across texts.

*Indirectly, this chapter supports **RQ 7**: “How effectively can existing and enhanced machine-learning methods, applied to an annotated data set, identify and analyse high-level narrative structures?” by providing the high-quality, standardised data required for the automated systems evaluated in Chapter 8. Without this data set, the models could not be trained or benchmarked, linking the annotation process to computational advancements in narrative analysis.*

7.1 Introduction

Computational approaches to narrative structure, such as those explored in the SANTA project, have focused on annotating narrative discourse in the form of narrative levels and scene segmentation, often using shorter texts to identify these elements (Reiter et al., 2019). However, analysing full texts, as done in this data set, is essential to capture the

complete arc of story structures. Furthermore, there may be genre-specific aspects to story arcs, due to the nature of the stories, such as in whodunits (as discussed in Chapter 4). This data set builds on these foundations to provide a resource for studying both general and genre-specific narrative properties, supporting applications in computational analysis and literary studies.

In Chapter 5 an initial round of annotations took place and Chapter 6 discussed and clarified the challenges during the annotation process. Now we extend the data set presented in Chapter 5 with extra annotated text (resulting in a data set of ten annotated texts). This data set includes the text annotated in Chapter 4 although now the complete annotation guidelines resulting from Chapter 5 and 6 are used to annotate the text.

7.2 Method

The collection and annotation of the material in the data set consists of a number of steps: selection and collection, clean-up, import and annotation, and packaging, which will be described in more detail here¹.

7.2.1 Steps

First, a collection of suitable narratives needs to be selected. For this, ten narratives are manually selected from Project Gutenberg². Project Gutenberg provides digital versions of books that are freely usable. Plain text (UTF-8) versions of the books listed in Table 7.1 are downloaded. Note that some narratives come from the same downloaded text file.

Second, the texts are cleaned up. The downloaded files contain non-narrative information, in other words, the colophon, which contains information about the presentation (such as publisher and year), title, table of contents, and foreword. These are removed manually before annotation. Chapter headings or divisions and section breaks were kept as these might contain useful information for the annotation process. Only the narrative itself (including its layout structure) is kept. See Table 7.1 for information on the size of each text³.

Third, the cleaned texts are imported into CATMA⁴. This tool allows, among other things, for manual annotation and analysis of texts. Based on the annotation guidelines as published

¹The data set is uploaded to the repository of the South African Centre for Digital Language Resources (SADiLaR), which can be found here: <https://hdl.handle.net/20.500.12185/693>, under the file name *Annotated Mystery Short Stories Data Set*. The data set is made available under a Creative Commons License Attribution-ShareAlike 4.0 International license.

²<https://www.gutenberg.org/>

³Sentence and word counts were computed using an online text analyser: <https://www.online-utility.org/text/analyzer.jsp>.

⁴<https://catma.de/>

Table 7.1: Information on the narratives in the data set.

Text	Author	# Sentences	# Words
The adventures of the Italian nobleman ^a	Agatha Christie	324	3839
The jewel robbery at the grand metropolitan ^a	Agatha Christie	564	5101
The mystery of Hunter's Lodge ^a	Agatha Christie	408	4618
A case of identity ^b	Sir Arthur Conan Doyle	451	7028
A scandal in Bohemia ^b	Sir Arthur Conan Doyle	686	8628
The red-headed league ^b	Sir Arthur Conan Doyle	632	9273
A question of passports ^c	Emmuska Orczy	257	3767
Story of the young man in holy orders ^d	Robert Louis Stevenson	296	5623
The chronicles of Addington Peace ^d	Bertram Fletcher Robin- son	542	6295
The mystery of the steel disk ^d	Broughton Brandenburg	488	7356
Total		4648	61528

^a Christie, A. (2020). Poirot investigates. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/61262/pg61262.txt>

^b Doyle, A. C. (2015). Adventures of sherlock holmes. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/48320/pg48320.txt>

^c Orczy, B. E. (2004). The league of the scarlet pimpernel. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/5805/pg5805.txt>

^d Patten, W. (2024). Great short stories, volume 1 (of 3) detective stories. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/74548/pg74548.txt>

in Heyns and Van Zaanen (2024) and discussed in Chapter 5, tag types are created in CATMA, which allows for the annotation of narrative structure. The annotation guidelines make provision for 18 unique tags as listed in Table 7.2 and defined in Chapter 5.

Following this, the narratives are read in detail and manually tagged. A total of 511 tags were assigned across all texts. The number of annotations per tag is indicated in brackets in Table 7.2.

Finally, the annotated narratives are stored in TEI XML and CSV formats. The CSV format is generated from the TEI XML files using the `teitocsv` package⁵. The plain UTF-8 text files together with the annotated CATMA files (TEI XML and CSV) are combined with a UTF-8 markdown readme file, that describes the data set, into a zip file that can be found on the repository of the South African Centre for Digital Language Resources (SADiLaR). In addition to the markdown readme file, the zip file contains the following directories:

- `original`: This directory contains the original texts of the ten mystery short stories as downloaded from Project Gutenberg. The names of these files correspond to the UTF-8 filenames downloaded from the Project Gutenberg website as indicated in Table 7.1.

⁵<https://github.com/komax/teitocsv>

Table 7.2: Information on the tags assigned in the data set.

Tagset	Tag	Tagset	Tag
Perspective		Acts	
	Focalisation (2)		Aftermath (9)
	Voice (22) (hodiegetic/heterodiegetic)		Confession (1)
Segment			
	Ellipsis (0)		Confrontation (4)
	Non-scene (Summary (26)/Description (3))		Introduction (10)
	Scene (Time, Place, Characters) (115)		Investigation (10)
			Reveal (10)
Anachronisms		Misc	
	Analepsis (11)		Clue (42)
	Prolepsis (0)		Detective thought (33)
Diegetic level			
	Extradiegetic (69)		
	Intradiegetic (124)		
	Metadiegetic (20)		

- `txt`: This directory contains the cleaned up texts (after removing colophons, titles, tables of contents, forewords) the ten mystery short stories that are used for the annotation process. The names of these files correspond to the names of the narratives as indicated in Table 7.1.
- `xml`: This directory contains the annotated versions of the ten mystery short stories in TEI XML format, generated using CATMA. The names of these files correspond to the names of the narratives as indicated in Table 7.1.
- `csv`: This directory contains the CSV files of the ten mystery short stories that are generated from the CATMA TEI XML files in the `xml` directory. The names of these files correspond to the names of the narratives as indicated in Table 7.1.

7.2.2 Sampling strategy

To select suitable stories, several criteria are considered. First, the narratives should fall in the whodunits genre. The reason for this is that stories in this genre typically follow a relatively rigid storyline. It is expected that this will lead to comparable narrative structures. Moreover, whodunits naturally reflect information gain over time as clues and revelations unfold, providing a unique opportunity to study narrative progression and temporal elements in detail. This aspect allows for deeper analysis of how suspense is built through the gradual disclosure of plot points, making the genre particularly suited

for computational exploration of time-based narrative techniques. This also means that genre-specific annotation can be applied as described in Chapter 5.

In addition, only whodunit stories written in a relatively narrow time frame (between 1891 and 1924)⁶ are selected, as even though whodunits are expected to follow a relatively fixed structure, this structure may develop over time. This period captures the formative consolidation of the genre, following Edgar Allan Poe's proto-whodunits, beginning with Arthur Conan Doyle's "A Scandal in Bohemia" in 1891, which sparked a surge in popularity and established key archetypes like logical deduction and eccentric detectives. Although authors such as Poe and Wilkie Collins were pioneers of the mystery genre, their style pre-dates the structured whodunit conventions that emerged later. Their work focuses on psychological intrigue rather than the puzzle-solving focus of the selected range. Furthermore, limiting the date range to 1924 ensures all selected works are in the public domain, allowing for unrestricted access and reuse via sources like Project Gutenberg, while also aligning with the structural consistency of the time frame for computational analysis.

Finally, due to practical (time) limitations, only stories consisting of at least 2 000 and at most 10 000 words were selected. The narratives should be long enough to capture the full structure of whodunits. At the same time, the texts should remain short enough for practical manual analysis as the annotation process is time and expertise intensive. Note that only annotating snippets as annotated by the SANTA project⁷ and their extended data set⁸ (which could be an alternative for long stories) will not uncover the complete structure of the narratives.

To streamline the cleaning process, collections with similarly formatted texts were prioritised, such as The Project Gutenberg eBook of Great Short Stories, Volume 1, The Project Gutenberg eBook of Poirot Investigates, The Project Gutenberg eBook of The Adventures of Sherlock Holmes and The Project Gutenberg eBook of The League of the Scarlet Pimpernel. From these, several stories from Poirot Investigates and Sherlock Holmes were selected to provide multiple examples of the same detective, enhancing consistency in annotation. The remaining stories were chosen for their diverse detective styles, offering a representative sample of how detectives can be portrayed. If not limited by manual annotation costs, a broader representation of detectives would have been considered. Hopefully this can be expanded in future work.

⁶This task deviates from the timespan mentioned in Chapter 4. All text used in Chapter 4 still fall in this new timespan, but the starting point was moved a bit earlier to accommodate some influential earlier work.

⁷<https://github.com/SharedTasksInTheDH>

⁸<http://lsx-events.informatik.uni-wuerzburg.de/stss-2021/task.html>

7.2.3 Quality control

The annotation was performed by the author of this thesis, who developed the annotation guidelines presented in Chapter 5 and refined in Chapter 6. Therefore the annotator has an in-depth understanding of the theoretical and practical requirements of each tag.

Quality control was conducted systematically through an iterative, multi-phase process. First, scenes and non-scenes were annotated, followed by anachronic structures, diegetic levels, perspective, clues, and finally acts, which proved the least complex. This was done to ensure consistent annotation throughout the texts in the data set.

The annotation of the texts on the different diegetic levels (extradiegetic where the author or narrator is outside the story, intradiegetic, where the narrator is telling the main story, and metadiegetic where stories are told in stories) is challenging as discussed in Chapter 6, necessitating iterative refinement. After a small subset of texts was annotated with narrative levels, the annotator encountered a case that did not fit the initial tagging rules relating to the distinction between intradiegetic and metadiegetic. The guidelines were refined, and all four previously annotated text were re-annotated from the start, ensuring alignment with the correct definitions.

In addition, all texts underwent a complete re-reading to verify tag consistency, particularly for complex structures such as embedded analepses and character interjections.

7.3 Reuse potential

In the field of computational narratology, this data set can aid in refining algorithms for automatic narrative segmentation, plot structure analysis, and genre classification. The detailed annotations of narrative elements, such as scene boundaries, diegetic levels, acts, focalisation, and voice, alongside genre-specific tags like detective thoughts and reveals, provide a structured resource for training and evaluating computational models.

As surveyed in Chapter 5, few comparable data sets exist for full-text, genre-specific literary annotation, making this resource particularly valuable for research in computational narratology. For instance, scene and non-scene tags can be used to train supervised models to detect boundaries in unseen texts, similar to the SANTA project's annotation-based approaches for narrative levels (Reiter et al., 2019) while act and clue tags enable extraction of plot arcs for classification tasks, building on schema like those in Eisenberg and Finlayson (2021) for temporal disruptions. The annotations in this data set build on existing frameworks like those developed in the SANTA project for narrative level and scene segmentation (Reiter et al., 2019).

Scholars in the fields of literary studies, linguistics, and digital humanities can leverage this data set for genre studies, particularly to explore how mystery short stories employ

narrative techniques like temporal disruptions to build suspense, as discussed in Eisenberg and Finlayson (2021). It also supports cross-genre comparisons by offering a model for genre-specific annotations, enabling researchers to contrast whodunit structures with other narrative forms, such as modernist novels (Kearns, 2020).

Additionally, the data set can enhance applications like narrative generation and summarisation, where understanding plot structure and temporal flow is important (Droog-Hayes et al., 2018). For example, patterns in clues and reveals could inform generative models to produce suspenseful sequences or aid summarisation by prioritising key acts and diegetic shifts, extending SANTA-inspired frameworks for automated story understanding (Gius et al., 2021). While for educational purposes, it can serve as a teaching resource to illustrate narrative techniques in mystery short stories and contrast this to other genres.

One has to keep certain limitations of the data set into account, however. The focus on a single genre may restrict immediate applicability to other narrative types, such as fantasy or historical fiction. The reliance on a single annotator's interpretations may introduce subjectivity, potentially requiring validation or adaptation for different research contexts. Furthermore, the data set's scope, limited to short stories, may not fully capture the complexity of longer narratives, posing challenges for scaling to larger texts.

7.4 Conclusion

This chapter presented a fully annotated data set of ten public-domain mystery short stories, systematically segmented and tagged according to the refined annotation framework developed in Chapters 5 and 6. By extending the pilot annotations to a complete, genre-coherent corpus spanning key authors of the whodunit tradition (Christie, Doyle, Stevenson, and others), this study provides a structured resource that captures both universal narratological structures-scenes, non-scenes, diegetic levels, focalisation, anachronisms, and genre-specific dynamics such as act progression, clue placement, and detective cognition. The resulting 4 648 sentences and over 61 000 words, annotated with 511 high-level tags across 18 categories, represent one of the most comprehensive manually annotated collections for full-text computational narratology to date.

The data set's strength lies in its dual utility: it serves as data for training and evaluating transformer-based models for automated scene detection, narrative level classification, and plot structure extraction (as explored in Chapter 8), while simultaneously enabling literary scholars to conduct quantitative analyses of suspense-building mechanisms, temporal disruption patterns, and act-based narrative rhythm in the whodunit genre. Unlike segment-based data sets (for example, SANTA), our full-text approach preserves global narrative arcs, allowing investigation of how local cues aggregate into macro-level plot progression, which is an important step toward bridging close and distant reading.

Despite its focused scope, limited to early 20th-century English-language whodunits annotated by a single annotator, the data set is a replicable, modular resource that can be scaled to larger corpora or adapted to other genres. Future extensions may incorporate inter-annotator validation, semi-automatic pre-labeling via the models developed in Chapter 8, or cross-genre expansion to test the transferability of act and clue schemas of the framework. Ultimately, this resource not only supports immediate computational tasks but lays foundational infrastructure for a new generation of hybrid literary-computational inquiry into how stories are structured, perceived, and generated.

Chapter 8

Automatic Narrative Discourse Annotation Systems

This chapter explores the application of fine-tuned Robustly Optimized BERT Pretraining Approach (RoBERTa) models for multi-label classification of narrative discourse elements. Based on a submitted article Heyns, N., & van Zaanen, M. (n.d.). Automatic Narrative Annotation [Submitted to Computational Humanities Research (under review)] three systems are trained on the SANTA data set, a mystery-specific data set, and a combined version, with a few-shot Ollama baseline for comparison. The goal is to automate costly manual annotation, enabling scalable analysis of scenes, narrative levels, focalisation, and temporal shifts across large literary corpora.

8.1 Introduction

The segmentation and annotation of narrative discourse in literary texts provide important information in the context of computational literary analysis. This process, which identifies and labels distinct scenes and other narrative elements, is important for understanding the structure and pacing of literary works, providing insights into the author's stylistic choices and the overarching narrative framework.

Chapter 5 developed and applied a modular annotation scheme tailored to the unique narrative structure of mystery novels, particularly the whodunit genre. This scheme, based on consolidated guidelines from various sources, allowed for the systematic annotation of narrative elements, and also accommodates for the genre-specific phases and character roles that define whodunits. The result is a genre-specific annotated data set that complements existing resources like the SANTA¹ data set.

¹<https://github.com/SharedTasksInTheDH>

Building on this foundation, the present chapter contains the designs and implementation of a system for the automatic annotation of narrative elements such as scenes, narrative levels, and focalisation, by using the manually annotated text as training data. By automating the annotation process, this chapter attempts to extend the application of the modular annotation scheme to larger data sets, making the task more scalable and efficient.

Three models are developed, by training a fine-tuned RoBERTa model using three different data sets for multi-label narrative tagging. In addition, a baseline system is implemented using the Ollama model (AI@Meta, 2024) to provide a comparative perspective on a few-shot annotation performance. By examining models trained on distinct and combined data sets, as well as the Ollama baseline, it is assessed how data composition and model architecture influence performance and generalisation in a literary analysis application. Section 8.3.1 will discuss the three data sets in detail.

This chapter poses two research questions, originally introduced in Chapter 1: **RQ 6**: “Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?” and **RQ 7**: “How effectively can existing and enhanced machine-learning methods, applied to an annotated data set, identify and analyse high-level narrative structures?”

The contributions of this chapter are threefold. First, different annotated data sets are consolidated to be used for the training of our systems. Second, three systems and a baseline system are designed that automate the annotation of narrative properties. Third, the system variants are evaluated. In addition, a curiosity is addressed about whether a large language model (LLM) can effectively annotate literary data, given its potential as the easiest, most cost-effective, and approachable method. Assessing the performance of a few-shot approach with the Ollama model, provide insights into its viability for scalable annotation tasks.

8.2 Background

This section reviews the existing work on narrative segmentation, focusing on two key areas: annotation guidelines and methods for automatic scene and narrative level detection.

8.2.1 Existing annotated data

Creating annotated datasets with clear annotation guidelines for narrative segmentation have been the focus of several research efforts, although only a few data sets are publicly available. Among these, two significant contributions come from the SANTA project.

The first being Gius, Reiter, and Marcus (2019), which aimed to develop comprehensive guidelines for annotating narrative structure based on narratology. In Section 2.4.2 we mention the project's participants include Barth (2019), Bauer and Lahrsow (2020), Eisenberg and Finlayson (2021), Hammond (2021), Kearns (2019), Ketschik et al. (2019), and Wirén and Ek (2021), who contributed to the creation and extension of these guidelines.

The second contribution comes in the form of the annotated data. The SANTA project established specific guidelines for annotating narrative levels, applying these guidelines to a corpus² that was made available for a shared task. These annotated narratives, presented in XML and CSV format, marked a step forward in publicly accessible data sets for narrative segmentation. This data set will be used in this chapter. An additional data set³ was made available in JSON format, encompassing 22 novels, including the annotated texts that contributed to the SANTA data set, thereby expanding the range of accessible resources. However, due to the presence of copyrighted material and the further effort to convert the JSON files into CSV format, this data set will be excluded from the current study.

Another significant effort was discussed in Section 2.4.2 in the form of Gaizauskas and Alrashid (2019), who introduced SceneML for annotating scenes, locations, characters, and time within narratives. Initially relying on XML, the project later adopted the Brat Annotation Tool (Alrashid & Gaizauskas, 2021) for a more intuitive annotation experience. However, despite its promise, SceneML lacked clarity on how to handle narrative description and multiple levels within the narrative, which limited its broader application, therefore this chapter will not use this data set.

Unfortunately, several other data sets mentioned by Chung et al. (2018), Kearns (2020), Newberry and Bailey (2019), and Rogers et al. (2024) remain unavailable to the public, limiting access to diverse annotation examples.

Building on the efforts of the SANTA project, Chapter 5 addresses the ambiguities in existing guidelines where authors used similar terms for different concepts or different terms for similar concepts. This work combined these contributions into a new, modular set of guidelines and introduced genre-specific annotations tailored to mystery novels. This allows for a more consistent approach to annotating both general narrative elements and the unique features of mystery narratives. This chapter will use this data set as well as the SANTA data set.

8.2.2 Existing methods

In recent years, several methods have been developed for automatic narrative discourse detection, each offering unique insights into how narratives can be computationally structured. This section will focus on methods that use annotated guidelines.

²<https://github.com/SharedTasksInTheDH>

³<https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/German-Scene-dataset>

Most work came from the second part of the SANTA project (the STSS). Participants explored various approaches to identifying shifts in narrative content, with many using sentence embeddings, or temporal, spatial, and entity features to determine scene boundaries.

Schneider et al. (2021) based their method on detecting significant changes in word embeddings, hypothesising that large shifts in semantic content between segments of text often correspond to scene transitions. By measuring these shifts over time, their model effectively pinpointed scene boundaries. A more complex approach was taken by Gombert (2021), who employed a twin BERT architecture to create a space where scene-bordering sentences could be distinguished from those within a scene. This model leveraged sentence embeddings as features for a gradient-boosted decision tree ensemble to predict scene changes, ranking highly in the STSS leaderboard.

Other participants, such as Barth and Dönicke (2021), focused on temporal, spatial, and entity-based features, building a system that detected only transitions between scenes but struggled to differentiate non-scenes due to a limited training set. In addition, Hatzel and Biemann (2021) used a transformer model enhanced with CR features, showing the importance of tracking characters across sentences to detect scene boundaries accurately.

While this review focuses on the key systems developed during the SANTA project (those most relevant to our current study) it is acknowledged that other methods may exist. However, the studies mentioned here are the only research efforts known to this author that annotated narratives to such an extent and with guidelines stemming from literary research. Our selection of articles also reflects a deliberate engagement with the STSS context, aligning our research with the ongoing dialogue within this shared task community.

8.3 Methodology

This section will develop models based on RoBERTa (Liu et al., 2019) to automatically annotate narrative structure information⁴. RoBERTa will be trained on three different data sets, the SANTA data set, and mystery data set from Chapter 7, and a combined version of both data sets. To train a model on a combined data set, which integrates the mystery narrative data set with the SANTA data set, the two annotation schemes were consolidated to enable the model to learn consistent narrative patterns across texts. Section 8.3.1 will explain the consolidation of the annotation schemes and Section 8.3.2 will explain the fine-tuning of the models.

To provide a comparative baseline for the fine-tuned RoBERTa models, the results are compared against an Ollama model, which is a large language model (LLM), utilising a few-shot learning approach for narrative annotation. This method leverages a small

⁴Replication data and code can be found on the following Github repository:
<https://github.com/Nuette/Automatic-narrative-annotation>.

set of annotated examples to guide the model’s predictions, offering an alternative to supervised fine-tuning. This approach is potentially more accessible and cost-effective. A structured prompt template (see Appendix C) was designed that includes definitions of key narratological elements (such as scenes, narrative levels, and focalisation) along with example annotations from the mystery data set. The prompt is tailored to elicit CSV-formatted outputs matching the annotation scheme, ensuring consistency with the RoBERTa training data. Details of the prompt design and example selection are discussed in Section 8.3.3, while the evaluation of Ollama’s performance is covered in Section 8.5, along with the performance of the fine-tuned models.

8.3.1 Data sets

The usefulness of a data set for this task of high-level structure analysis is determined by the annotation scheme used. If the annotation schemes of the data sets are different, then the data sets cannot simply be combined. The annotation schemes will first need to be consolidated. This section describes the consolidation process and the resulting unified data set.

Consolidating the existing annotation schemes (used in the SANTA project and the mysteries data set) is challenging, because each SANTA article employs a different annotation scheme with varying terminologies and levels of detail. To develop consolidated tags, the SANTA guidelines were referenced as well as the mystery narrative guidelines and the annotation types were unified as much as possible. For instance, the mystery narrative data’s tag `Narrative_level` includes properties for extradiegetic, intradiegetic, and metadiegetic levels, but not all SANTA articles annotate narrative levels, or they only cover extradiegetic and intradiegetic levels. To address this, a unified tagset was created, mapping equivalent concepts across the SANTA schemes to the mystery narrative data set. Table 8.1 illustrates this mapping, detailing how each SANTA article’s tag aligns with the consolidated tagset. This highlights both direct equivalences and gaps where certain SANTA articles omit specific concepts. Each row in Table 8.1 corresponds to one of the consolidated tags, with columns showing the equivalent or related tags (if any) from the SANTA project articles⁵.

The SANTA data, stored in CSV format, includes columns for the annotator’s initials, tag name, and the start and end positions of the text. These tags were mapped to the appropriate columns in the consolidated data set, using binary values to indicate their presence, see Table 8.1 for details. The annotator initials were also kept in the data set. In cases where the same annotator assigned multiple tags to the same text span (for example, narrative level and focalisation), these were merged into a single row with binary

⁵As the SANTA project was a shared task multiple articles were published relating the authors own guidelines that they annotated the text according to. There were eight articles in total that took part in the project, but the third article was withdrawn. Therefore their guidelines and data were not used during this article.

values to reflect the presence of each tag. In addition, the text was extracted from the XML files and start and end positions were replaced with the actual text. The consolidated data sets are available in a Github repository⁶.

Table 8.1: Overview of the consolidated tag (in the left-most column) and the corresponding tags from the different SANTA articles.

Tag	SANTA 1	SANTA 2	SANTA 4	SANTA 5
Scene	original	none	1	none
Summary	none	none	n	none
Descriptive_passage	none	none	none	none
Analepsis	none	analepsis	none	none
Prolepsis	none	prolepsis	none	none
Extradiegetic	embedded_1	level degree=1	1	l1
Intradiegetic	embedded_2	level degree=2	2	l2
Metadiegetic	none	none	3	l3
Focalisation	none	fid_pers={character_name}	none	none
Voice_homodiegetic	none	none	none	n1
Voice_heterodiegetic	none	none	none	n2
Tag	SANTA 6	SANTA 7	SANTA 8	
Scene	none	characters_0	none	
Summary	none	nc	none	
Descriptive_passage	none	narrator_0	none	
Analepsis	none	none	none	
Prolepsis	none	none	none	
Extradiegetic	narrative_level number=1	none	a	
Intradiegetic	narrative_level number=2	turn	b	
Metadiegetic	narrative_level number=3	none	c	
Focalisation	none	foc_unr	none	
Voice_homodiegetic	narrator_participation=autodiegetic	none	none	
Voice_heterodiegetic	narrator_participation=heterodiegetic	voice_3	none	

The consolidation of the data sets allows for the creation of a unified annotation scheme, but it also reveals potential biases stemming from the distinct characteristics of the source materials, particularly in terms of genre and era representation. The SANTA data set, developed to encompass a broad range of narrative phenomena, is derived from a heterogeneous corpus that includes a variety of genres, publication dates, and text lengths.

⁶<https://github.com/Nuette/Automatic-narrative-annotation>

In contrast, the mystery data set is drawn exclusively from the golden era of mystery narratives (approximately around 1900), comprising classic whodunits characterised by relatively structured plot arcs. The mystery data set is also significantly smaller than the SANTA data set, raising the risk of biasing the model toward the SANTA data's broader patterns. The SANTA data set comprises 2 343 text segments derived from thirteen unique texts, annotated across seven distinct annotation sets, each reflecting a different focus by the contributing article. In contrast, the mystery data set contains 289 segments from ten unique texts. In addition, the SANTA data annotates shorter segments of texts, with on average 7 468.45 words and 470.68 sentences per text, and segments contain 290.19 words on average. The mystery data consists of full length mystery short stories averaging 1 030.20 sentences and 12 906.60 words per text. Segments contain 449.71 words on average.

The difference in sample lengths and volumes, raises concerns that the models may be disproportionately influenced by the larger volume of SANTA data, potentially skewing performance toward its broader and less genre-specific patterns. This concern will be discussed in more detail in Section 8.4 and Section 8.5, where its impact on the model performance is explored. While the long-term goal is to enhance model robustness through more balanced sampling across eras and genres in future work, the current focus is to explore whether padding the existing data with genre-specific annotations can improve the models' performance. This initial investigation aims to assess the potential benefits of tailoring the data set to the whodunit genre before expanding its scope.

8.3.2 Model development

This chapter presents the fine-tuning of RoBERTa (Liu et al., 2019) for the task of multi-label classification of high-level narrative elements. RoBERTa is a transformer-based language model that builds directly on BERT but improves its pre-training through several optimisations. It is trained on a larger corpus (160 GB of text, including BooksCorpus, CC-News, OpenWebText, and Stories), uses dynamic masking (where masked tokens change in each training epoch), removes the next-sentence prediction objective, and is trained with larger batch sizes and for more steps. These changes result in significantly better performance on downstream NLP tasks, particularly in understanding long-range dependencies and nuanced linguistic patterns essential for detecting subtle narrative structures in literary texts.

All models are fine-tuned with consistent hyper-parameters: 20 epochs (with early stopping patience of 5), a batch size of 16, warmup steps of 500, and a weight decay of 0.01. Evaluation is performed every 100 steps using the Adam optimiser evaluating the F1 metric. The training utilises the Adam optimiser within the Hugging Face Transformers library, executed in a Python environment. Data pre-processing includes tokenisation with a maximum length of 128 tokens using the RoBERTa tokeniser from the FacebookAI/roberta-base model, ensuring manageable memory usage and consistency across data sets. Training

is conducted on a single device, using PyTorch with the accelerate library (version 0.26.0) to optimise performance. The RobertaForSequenceClassification model is initialised and configured for multi-label classification where each token in a sentence can be assigned zero or more narrative labels (see Table 7.2). This is achieved by adding a linear classification layer on top of the final hidden states and training the entire model end-to-end using binary cross-entropy loss across all labels.

A five-fold cross-validation strategy is employed to robustly assess model performance across distinct folds of the data, ensuring the evaluation of generalisation. This process is applied separately for the three data sets: the SANTA data set, the mystery data set, and a combined data set integrating both, allowing us to evaluate the impact of data composition on model effectiveness. Each fold's training and evaluation data sets are created using the data sets library, with features defined as text (string) and labels (sequence of float32 values).

8.3.3 Baseline model

To establish a robust baseline, a structured prompt template was developed that leverages the modular guidelines from Chapter 5, tailored specifically to the whodunit genre. This template instructs the Ollama hosted Meta Llama 3.1 8B⁷ (an instruction-tuned variant of a large language model (LLM)) to act as an expert in narratology and produce annotations in a CSV format with the same columns as the training data created in the data sets section.

The prompt begins with definitions for each narratological tag as defined in Chapter 5, providing clear criteria for annotation. These definitions ensure the model aligns with our annotation scheme, maintaining consistency with the manually annotated training data used for the RoBERTa models.

To enhance the few-shot learning capability, three carefully selected examples from the mystery data set are incorporated into the prompt. These examples, showcase diverse narrative segments with corresponding annotations, such as scenes, temporal shifts or different narrative levels. The model is prompt to output a single CSV row without additional comments. The prompt as well as the examples are given in Appendix C.

The temperature parameter is set to 0 in the Ollama API call to minimise randomness and enforce deterministic responses.

The final prompt file, which encapsulates this design, is available alongside the code for the fine-tuned RoBERTa models in the project repository⁸, promoting transparency and enabling replication of our experiments. This approach allows researchers to adapt the prompt for other data sets or annotation tasks, contributing to the broader field of computational narratology.

⁷<https://ollama.com/library/llama3.1:8b>

⁸<https://github.com/Nuette/Automatic-narrative-annotation>

For practical purposes, the LLM setup is comparable to the fine-tuned RoBERTa models, since both are executing multi-label classification on the same narrative texts. The Ollama baseline highlights the advantage of data-specific fine-tuning through the Ollama model's dependency on the few-shot examples versus the RoBERTa model's extensive fine-tuning on the annotated training data.

To evaluate the baseline's generalisation, the Ollama model was tasked with annotating all rows in the combined data set, excluding the three initial examples provided in the prompt. This approach tests the model's ability to extrapolate beyond the given examples, providing an assessment of its few-shot learning performance on unseen data. This extensive application highlights the baseline's limitations when scaling to a larger, diverse corpus, reinforcing the value of the fine-tuned data-driven approach of the RoBERTa models.

8.4 Implementation

This chapter trained models using the three consolidated data sets discussed in Section 8.3.1 using a five-fold cross-validation approach to ensure robust performance across different data partitions.

The SANTA model was trained on the SANTA data only. Training started with a random seed of 1 for reproducibility. Each fold took approximately three hours to train and covered on average 1 400 out of 2 300 steps across 12 out of 30 epochs. The total training time per fold varied due to data set size and computational load. Early stopping triggered after five epochs of no micro-averaged F1 score improvement, preventing further training when performance plateaus, to avoid overfitting.

Each fold of the mystery model was trained on the mystery data set only and took around 30 minutes to train, with early stopping determining the optimum performance of the model on step 300 for 20 epochs. Compared to the SANTA model, the faster convergence can be explained by the smaller size of the data set.

The combined model was trained on the combined data set (consisting of the SANTA and the mystery data sets). The training took approximately four hours per fold, with early stopping activating after around 1 800 steps and 13 epochs, which reflects the larger data set size compared to the SANTA data set and mystery data set.

The significantly faster training time of the mystery model, averaging around 30 minutes per fold, compared to three hours for the SANTA model and four hours for the combined model, can primarily be attributed to the smaller size of 289 segments of the mystery data set. For comparison, the SANTA data set has 2 343 segments and the combined data set has 2 632 segments. The SANTA model thus has fewer samples to process during each epoch and therefore fewer overall iterations before convergence or early stopping is triggered. However, in addition to the smaller data set, other potential contributing factors

to the differences in training speed include differences in data complexity. The text of the mystery data set spans only four authors and is characterised by the structured nature of golden-era whodunits. This data set likely exhibits less variation in text spans, reducing the complexity of tokenisation and forward passes through the RoBERTa model.

A notable challenge in this implementation arises from the tokenisation limit of 128 tokens imposed by the Hugging Face Transformers library. The mystery data set, which contains full length mystery short stories with an average of 449.71 words per segment, is particularly affected. Truncation of the text may lead to the discarding of important narrative context beyond the limit. This could impact the model's ability to accurately detect narrative elements such as scene boundaries or focalisation in longer segments, where key events or shifts might occur outside the tokenised range. While this limitation is less pronounced in the SANTA data set due to its shorter segment lengths of 290.19 words per segment, it remains a concern for the SANTA data as well. Future iterations should explore strategies such as segmenting longer texts into overlapping chunks or employing models with extended context windows to preserve essential information.

8.5 Evaluation

A five-fold cross-validation strategy is employed to robustly assess model performance across distinct folds of the data, ensuring the evaluation of generalisation. This process is applied separately to the three RoBERTa models, trained on the mystery data set, the SANTA data set, and the combined data set. Each fold's training and evaluation data sets are created using the data sets library, with features defined as text and labels.

The performance of the models is evaluated using a combination of metrics, including micro-averaged F1 score, ROC-AUC score, and accuracy, computed at the end of each evaluation step. These metrics provide a view of the models' ability to classify narrative tags across the multi-label framework, with F1 balancing precision and recall, ROC-AUC assessing discriminative power, and accuracy reflecting overall correctness.

The first model, trained on the SANTA data set with 11 labels, exhibits varying performance across its five-fold cross-validation, reflecting the data set's 2 343 segments and diverse narrative structures. The results of the best metrics per fold, along with their averages and standard deviations for the SANTA model, are summarised in Table 8.2.

The best-performing steps for each fold highlight the ability of the model to adapt. Fold 2 recorded the overall best F1 score of 71.78 at step 600, alongside a ROC-AUC of 92.76 and accuracy of 56.50, finished in two hours and 17 minutes over nine epochs.

The variability in training times (ranging from two to four hours) and peak metrics across folds underscores differences in data partitioning and computational demands, with early stopping effectively curbing overfitting as validation loss rose in later steps.

The progressive improvement in F1 scores across folds, supported by declining training losses, reflects effective learning, though the increasing validation loss in later steps hints at potential overfitting, which the early stopping mechanism helps mitigate.

Overall, the model demonstrates robust adaptation to the SANTA data set's narrative tag prediction task. The performance of the model is summarised by average metrics across folds, with an average F1 score of 69.89 (stddev 0.0202), ROC-AUC of 92.10 (stddev 0.0094), and accuracy of 54.14 (stddev 0.0159). These averages reflect consistent label prediction and strong discriminative power, whereas the low standard deviations indicate stable performance across folds.

Table 8.2: Best metrics per fold for the SANTA model.

Fold	Best Step	F1 Score	ROC-AUC	Accuracy
1	1200	68.70	91.12	53.30
2	600	71.78	92.76	56.50
3	600	71.00	92.77	54.70
4	1200	66.49	90.80	51.71
5	1300	71.50	93.05	54.49
Average	—	69.89	92.10	54.14
Stddev	—	0.0202	0.0094	0.0159

For the second model, trained on the mystery data set, the results of the best metrics per fold are summarised in Table 8.3. Fold 3 stands out, achieving the highest F1 score of 60.16 at step 200, accompanied by a ROC-AUC of 85.51 and an accuracy of 56.14, completed in approximately 30 minutes. This fold suggests the model effectively captures genre-specific traits, such as the structured plot arcs typical of golden-era mystery narratives, though the performance remains modest given the limited scope of the data set.

Rapid convergence is also evident in fold 2, with an F1 score of 59.20 at step 200, a standout ROC-AUC of 91.06, and an accuracy of 53.45, reflecting efficient learning on the compact data set. The smaller size and reduced complexity, driven by the formulaic nature of mystery plots, enable quick optimisation, allowing the model to adapt to predictable patterns without extensive iterations.

Training consistently halts at step 300 across all folds, a behaviour attributed to `EarlyStoppingCallback`, which terminates training when the F1 score plateaus for five consecutive evaluations (spanning 500 steps with `eval_steps=100`). However, the observed stop at step 300, corresponding to the third evaluation, indicates that the F1 score likely peaks early (for example, at step 200) and remains stagnant through step 300, triggering early stopping after only two non-improving evaluations. This premature cessation reflects the limited learning potential of the data set, where the structured simplicity and small volume lead to early saturation of performance gains. The average F1 score of 59.02 (stddev 0.0085) and the accuracy of 52.71 (std 0.0229) across folds, as shown in Table 8.3,

underscores this stability, although the higher ROC-AUC variability 87.82 (stddev 0.0270) suggests inconsistent discrimination ability.

The limited step range (only three evaluations) raises questions about the generalisation of the model compared to the more extensive evaluations of the SANTA and combined models (discussed after this), as the smaller data set may not fully represent the broader narrative variability seen in larger corpora.

The quick adaptation of the model is further supported by declining training losses, yet the rising validation loss in later steps indicates a risk of overfitting, managed by the early stopping mechanism after minimal epochs. This efficiency, driven by the compact 289 segment data set, contrasts with the longer training demands of larger data sets, offering a trade-off between speed and learning.

Table 8.3: Best metrics per fold for the mystery model.

Fold	Best Step	F1 Score	ROC-AUC	Accuracy
1	300	58.99	84.21	51.72
2	200	59.20	91.06	53.45
3	200	60.16	85.51	56.14
4	300	57.52	87.75	53.12
5	300	59.21	90.59	49.12
Average	—	59.02	87.82	52.71
Stddev	—	0.0085	0.0270	0.0229

The results of the best metrics per fold for the combined model are summarised in Table 8.4. The combined model, trained on the combined data set, demonstrates a robust performance across its five-fold cross-validation, with fold 4 achieving the highest F1 score of 70.04 at step 1 200, alongside a ROC-AUC of 94.15 and an accuracy of 54.10, completed in three hours and 11 minutes. This fold highlights the ability of the model to leverage the diverse patterns within the merged data set, although it falls slightly short of the peak performance of the SANTA model. Since the SANTA model was trained solely on the SANTA data, this suggests a trade-off in specialisation. This gap may reflect the focus of the SANTA model optimisation on the unique narrative structures of SANTA, whereas the combined model must generalise across both data sets, potentially diluting its precision. In addition, annotation consistency could contribute, as the SANTA data, annotated collaboratively in a shared task, likely benefited from extensive guideline discussions ensuring uniform tag interpretations. In contrast, the mystery data, based on these guidelines but annotated independently, may exhibit slight interpretive differences. This variability could introduce noise into the training of the combined model, challenging its ability to reconcile divergent label distributions. Regardless, fold 3 also stands out with a ROC-AUC of 95.32 and accuracy of 59.86, reflecting the strong discriminative power of the model.

The average F1 score across folds reaches approximately 68.60, with a consistent ROC-AUC of 94.01 and an average accuracy of 55.51, as detailed in Table 8.4. These results indicate consistent performance, with notable variability in accuracy hinting at data set heterogeneity.

The varied step counts for peak performance (ranging from 800 to 1 800) suggest that the increased data complexity and volume necessitate longer optimisation, with the model adapting to a broader range of narrative structures, compared to the mystery model. However, the plateauing F1 scores across folds indicate that additional epochs may offer diminishing returns without further hyper-parameter tuning.

This extended training, compared to the rapid convergence of the mystery model on a smaller data set, underscores a trade-off between learning and computational demand, with the combined data set enabling stronger generalisation potential.

Table 8.4: Best metrics per fold for the combined model.

Fold	Best Step	F1 Score	ROC-AUC	Accuracy
1	1800	66.92	92.95	51.71
2	900	69.62	93.61	59.67
3	800	68.21	95.32	59.86
4	1200	70.04	94.15	54.10
5	1400	68.82	94.02	52.19
Average	—	68.60	94.01	55.51
Stddev	—	0.0131	0.0078	0.0357

As shown in Table 8.5, the baseline model, relying on limited examples without extensive training, achieves an F1 score of 44.07, a ROC-AUC of 65.57, and an accuracy of 16.67. These results highlight a significant performance gap compared to the fine-tuned models, with the F1 score, ROC-AUC and accuracy falling well below the averages of the SANTA and combined models summarised in Table 8.5. This disparity underscores the baseline's limited ability to capture complex narrative patterns, reflecting the inefficiency of few-shot learning in this context where detailed annotations substantially enhance model performance. The low metrics of the baseline suggest that, despite the reduced annotation effort, the trade-off heavily favors full annotation for achieving robust classification outcomes in narratology tasks.

Table 8.5: Average performance metrics for each model and the baseline.

Metric	SANTA	Mystery	Combined	Baseline
F1 Score	69.89	59.02	68.60	44.07
ROC-AUC	92.10	87.82	94.01	65.57
Accuracy	54.14	52.71	55.51	16.67
Average Training Time per Fold (minutes)	180	30	270	—

Collectively, these results highlight the trade-offs between data set size, training duration, and performance optimisation across the three models. The SANTA model's extensive training times (ranging from two to four hours per fold) underscore the computational cost of handling 2 343 segments, yet the peak F1 scores (up to 71.00) suggest significant gains from fine-tuning on a large, diverse corpus. In contrast, the shorter training times for the mystery data set (averaging around 30 minutes) may yield lower but more rapid improvements, reflecting its smaller size of 289 segments. The results of the combined data set further clarify whether increased data volume enhances generalisation, though at the expense of extended training durations. This trade-off between annotation effort and accuracy gains can be assessed by comparing these fine-tuned models with a few-shot baseline to determine the cost-effectiveness of full annotations versus few-shot approaches.

8.6 Discussion

The comparative analysis of the three fine-tuned RoBERTa models reveals significant insights into the influence of data set composition and training dynamics on narrative tagging performance.

The SANTA model achieves an average F1 score of 69.89, the highest among the models, reflecting its specialised adaptation to the SANTA corpus's unique narrative structures, supported by an average ROC-AUC of 92.10 and accuracy of 54.14 over 180 minutes per fold. However, the increasing validation loss suggests a potential overfitting, managed by the early stopping mechanism, indicating that whereas the SANTA data set offers a strong foundation, its complexity may limit further gains without additional regularisation or data augmentation.

The mystery model delivers an average F1 score of 59.02, ROC-AUC of 87.82, and accuracy of 52.71, completed in just 30 minutes per fold, highlighting its efficiency due to the smaller data set size and simpler, formulaic mystery patterns. However, the rapid convergence and limited evaluation steps, suggest overfitting to these patterns, potentially compromising generalisation to broader contexts, despite a respectable discriminative ability.

The combined model achieves an average F1 score of 68.60, with a leading ROC-AUC of 94.01 and accuracy of 55.51 over 270 minutes per fold, reflecting the benefits of diverse data integration and extended training. The plateauing F1 score indicates a performance ceiling within current hyper-parameters, yet the improved ROC-AUC and accuracy underscore enhanced generalisation, though at a higher computational cost.

This performance landscape extends to a comparison with the baseline model, which, relying on a few-shot approach with minimal annotation effort, yields an average F1 score of 44.07, ROC-AUC of 65.57, and accuracy of 16.67. The substantial gap between

the baseline and fine-tuned models, particularly the baseline's lower F1 and ROC-AUC compared to the SANTA and combined models, highlights the baseline's limited capacity to capture complex narrative patterns, affirming the cost-effectiveness of full annotations for robust classification. The SANTA model's superior F1 score reflects its specialised precision, while the combined model's higher ROC-AUC and accuracy demonstrate the advantage of broader data exposure, though its longer training time (270 minutes) contrasts with the implicit efficiency of the baseline.

Overall, these results underscore the trade-offs between data set size, training duration, and generalisation, with the baseline serving as a benchmark for minimal-effort scenarios. In contrast the fine-tuned models, especially the combined model, suggest that integrating diverse data sets enhances performance despite increased computational demand.

8.7 Future work

There are several avenues that can be explored for future work. Given the influence of the amount of data available, one avenue to explore for future work is to use larger or more specialised pre-trained models, such as RoBERTa-large.

On a practical level, given the identified limitation of the Hugging Face Transformers library's 128 token maximum, future research could investigate models with extended context windows, such as Longformer or BigBird, to accommodate longer text segments, which is especially important for the mystery data set where some segments exceed 400 words. This extension would preserve essential narrative context, such as extended dialogues or temporal shifts, which are currently truncated, potentially improving annotation accuracy for tags like analepsis or focalisation.

In addition, incorporating advanced regularisation techniques could address overfitting observed in later training steps, particularly for additional columns such as acts or character presence, which may introduce further complexity. Exploring ensemble methods, where predictions from the SANTA, mystery, and combined models are integrated, could leverage the strengths of each data set, diverse patterns from SANTA's 2 343 segments, rapid convergence from mystery's 289 segments, and balanced generalisation from the combined set, potentially yielding superior results.

Hybrid approaches blending rule-based narratology rules with AI predictions could also enhance performance, combining human expertise with model-driven insights to refine tag assignments.

Finally, developing interactive tools for literary scholars would broaden the practical impact of this work. Integrating these models with visualisation platforms like Voyant Tools could enable real-time annotation, exploration of narrative structures, and comparative analysis across data sets, facilitating adoption in digital humanities workflows. Such tools could also support iterative feedback loops, allowing scholars to adjust token limits or segment

strategies based on observed performance, further aligning AI capabilities with literary analysis needs.

8.8 Conclusion

Manual annotation of structural aspects of literary narratives is a time and expertise intensive task. However, such annotation is essential to investigate how people tell stories, among other things. Automatic analyses (even as a pre-processing step for manual correction) may help this annotation process.

This chapter investigated the performance of a number of such automated systems that are trained on different data sets, the larger SANTA data set, the small mystery data set, and a combined data set. The results are also compared against a LLM baseline. The results of the different systems indicate that in general, a fine-tuned RoBERTa model is well-suited for narrative tag prediction, with potential for refinement through additional data, hyper-parameter optimisation, and advanced modelling techniques.

The average performance metrics reveal the models' adaptability to varying narrative structures despite challenges, such as the Hugging Face Transformers' 128-token limit, which truncates longer texts like those in the mystery data set exceeding 2 000 words. The SANTA model achieves the highest average F1 score of 69.89 (stddev 0.0202), reflecting its specialised precision on this diverse corpus, with an average ROC-AUC of 92.10 (stddev 0.0094) and accuracy of 54.14 (stddev 0.0159) over 180 minutes per fold. The mystery model delivers an average F1 score of 59.02 (stddev 0.0085), ROC-AUC of 87.82 (stddev 0.0270), and accuracy of 52.71 (stddev 0.0229) in just 30 minutes per fold, highlighting efficiency but limited generalisation due to the smaller, formulaic data set. The combined model strikes a balance with an average F1 score of 68.60 (stddev 0.0131), the highest ROC-AUC of 94.01 (stddev 0.0078), and accuracy of 55.51 (stddev 0.0357) over 270 minutes per fold, showcasing the benefits of data integration for enhanced discriminative power and generalisation. In contrast, the baseline model, relying on minimal annotation, yields a notably lower average F1 score of 44.07, ROC-AUC of 65.57, and accuracy of 16.67, underscoring the important role of extensive annotations in achieving robust performance.

This study directly addresses two research questions posed in Chapter 1 and highlighted in Section 8.1.

To address **RQ 6**: "Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?", this chapter demonstrates that purpose-built, fine-tuned RoBERTa systems trained on consolidated, genre-aware data sets, serve as effective entry points for automating narrative discourse analysis, laying a scalable foundation for future extensions to additional structural dimensions (for example, character roles, acts).

To address **RQ 7**: “How effectively can existing and enhanced machine-learning methods, applied to an annotated data set, identify and analyse high-level narrative structures?”, this chapter shows that existing and enhanced machine-learning methods, when applied to our refined data set, effectively identify narrative structures, with the combined model achieving strong generalisation (F1 68.60, ROC-AUC 94.01) and outperforming both the few-shot LLM baseline and genre-specific variants.

This study affirms the effectiveness of fine-tuned RoBERTa models for narrative tagging, reinforcing the power of transfer learning in computational narratology. The baseline results emphasises the cost-effectiveness of full annotation over few-shot approaches.

The variability in training times and the observed overfitting, mitigated by early stopping, suggest that further refinement is feasible through hyper-parameter tuning, such as adjusting learning rates or batch sizes, and incorporating advanced regularisation to handle additional columns like acts or character presence.

Future work could explore larger pre-trained models with extended context windows, such as Longformer or BigBird, to address the token limitation and preserve narrative context in longer segments, potentially boosting performance on downstream tasks like focalisation or analepsis annotation. Ensemble methods, combining predictions from all three models, could harness their complementary strengths, while hybrid approaches integrating rule-based narratology with AI could enhance precision.

By bridging AI capabilities with literary analysis, this work advances the evolving field of digital humanities, enabling more scalable and insightful examinations of narrative structures across eras and genres. The models’ performance lays a foundation for automated tools that augment human interpretation, fostering deeper engagement with literary heritage and paving the way for broader applications in text analysis and cultural preservation.

Chapter 9

Discussion

9.1 Introduction

This thesis explored the computational analysis of high-level narrative structures. A narrative constitutes two aspects: the temporal-causal ordering that defines the story (*histoire*) and the narrative discourse (*récit*). Part I of this thesis focused on the first by investigating the identification and annotation of story (in the form of narrative arcs). Part II focused on narrative discourse, for example, scenes and levels.

By addressing the objectives outlined in the introduction, the work contributed to bridging the gap between narratological theory and computational methods, demonstrating both the potential and limitations of automated approaches in literary studies. This chapter synthesises the key findings, relating them to existing literature, discussing implications and limitations, and proposing avenues for future work.

9.2 Objectives and key findings

The primary objective of this research was to develop and evaluate computational systems for detecting high-level narrative structures, while establishing standardised annotation guidelines to support these efforts. This was achieved in multiple stages. Chapter 2 provided a literature review to map the field, and two forms of high-level narrative structure were identified to investigate further: narrative arcs (story) and narrative discourse. Systems for arc identification that address story (*histoire*), and scene and narrative level identification address narrative discourse. In addition, the need for a standardised annotation format and guidelines was identified.

The literature review in Chapter 2 revealed a fragmented landscape of terminology and approaches. Terms like “plot”, “narrative level”, “scene”, and “temporal aspect” are often used inconsistently, drawing from theorists such as Freytag (1894), Genette et al. (1980),

and Propp (1968). The correlation between terms highlighted the interconnectedness of narrative elements, for example, scenes often aligned with temporal shifts.

Annotating strategies varied greatly, ranging from simple annotations in CSV files to complex XML TEI annotations. Some efforts emphasised IAA through guidelines, although the guidelines also varied greatly from annotation to annotation, resulting in small, fragmented data sets that cannot be combined easily.

Chapter 2 achieved two of the objectives of this thesis. Narratological concepts and existing computational approaches were investigated. Computationally under-explored aspects of high-level narrative structure identification were identified.

Part I, conceived plot as thematic narrative arcs (macro-level shifts in story topics that trace the causal progression of events), motivated by the claim that sentiment-based approaches to arc identification mix emotional tone with causal structure. This thesis argued that topic transitions directly reflected shifts in *what happens*, the core of *histoire*. Part I focused on thematic narrative arcs. An annotation format and also a small annotated data set consisting of four texts were created to support model training and evaluation in Chapter 4. Existing methods to computationally detect and analyse narrative structures were investigated, and computational techniques to automate the identification of narrative structure were applied in Chapter 3. This chapter introduced the STI, RSTI, and MTI systems that use LDA to identify the boundaries between narrative parts.

The STI system (Single Transition Identification) identified a single boundary in a text by first dividing the text into snippets and assigning an LDA topic to each snippet, producing a sequence of topics. It then computed the intersection of topics before and after each possible boundary position and selected the position with the minimal intersection as the transition point, indicating a significant topic shift.

The STI system excelled in detecting single boundaries with low LDA topics (RMSE of 0.0 for two topics), but performance declined with more topics due to increased topic fragmentation. The main limitation of this system was that it could only identify a single boundary in a text. Real text contains multiple boundaries, therefore the STI system was extended with the RSTI system so that it could detect multiple boundaries.

The RSTI system started out the same as the STI system. The text was divided into snippets, LDA assigned a topic to each, and a sequence of topics was produced. However, instead of using intersection, the system computed entropy at each possible boundary position. The position with the lowest entropy was chosen as the first transition. Once found, the text was split into two parts at this point. This process was then repeated independently on each half to find the next transition. This recursion continued, allowing the system to identify more than one boundary. If multiple positions tied for minimum entropy, one was selected at random. This approach considered topics on both sides of the boundary and therefore better captured narrative shifts compared to the original STI system described. Although the RSTI system successfully extended the original STI

approach to detect multiple boundaries through recursive splitting, it inherited limitations from the STI system's design. The original STI system was developed to identify a single transition, and recursive application may not optimally capture complex, non-linear narrative structures in real literary texts.

An alternative, the MTI system, was developed with the intent of identifying multiple transitions in a text. The MTI system was a collation-based clustering method that could handle multiple transitions in a text effectively, identifying transitions where topic groups shift. These findings aligned with prior work (Blei et al., 2003), but innovated by applying LDA extrinsically to boundary detection rather than intrinsic topic coherence.

Up to this point all systems were evaluated with data created synthetically by concatenating snippets of different text to simulate a transition point in the text. To evaluate the systems, RMSE was used, a metric that measures the average magnitude of the errors between predicted boundary positions, identified by the system and actual boundary positions in the synthetic data or human annotations. RMSE was chosen because it penalises larger errors more heavily and provided a continuous measure of distance, allowing for assessment of near-misses rather than binary correct/incorrect judgements. It was calculated as the square root of the mean of the squared differences between predicted and actual values.

Chapter 4 compared the MTI system to manually annotate arcs in whodunit short stories. Human annotators achieved moderate IAA (RMSE values of 0.06 for the three-act structure and 0.19 for the four-part structure), reflecting some subjectivity in arc boundaries. Comparing human and MTI annotations, the system showed strong alignment with human annotations (RMSE of 0.09 for the three-act structure and 0.144 for the four-part structure). Although the difference was slight, both human and system annotations revealed that whodunit short stories followed a three-act structure rather than a four-act structure. The MTI system's strong alignment with human annotations, validates the MTI approach for arc detection.

Part II shifted to narrative discourse. Chapter 5 consolidated guidelines from sources like the SANTA project (Gius, Reiter, & Marcus, 2019), defining scenes as continuous segments with equal discourse and story time, stable location, consistent character constellation, and unified action. Narrative levels were identified as shifts in narrator or diegetic embedding (for example, stories within stories). The scheme also included anachronisms (temporal deviations such as analepses and prolepses) and focalisation (perspective anchoring).

Whodunit short stories were selected as the primary genre for annotation due to their highly formulaic structure, which provides clear, recurrent patterns in narrative arcs (setup, confrontation, resolution) and discourse progression (clue placement, red herrings, detective reasoning), ideal for developing and evaluating both annotation guidelines and automated systems. Their constrained length facilitates full-text analysis and segment-level annotation, while their literary significance ensures relevance beyond experimental convenience.

Genre-specific annotations were introduced for whodunits to capture elements, such as the presentation of clues, detective reasoning, and genre specific arc tags, that are central to the narrative logic of the genre but absent from general discourse frameworks. These tags were designed as a modular extension to the core annotation scheme, enabling selective activation depending on genre without altering the foundational tagset. This modularity ensured broad applicability across narrative types while allowing fine-grained analysis in formulaic genres like detective fiction.

The initial annotation process of Chapter 5 encountered several theoretical and practical challenges discussed in Chapter 6, which resulted in further refinements before the annotation of the full mystery data set in Chapter 7.

Key issues included: (1) the container principle, in which embedded narrative levels (for example, a character's recounted memory) created ambiguous nesting boundaries, complicating segment delimitation; (2) scene versus non-scene classification, particularly in passages with implied action or summarised events, in which temporal progression was unclear despite stable setting and characters; (3) focalisation ambiguity, arising from unmarked shifts in perceptual perspective (for example, omniscient intrusions in character-bound narration) or gradual transitions not aligned with sentence boundaries; and (4) misleading character descriptions, in which a character is described as having a specific role, but the role changes or is revealed to be different later on. These challenges revealed limitations in strictly sequential, segment-based annotation and prompted adjustments such as allowing overlapping or hierarchical tags and introducing explicit markers for summary and description.

The resulting data set presented in Chapter 7 combined SANTA (2 343 segments) and mystery annotations (289 segments), providing a unified resource for training. This data set was used in Chapter 8 to create a few-shot Ollama baseline model, as well as fine-tuned RoBERTa models. Three RoBERTa models were fine-tuned: a SANTA model (fine-tuned only on the SANTA data in the collection), a mystery model (fine-tuned only on the mystery data in the collection) and a combined model (fine-tuned on the combination of both collections).

The SANTA model achieved the highest average F1 (69.89), reflecting its specialised adaptation to the SANTA corpus's unique narrative structures. The mystery model, despite training on a smaller and more formulaic data set, attained a respectable average F1 of 59.02, demonstrating efficiency (30 minutes per fold) but limited generalisation due to data sparsity. The combined model balanced both, yielding an average F1 of 68.60 and the highest ROC-AUC of 94.01, highlighting the benefit of data integration for improved discriminative power and cross-corpus robustness.

These results showed the trade-offs between data set size, training duration, and generalisation. The Ollama baseline served as a benchmark for minimal-effort scenarios, whereas the fine-tuned models, especially the combined model, suggested that integrating diverse data sets enhanced performance despite increased computational demand.

9.3 Limitations

Despite these advances, limitations persist. The STI/MTI systems rely on LDA's bag-of-words assumption, ignoring word order and semantics, which limits nuance in complex narratives. Annotation subjectivity remains, even with guidelines, as IAA never reached perfect levels. Data sets are small, spanning minimal genres and are English-centric, potentially biasing models against diverse or multilingual texts. The RoBERTa models, while effective, are computationally intensive and may overfit to training data. Finally, evaluation focused on accuracy, but real-world utility (for example, in narrative generation) requires further testing.

9.4 Contributions

This thesis contributes the following:

1. Three novel systems for arc detection (the STI, RSTI and MTI systems) in Chapter 3.
2. Consolidated, modular annotation guidelines in Chapter 5.
3. A unified publicly available data set that can be used for future studies described in Chapter 7.
4. Fine-tuned models for narrative discourse identification described in Chapter 8.

9.5 Future work

This thesis represents an initial investigation into the computational analysis of narrative structures. While the developed annotation framework, data sets, and fine-tuned models established a solid empirical and methodological foundation, the surface has barely been scratched. The field of computational narratology remains rich with untapped potential, and the work presented here serves as a starting point rather than a conclusion.

Future work could expand data sets to include more text and also perhaps different genres would enhance the data set's usability. Also, incorporating longer-form narratives (for example, novels), non-Western literary traditions, or multimodal sources (for example, graphic novels, screenplays) would increase the data set's representativeness and test the robustness of the annotation scheme and models under varied structural conditions.

The models developed in this study demonstrate robust performance in segment-level discourse annotation. However, several avenues remain for enhancing both predictive accuracy and interpretability. First, multi-task learning (MTL) frameworks could jointly

optimise discourse tagging with NER, CR, and relation extraction. Integrating character-centric NER (for example, distinguishing ‘PER:Detective’, ‘PER:Suspect’, ‘PER:Victim’) would enrich contextual embeddings, enabling the model to better capture focalisation shifts and character-bound narration. Similarly, character interaction networks, constructed via dynamic co-occurrence graphs or knowledge-enhanced relation triples, could serve as auxiliary inputs to transformer architectures, allowing the model to infer narrative levels from diegetic embedding patterns and detect embedded speech or thought more reliably.

Second, graph neural networks (GNNs) offer a natural extension for modelling narrative graphs, where nodes represent discourse segments and edges encode temporal, causal or focal links. Such architectures could explicitly learn scene continuity (via spatial-temporal stability) and anachronism detection (via deviation from chronological flow), addressing current limitations in handling non-linear or nested structures. Incorporating event extraction as a pre-task would further ground scene delimitation in verb-driven semantic roles, reducing ambiguity between scene, summary, and description.

Third, long-context models such as Longformer, BigBird, or reformer-based variants should be explored to overcome the 128-token truncation imposed by RoBERTa. These architectures preserve full-segment coherence, which is crucial for whodunit texts exceeding 2 000 tokens, and enable end-to-end sequence labelling without artificial segmentation, potentially improving F1 on low-frequency tags (for example, *prolepsis*, *clue*).

Finally, ensemble strategies combining the SANTA, mystery, and combined RoBERTa models with rule-based narratological heuristics (for example, punctuation-driven level shifts, clue-introducing discourse markers) could yield hybrid systems with both high precision and theoretical transparency. Extending the modular tagset to include act-level arc labels such as *setup*, *confrontation*, *resolution* and red-herring markers would facilitate genre-aware downstream applications such as automated plot summarisation or suspense profiling. These directions collectively aim to bridge computational scalability with the nuanced demands of literary structural analysis.

Beyond technical refinement, real-world applications in AI-assisted literary analysis, digital humanities platforms, or generative literature could validate practical impact. Deploying these models in interactive tools for scholars, educators, or writers would bridge computational scalability with interpretive depth, fostering sustained interdisciplinary collaboration between NLP and literary theory. Ultimately, this work establishes a scalable foundation for computational narratology, paving the way for broader, more nuanced exploration of narrative structures across eras, cultures, and media.

Chapter 10

Conclusion

This thesis presents a framework for the computational analysis of narrative structures, integrating theoretical narratology with practical NLP techniques.

After an extensive literature review, that identifies terminology, existing methodologies, and available data, this thesis zooms in on two main areas of computational narrative structure analysis for further investigation: Narrative arcs as well as narrative discourse. These two fields are investigated by reviewing existing work and identifying areas for improvement. For narrative arcs discussed in Part I the main short-coming was a lack of research on thematic arcs. Three different systems are developed to identify thematic narrative arcs using LDA. The final (MTI) system is evaluated by comparing the results with human annotated text. The system achieves RMSE values of 0.092 and 0.144 when compared to the three-act and four-part structure theories annotated by human annotators. These results are comparable to human IAA.

Part II focuses on narrative discourse, and identifies a lack of unified guidelines to develop a data set that can be used for this task. Existing guidelines and data sets are consolidated and an initial transformer-based model is trained on the data to showcase its usability.

10.1 Problem statement

This thesis systematically addresses the problem statement from Chapter 1 that *“The computational analysis of literary narrative structures, specifically high-level narrative structures such as plot, scene, or temporal event sequences, is limited by a scarcity of standardised annotated data sets and fragmented machine-learning approaches. Existing data sets vary widely in annotation standards due to differing research goals, hindering comparative analyses. Similarly, techniques for detecting narrative structures are often incompatible, lacking a unified framework. These gaps restrict large-scale study of narrative structures, impeding advancements in computational narratology.”*

This thesis resolved this problem through a multi-phase solution:

1. A comprehensive survey of existing resources in Chapter 2
2. The development of modular annotation guidelines in Chapter 4 and Chapter 5
3. The creation and public release of a consolidated data set in Chapter 7.
4. The development of computational narrative structure detection methods in Chapter 3 and Chapter 8.

By delivering a unified annotation framework, a consolidated public-domain data set, and detection systems, the work directly mitigated these barriers and enabled large-scale, reproducible research in computational narratology through cross-genre, cross-study comparability and scalable model training.

10.2 Scope

The scope of *“identifying narrative structure, with an emphasis on developing tools and resources to identify and annotate high-level structural elements.”* is fully realised. Chapter 2 surveys 90 relevant articles, revealing inconsistencies in scope, granularity, and format. Chapters 4 and 5 deliver a modular annotation framework.

Chapter 7 presented the consolidated public data set developed by following the guidelines from Chapter 5 and 6, merging the SANTA data set (2 343 segments) with 289 newly annotated whodunit segments. The resulting data set contained core tags (scene, non-scene, ellipsis, narrative level, anachronism, focalisation) and optional genre modules (for example, whodunit clues).

10.3 Research purpose and questions

This chapter presents a clear answer to the primary research questions:

1. **RQ 1:** “What computational approaches have been developed for analysing high-level narrative structures in literary texts?”. This question was addressed in Chapter 2. A systematic review classifies 90 studies where two dominant paradigms were identified. The first being plot modelling (we situate plot as story in Part I) through unsupervised methods like sentiment analysis, topic modelling or character networks. The second being discourse segmentation through supervised tasks.
2. **RQ 2:** “Which under-explored aspects of high-level narrative structure are both foundational to overall story organisation and amenable to automated detection?”.

This question was again addressed in Chapter 2. Plot has generally been modelled with sentiment analysis or character networks although thematic narrative arcs that make up the plot is rarely investigated although as we argued in Part I thematic shifts are inherent to the text and directly signal changes in key story elements that advances the narrative and represent its segments. Narrative discourse elements like scenes, and diegetic levels were also identified as foundational yet under-automated.

3. **RQ 3:** “What are the primary challenges in annotating high-level narrative elements, across diverse literary texts?”. This question was also addressed in Chapter 2. The literature reveals systemic barriers including the absence of consolidated, interoperable data sets, fragmented or ambiguous annotation guidelines across projects, and the high cost and expertise required for manual structural annotation, particularly for abstract phenomena like focalisation, narrative levels, and anachronisms. These were compounded by practical issues identified during annotation in Chapter 6, such as boundary ambiguity, the container principle in nested levels, and unmarked focalisation shifts.
4. **RQ 4:** “Can clear, standardised annotation guidelines be developed to consistently capture narratological concepts?”. This question was addressed in Chapter 4 that established a standardised format for thematic arc boundaries using three-act and four-part models. Chapter 5 unified existing annotation guidelines by consolidating overlapping definitions of terms for narrative discourse elements and proposing a standard annotation scheme. Chapter 6 tested the scheme in practice, identifying and resolving ambiguities (for example, container principle, scene/non-scene transitions, unmarked focalisation) through iterative refinement that resulted in a scalable protocol.
5. **RQ 5:** “Can modular, task-specific annotation schemes (for example, genre-specific annotations) be integrated into the standardised guidelines so that unique narrative features can be captured?”. This question was addressed in Chapter 4 and Chapter 5. whodunit extensions (clue, red-herring, detective-thought and whodunit specific arcs) were seamlessly integrated as optional tags, preserving core schema compatibility.
6. **RQ 6:** “Can purpose-built systems serve as strategic entry points for addressing gaps in current research while establishing a foundation for future investigation of additional structural dimensions?”. This question was addressed in Chapters 3, 4 and 8. The STI, RSTI and MTI systems and RoBERTa classifiers provide scalable entry points, with modular outputs supporting extension to character networks or sentiment arcs.

10.4 Objectives of the study

Collectively, answering the research questions fulfilled the thesis objectives:

1. *Investigate narratological concepts and existing computational approaches* to computationally detect and analyse narrative structures, as well as *identify computationally under-explored aspects of high-level narrative structure identification*. This was achieved in Chapter 2: A systematic review of 90 publications identified core narratological constructs (for example, plot, narrative level, scene/event, anachronism, focalisation, social networks) and mapped existing computational approaches, including sentiment-based arc detection, LDA topic modelling, rule-based segmentation, and early neural methods, revealed critical gaps in standardised terminology, interoperable data sets, and scalable annotation strategies. Correlation analysis of 18 labels exposed conceptual overlaps and conflicts, while evaluation of prior systems highlighted limitations in handling nested levels, temporal deviation, and genre-specific patterns, directly informing the design of subsequent annotation guidelines and automated models.
2. *Create a standardised annotation format*. This was partially achieved in Chapters 4 and fully realised in Chapter 5 where an XML-based schema with mandatory core tags and extendible modules was developed.
3. *Create an annotated data set to support model training and evaluation, with a focus on integrating existing data sets into a publicly accessible data set* This is achieved in Chapter 8 where 2 632 segments from 101 (23 unique) texts were harmonised and publicly released.
4. *Apply computational techniques (for example fine-tuned models) to automate the identification of narrative structure*. This was achieved in Chapter 3 through the STI, RSTI and MTI systems for arc detection and Chapter 8 through fine-tuned RoBERTa models for narrative discourse classification.

10.5 Contribution of the study

The work bridged a gap between qualitative literary theory and quantitative computational methods, demonstrating that structured narratives can be reliably parsed with a combination of topic modelling and transformer architectures.

The study addressed its core research questions while contributing novel methodologies and resources to the field of digital humanities.

10.5.1 Practical contributions

1. The *Standardisation of narrative level annotation schemes specifically with regards to high-level narrative structure* was contributed in Chapters 4 and 5)

2. *The Creation and public release of an annotated literature data set, incorporating high-level narrative structure information (identified in Part II clearly labelled and contextualised within narratives was contributed in Chapter 7)*
3. *The development of new tools, for example, through applying topic modelling and machine-learning techniques to automate the identification of structural components of narratives, that can be used by the literary research community to aid them in their research in the field of narratology was contributed to in the form of the STI, RSTI and MTI methods from Chapter 3 and fine-tuned RoBERTa models from Chapter 8.*

10.5.2 Theoretical contributions

1. This study refined the theoretical understanding of high-level narrative structures, by formalising narratological concepts for computational analysis, providing a clearer framework for how these structures can be systematically represented and studied across diverse literary texts.
2. Ultimately, this study combined human understanding of narratives with computational tools to uncover new insights about how narratives are structured and how computational methods can better analyse literature, while creating a clear, reliable method for future research in computational narratology.

The thesis therefore resolved the initial problem by delivering standardised, modular, public resources and consolidated, enhanced detection methods, transforming fragmented efforts into a cohesive foundation for future computational narratology.

The implications can extend beyond the academic sphere. For literary scholars, these tools enable large-scale comparative studies, for example, tracking arc complexity across historical periods, genres, or cultures. For educators, visualised narrative structures can enrich pedagogy. For computational creativity, the models and data sets support AI-driven story generation grounded in classical narratological principles.

The focus on whodunits serves a deliberate methodological purpose. The whodunit genre follows a strict structure that allows models to highlight the patterns in the text. The results from Chapter 4 and Chapter 8 illustrate how genre constraints can enhance model performance. This suggests that the applications to other tightly structured genres will be equally successful. For example fantasy tends to follow the Hero's Journey (Campbell, 2008) as discussed in Chapter 4, or modern film and television screenplays tend to stick to the seven-point structure discussed in Chapter 4. In these examples the genre forces the writer to follow a consistent pattern that the models can better identify.

10.6 Conclusion

In conclusion, this thesis marks a meaningful step toward computationally grounded narratology. By providing both theoretical clarity and practical tools, it equips researchers to explore narrative structure at scale, transforming how we read, teach, and generate stories in the digital age. The work done in this thesis provides a foundation for the journey from human intuition to machine interpretation where the path forward is now clearer, more rigorous, and better annotated.

10.7 Ethics clearance

This study has a low-risk ethics approval from the ECLM. The ethics approval number is NWU-00732-21-A7 and expires on 25 March 2026.

References

- Abbott, H. (2007). *The Cambridge Companion to Narrative*. Cambridge University Press. <https://doi.org/10.1017/CCOL0521856965.003>
- Abrams, M., & Harpham, G. (2011). *A glossary of literary terms* (10th ed.). Cengage Learning.
- Abrusán, M. (2021). The language of fiction. In *Computing perspective shift in narratives* (pp. 325–348). Oxford Academic. <https://doi.org/10.1093/oso/9780198846376.003.0013>
- Adanay, M. C., & Sporleder, C. (2015). Clustering of Novels Represented as Social Networks. *Linguistic Issues in Language Technology*, 12. <https://aclanthology.org/2015.lilt-12.4>
- Agarwal, A. (2011). Social Network Extraction from Texts: A Thesis Proposal., 111–116.
- AI@Meta. (2024). *Llama 3 model card* (tech. rep.). Meta. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
- Aksan, M., & Aksan, Y. (2015). Multi-word Expressions in Genre Specification. *Mersin Üniversitesi Dil ve Edebiyat Dergisi*, 12, 1–42.
- Alrashid, T., & Gaizauskas, R. (2023). ScANT: A Small Corpus of Scene-Annotated Narrative Texts. *Text2Story@ECIR*. <https://api.semanticscholar.org/CorpusID:258334765>
- Alrashid, T., & Gaizauskas, R. J. (2021). A Pilot Study on Annotating Scenes in Narrative Text using SceneML. *Text2Story@ECIR*. <https://api.semanticscholar.org/CorpusID:235076773>
- Altman, J. (1982). *Epistolarity: Approaches to a form*. Ohio State University Press.
- Aurnhammer, C., Cuppen, I., van de Ven, I., & van Zaanen, M. (2019). Manual Annotation of Unsupervised Models: Close and Distant Reading of Politics on Reddit. *DHQ: Digital Humanities Quarterly*, 13(3).
- Babelyuk, O. (2019). Structural peculiarities of detective stories. *Vectors of the development of philological sciences at the modern stage*.
- Bae, B.-C., & Young, R. M. (2008). A Use of Flashback and Foreshadowing for Surprise Arousal in Narrative Using a Plan-Based Approach. In U. Spierling & N. Szilas (Eds.), *Interactive storytelling* (pp. 156–167). Springer Berlin Heidelberg.
- Bakhtin, M. (1981). Forms of time and the chronotope in the novel (M. Holquist, Ed.). *The Dialogic Imagination: Four Essays by M. M. Bakhtin*, 84–258.

- Bal, M., & van Boheemen, C. (2009). *Narratology: Introduction to the theory of narrative*. University of Toronto Press.
- Baroni, R. (2007). *La Tension narrative. Suspense, curiosité et surprise*. Seuil. <https://lib.ugent.be/catalog/rug01:001384449>
- Baroni, R., & Revaz, F. (2016). *Narrative sequence in contemporary narratology*. Ohio State University Press.
- Barth, F. (2019). Annotation Guideline No. 5: Annotation Guidelines for Narrative Levels and Narrative Acts. *Journal of Cultural Analytics*, 4(3). <https://doi.org/10.22148/16.056>
- Barth, F. (2021). Annotation Guidelines for Narrative Levels and Narrative Acts V2. *Journal of Cultural Analytics*, 6. <https://doi.org/10.22148/001c.30701>
- Barth, F., & Dönicke, T. (2021). Participation in the KONVENS 2021 Shared Task on Scene Segmentation Using Temporal, Spatial and Entity Feature Vectors. *Shared Task on Scene Segmentation*.
- Bauer, M., & Lahrso, M. (2020). Annotation Guideline No. 6: SANTA 6 Collaborative Annotation as a Teaching Tool Between Theory and Practice. *Journal of Cultural Analytics*, 4(3). <https://doi.org/10.22148/001c.11747>
- Becker, M., & Egetenmeyer, J. (2018). A prominence-based account of temporal discourse structure. *Lingua*, 214, 28–58. <https://doi.org/10.1016/j.lingua.2018.08.002>
- Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, 34, 177–210.
- Bell, J. S. (2004). *Plot & structure - techniques and exercises for crafting and plot that grips readers from start to finish*. Writer's Digest Books.
- Bi, Y., Guo, P., Guo, P., & Wang, X. (2018). Comparative Analysis of Writing Styles in the Story of the Stone based on Character Networks. *37th Chinese Control Conference (CCC)*, 9679–9685. <https://doi.org/10.23919/ChiCC.2018.8483089>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Botelho, L. M. (2021). A guided journey through non-interactive automatic story generation. *ArXiv*, *abs/2110.11167*. <https://api.semanticscholar.org/CorpusID:239050064>
- Brontë, C. (1847). *Jane Eyre: An Autobiography*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/1260>
- Brooks, P. (1992). *Reading for the Plot: Design and Intention in Narrative*. Cambridge; London: Harvard University Press.
- Burgers, J. (2021). Imagining the Continuously Present Past: Visualizing William Faulkner's Narratives and Digital Yoknapatawpha. *Digital Humanities Quarterly*, 15. <https://api.semanticscholar.org/CorpusID:238210691>
- Bushell, S., Butler, J., Hay, D., & Hutcheon, R. (2022). Digital Literary Mapping: I. Visualizing and Reading Graph Topologies as Maps for Literature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 57, 11–36. <https://doi.org/10.3138/cart-2021-0008>
- Campbell, J. (2008). *The hero with a thousand faces*. New World Library. <https://books.google.co.za/books?id=I1uFuXlvFgMC>

- Carroll, L. (1871). *Through the looking-glass*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/12>
- Cawelti, J. G. (2014). *Adventure, Mystery, and Romance Formula Stories as Art and Popular Culture*. The University of Chicago Press. <http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9780226148700>
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22). Curran Associates, Inc.
- Chen, M., & Bunesco, R. C. (2021). Changing the Narrative Perspective: From Deictic to Anaphoric Point of View. *Inf. Process. Manag.*, 58, 102559. <https://api.semanticscholar.org/CorpusID:232146891>
- Choi, F. Y., Wiemer-Hastings, P., & Moore, J. (2001). Latent Semantic Analysis for Text Segmentation. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. <https://aclanthology.org/W01-0514>
- Christie, A. (2020). *Poirot investigates*. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/61262/pg61262.txt>
- Chung, J. J. Y., Williams, J. J., & Kim, J. (2018). Collaborative crowdsourcing between experts and crowds for chronological ordering of narrative events. *Korean HCI Society Conference*, 621–626.
- Church, K. W. (1993). Char_align: A Program for Aligning Parallel Texts at the Character Level. *31st Annual Meeting of the Association for Computational Linguistics*, 1–8. <https://doi.org/10.3115/981574.981575>
- Collins, W. (1854). *Hide and seek*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/7893>
- Collins, W. (1868). *The moonstone*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/155>
- Crampes, M., Veuillez, J. P., & Ranwez, S. (1998). Adaptive narrative abstraction. *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space—Structure in Hypermedia Systems: Links, Objects, Time and Space—Structure in Hypermedia Systems*, 97–105. <https://doi.org/10.1145/276627.276638>
- Dannenberg, H. (2008, January). *Coincidence and counterfactuality: Plotting time and space in narrative fiction*. University of Nebraska Press.
- DeBuse, M. A., & Warnick, S. (2023). Plot extraction and the visualization of narrative flow. *Natural Language Engineering*, 30, 480–524. <https://doi.org/10.1017/S1351324923000232>
- Deng, J., Zhong, C., Sun, S., & Wang, R. (2022). A spatio-temporal emotional framework for knowledge extraction and mining in digital humanities. *Aslib Journal of Information Management*, 74. <https://doi.org/10.1108/AJIM-09-2021-0278>
- Dery, J., & Koenig, J.-P. (2014). A Narrative-Expectation-Based Approach to Temporal Update in Discourse Comprehension. *Discourse Processes*, 52, 1–26. <https://doi.org/10.1080/0163853X.2014.966293>

- de Vree, F. (2016). *Using social co-occurrence networks to analyze Biblical narrative* [Master's thesis, VU Amsterdam, Department of Computer Science].
- Dickens, C. (1855). *Little dorrit*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/963>
- Dickens, C. (1861). *Great expectations*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/1400>
- Dickens, C. (1865). *Our mutual friend*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/883>
- Dostoevsky, F. (1866). *Crime and punishment*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/2554>
- Doyle, A. C. (1887). *A study in scarlet*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/244>
- Doyle, A. C. (2015). Adventures of sherlock holmes. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/48320/pg48320.txt>
- Droog-Hayes, M., Wiggins, G., & Purver, M. (2018). Automatic Detection of Narrative Structure for High-Level Story Representation. *5th AISB Symposium on Computational Creativity*.
- Dumas, A. (1846). *The count of monte cristo*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/1184>
- Eisenberg, J. D., & Finlayson, M. (2021). Narrative Boundaries Annotation Guide. *Journal of Cultural Analytics*, 6. <https://doi.org/10.22148/001c.30698>
- Eliot, G. (1871). *Middlemarch*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/145>
- Fields, S., Cole, C., Oei, C., & Chen, A. (2022). Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century OttomanIraqi personal diaries. *Digital Scholarship in the Humanities*, 38. <https://doi.org/10.1093/llc/fqac047>
- Finlayson, M. (2016). Inferring Propp's functions from semantically annotated text. *The Journal of American Folklore*, 129, 55. <https://doi.org/10.5406/jamerfolk.129.511.0055>
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–. <https://doi.org/10.1037/h0031619>
- Forster, E. M. (1927). *Aspects of the Novel*. New York: Harcourt, Brace.
- Frermann, L., Cohen, S. B., & Lapata, M. (2017). Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association for Computational Linguistics*, 6, 1–15. <https://api.semanticscholar.org/CorpusID:5029790>
- Freytag, G. (1894). *Freytag's Technique of the drama: an exposition of dramatic composition and art*. (E. MacEwan, Trans.; 6th ed.). Chicago : Scott, Foresman.
- Fudolig, M., Alshaabi, T., Cramer, K., Danforth, C., & Dodds, P. (2022, August). *Characterizing narrative time in books through fluctuations in power and danger arcs* [Preprint].
- Fudolig, M., Alshaabi, T., Cramer, K., Danforth, C., & Dodds, P. (2023). A decomposition of book structure through ousiometric fluctuations in cumulative word-time. *Humani-*

- ties and Social Sciences Communications*, 10(1). <https://doi.org/10.1057/s41599-023-01680-4>
- Gaizauskas, R., & Alrashid, T. (2019). SceneML: A Proposal for Annotating Scenes in Narrative Text. *Workshop on Interoperable Semantic Annotation*, 13. https://sigsem.uvt.nl/isa15/ISA-15_proceedings.pdf#page=19
- Gaskell, E. C. (1853). *Cranford*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/57539>
- Genette, G., Lewin, J. E., & Culler, J. D. (1980). Narrative discourse : An essay in method. *Comparative Literature*, 32, 413.
- Georgescu, M., Clark, A., & Armstrong, S. (2008). A Comparative Study of Mixture Models for Automatic Topic Segmentation of Multiparty Dialogues. *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*. <https://aclanthology.org/I08-2133>
- Gius, E., Jannidis, F., Krug, M., Zehe, A., Hotho, A., Puppe, F., Krebs, J., Reiter, N., Wiedmer, N., & Konle, L. Detection of Scenes in Fiction. In: *Book of abstracts of the digital humanities conference*. Utrecht, Netherlands, 2019, July.
- Gius, E., Willand, M., & Reiter, N. (2021). On organizing a shared task for the digital humanities conclusions and future paths. *Journal of Cultural Analytics*, 6(4), 1–28. <https://doi.org/10.22148/001c.30697>
- Gius, E., Jannidis, F., Krug, M., Zehe, A., Hotho, A., Puppe, F., Krebs, J., Reiter, N., Wiedmer, N., & Konle, L. Detecting Scenes in Fiction. In: *Proceedings of the digital humanities conference*. Utrecht University. Digital Humanities, 2019, July.
- Gius, E., Reiter, N., & Marcus, W. (2019). Cultural Analytics: A Shared Task for the Digital Humanities: Annotating Narrative Levels. *Journal of Cultural Analytics*, 4(3). <https://doi.org/https://doi.org/10.22148/16.049>
- Gius, E., & Vauth, M. (2023). Towards an Event Based Plot Model. A Computational Narratology Approach. *Journal of Computational Literary Studies*, 1(1). <https://doi.org/https://doi.org/10.26083/tuprints-00023251>
- Gombert, S. (2021). Twin BERT Contextualized Sentence Embedding Space Learning and Gradient-Boosted Decision Tree Ensembles for Scene Segmentation in German Literature. *Shared Task on Scene Segmentation*.
- Gonzalez, A. (2023). Mysteriously murderous manors of crime fiction: An examination of setting in knives out. *Digital Literature Review*, 10, 75–84. <https://doi.org/10.33043/DLR.10.1.75-84>
- Greimas, A. J. (1984). *Structural semantics: An attempt at a method*. University of Nebraska Press.
- Grosz, B. J., & Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3), 175–204.
- Gupta, A., Srinivasan, P., Shi, J., & Davis, L. S. (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2012–2019.
- Hammond, A. (2021). Annotation Guidelines for Narrative Levels. *Journal of Cultural Analytics*, 6. <https://doi.org/10.22148/001c.30704>

- Hardy, T. (1895). *Jude the obscure*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/153>
- Harris, Z. (1954). Distributional structure. *Word*, 10(2-3), 146–162. https://link.springer.com/chapter/10.1007/978-94-009-8467-7_1. https://doi.org/10.1007/978-94-009-8467-7_1
- Hatzel, H. O., & Biemann, C. (2021). LTUHH@STSS: Applying Coreference to Literary Scene Segmentation. *STSS@KONVENS*. <https://api.semanticscholar.org/CorpusID:243990461>
- Heyns, N. (2024). Post-digital narrative analysis. In *The Routledge Handbook of AI and Literature*. Routledge.
- Heyns, N., & van Zaanen, M. (2021). Finding topic boundaries in literary text. *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*. https://dh2021.digitalhumanities.org.za/wp-content/uploads/2021/11/DHASA_book_of_abstracts.pdf
- Heyns, N., & van Zaanen, M. (2022). Detecting multiple transitions in literary text. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3375–3381. <https://aclanthology.org/2022.lrec-1.360.pdf>
- Heyns, N., & van Zaanen, M. (2024). Automatic narrative structure identification in literary texts. *Digital Humanities Benelux Journal*, 6, 71–90. https://journal.dhbenelux.org/wp-content/uploads/2024/11/5_Heyns_individual.pdf
- Heyns, N., & van Zaanen, M. Refining Narrative Annotation for Segmentation: Challenges and Debates. In: *Comparative Literature Goes Digital DH2025 Mini-Conference*. 2025.
- Heyns, N., & Van Zaanen, M. (2024, November). Annotating mystery novels: Guidelines and adaptations. In Y. K. Lal, E. Clark, M. Iyyer, S. Chaturvedi, A. Brei, F. Brahman, & K. R. Chandu (Eds.), *Proceedings of the the 6th workshop on narrative understanding* (pp. 55–66). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.wnu-1.9>
- Heyns, N., & Van Zaanen, M. (2025). Annotated Mystery Narratives Data Set. *Journal of Open Humanities Data*.
- Heyns, N., & van Zaanen, M. (n.d.). *Automatic Narrative Annotation* [Submitted to *Computational Humanities Research* (under review)].
- Hirschberg, J., & Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Comput. Linguist.*, 19(3), 501–530.
- Huang, L., & Huant, L. (2013). Optimized event storyline generation based on mixture-event-aspect model. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 726–735.
- Innes, M., & Brookman, F. (2023, December). Whodunit, howdunit, whydunit? how detectives think about and investigate complex hard to solve murders. Routledge. <https://doi.org/10.4324/9781003195283-26>
- James, P. D. (1997). Murder and mystery: The craft of the detective story. <https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-108256>

- Jayannavar, P., Agarwal, A., Ju, M., & Rambow, O. (2015). Validating Literary Theories Using Automatic Social Network Extraction. *CLfL@NAACL-HLT*, 32–41. <https://doi.org/10.3115/v1/W15-0704>
- Jockers, M. L. (2015). Syuzhet: Extract sentiment and plot arcs from text [Available at: <https://github.com/mjockers/syuzhet>. Accessed 30 January 2023].
- Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41, 750–769.
- Joon, J., Chung, Y., Williams, J. J., & Kim, J. (2017). *Collaborative Crowdsourcing Between Experts and Crowds for Chronological Ordering of Narrative Events* [Workshop]. <https://api.semanticscholar.org/CorpusID:238996050>
- Kearns, E. (2019). For Annotating Anachronies and Narrative Levels in Fiction. *Journal of Cultural Analytics*, 4. <https://doi.org/10.22148/16.052>
- Kearns, E. (2020, July). Annotating and quantifying narrative time disruptions in modernist and hypertext fiction. In C. Bonial, T. Caselli, S. Chaturvedi, E. Clark, R. Huang, M. Iyyer, A. Jaimes, H. Ji, L. J. Martin, B. Miller, T. Mitamura, N. Peng, & J. Tetreault (Eds.), *Proceedings of the first joint workshop on narrative understanding, storylines, and events* (pp. 72–77). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.nuse-1.9>
- Kearns, E. (2021). Annotation Guidelines For Narrative Levels, Time Features, and Subjective Narration Styles in Fiction (SANTA 2). *Journal of Cultural Analytics*, 6. <https://doi.org/10.22148/001c.30699>
- Ketschik, N., Krautter, B., Murr, S., & Zimmermann, Y. (2019). Annotation guideline no. 4: Annotating narrative levels in literature. *Journal of Cultural Analytics*, 4(3). <https://doi.org/10.22148/16.055>
- Ketschik, N., Krautter, B., Murr, S., & Zimmermann, Y. (2021). On the theory of narrative levels and their annotation in the digital context. *Journal of Cultural Analytics*, 6, 69–97.
- Kim, H. (2022). Sentiment Analysis: Limits and Progress of the Syuzhet Package and ItsLexicons. *DHQ: Digital Humanities Quarterly*, 16.
- Kodinariya, T., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 90–95.
- Koliasa, O. (2019, January). Detective stories: From classic to postmodern. definition and history. <https://doi.org/10.36059/978-966-397-124-7/95-113>
- Konle, L., & Jannidis, F. (2022). Modeling Plots of Narrative Texts as Temporal Graphs. *CEUR Workshop Proceeding*, 3290, 318–336.
- Konstantopoulos, G. (2023). Chapter 7 Charting Emotional Structure in Cuneiform Literature. In *The shape of stories: Narrative structures in cuneiform literature* (pp. 144–184). Brill. https://doi.org/10.1163/9789004539761_008
- Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., & Weimar, L. (2015). Rule-based Coreference Resolution in German Historic Novels. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 98–104. <https://doi.org/10.3115/v1/W15-0711>

- Kurfali, M., & Wirén, M. (2021). Breaking the Narrative: Scene Segmentation through Sequential Sentence Classification. *STSS@KONVENS*. <https://api.semanticscholar.org/CorpusID:243990463>
- Labatut, V., & Bost, X. (2019). Extraction and Analysis of Fictional Character Networks: A Survey. *CoRR*, *abs/1907.02704*.
- Labov, W., & Waletzky, J. (1997). Narrative Analysis: Oral Versions of Personal Experience. *Journal of Narrative & Life History*, 7, 3–38.
- Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. *Essays on the Verbal and Visual arts, Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society*, 12–44.
- Labov, W. (2013). *The Language of Life and Death: The Transformation of Experience in Oral Narrative*. Cambridge University Press.
- Lee, E. (2017). Pluperfect in discourse: When and why do we go back in time? *Journal of Pragmatics*, 121, 76–90. <https://doi.org/https://doi.org/10.1016/j.pragma.2017.10.001>
- Lee, J. (2017). Shakespeares Tragic Social Network; or Why All the Worlds a Stage. *DHQ: Digital Humanities Quarterly*, 11(2).
- Lee, O., You, E., & Kim, J. (2021). Plot Structure Decomposition in Narrative Multimedia by Analyzing Personalities of Fictional Characters. *Applied Sciences*, 11, 1645. <https://doi.org/10.3390/app11041645>
- Li, B., Cardier, B., Wang, T., & Metze, F. (2018). Annotating High-Level Structures of Short Stories and Personal Anecdotes. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. <https://aclanthology.org/L18-1520>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July). Roberta: A robustly optimized bert pretraining approach. <https://doi.org/10.48550/arXiv.1907.11692>
- Lukin, S., Bowden, K., Barackman, C., & Walker, M. (2016). Persona Bank: A Corpus of Personal Narratives and Their Story Intention Graphs. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1026–1033. <https://aclanthology.org/L16-1163>
- Lukin, S., & Walker, M. (2019). A Narrative Sentence Planner and Structurer for Domain Independent, Parameterizable Storytelling. *Dialogue and Discourse*, 10, 34–86. <https://doi.org/10.5087/dad.2019.103>
- Macovei, A. (2017). Capturing the Evolution of Characters in Novels. *2017 21st International Conference on Control Systems and Computer Science (CSCS)*, 545–552. <https://doi.org/10.1109/CSCS.2017.83>
- Mahmoodi, M. A., & Savab, F. (2023). Semiotic ladder: The schema of producing meanings in narrative. *Semiotica*, 2023(253), 51–70. <https://doi.org/doi:10.1515/sem-2021-0142>

- Malmgren, C. (2004). Anatomy of murder: Mystery, detective, and crime fiction. *The Journal of Popular Culture*, 30, 115–135. https://doi.org/10.1111/j.0022-3840.1997.3004_115.x
- Mani, I. (2005). Narrative Summarization. *Journal Traitement Automatique Des Langues (TAL)*, 45, 1–24.
- Mansouri, A., Affendey, L. S., & Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2), 339–344.
- Mill, J. S. (1861). *Utilitarianism*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/11224>
- Miller, B., & Park, J. S. (2020). Computing narrative. *CEUR Workshop Proceedings*, 2723, 182–190.
- Milutinovi, D. (2020). Grammar of the detective genre. *Facta Universitatis, Series: Linguistics and Literature*.
- Min, S., & Park, J. (2016a, March). Complex Networks VII. Springer. https://doi.org/10.1007/978-3-319-30569-1_19
- Min, S., & Park, J. (2016b). Mapping Out Narrative Structures and Dynamics Using Networks and Textual Information. *CoRR*, abs/1604.03029. <https://dblp.org/rec/journals/corr/MinP16.bib>
- Min, S., & Park, J. (2016c). Narrative as a Complex Network : A Study of Victor Hugo's Les Misérables. *PROCEEDINGS OF HCI KOREA 2016*, 100–107. <https://doi.org/10.17210/hcik.2016.01.100>
- Min, S., & Park, J. (2019). Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling (T. Gross, Ed.). *PLOS ONE*, 14(12), e0226025. <https://doi.org/10.1371/journal.pone.0226025>
- Montgomery, L. M. (1908). *Anne of green gables*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/64365>
- Muzny, G., Algee-Hewitt, M., & Jurafsky, D. (2017). Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32, ii31–ii52. <https://doi.org/10.1093/llc/fqx031>
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3–26.
- Newberry, K. M., & Bailey, H. R. (2019). Does semantic knowledge influence event segmentation and recall of text? *Memory and Cognition*, 47, 1173–1187. <https://doi.org/10.3758/s13421-019-00926-4>
- Novakova, I., & Siepmann, D. (2020, January). Phraseology and Style in Subgenres of the Novel. Springer. https://doi.org/10.1007/978-3-030-23744-8_1
- Orczy, B. E. (2004). The league of the scarlet pimpernel. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/5805/pg5805.txt>
- Ouyang, J., & McKeown, K. (2015, September). Modeling Reportable Events as Turning Points in Narrative. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2149–2158). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D15-1257>

- Ouyang, J., & McKeown, K. (2014). Towards automatic detection of narrative structure. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4624–4631.
- Patten, W. (2024). Great short stories, volume 1 (of 3) detective stories. Retrieved November 19, 2025, from <https://www.gutenberg.org/cache/epub/74548/pg74548.txt>
- Pevzner, L., & Hearst, M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28, 19–36.
- Piper, A., & Bagga, S. (2022). Toward a Data-Driven Theory of Narrativity. *New Literary History*, 54, 879–901. <https://doi.org/10.1353/nlh.2022.a898332>.
- Piper, A., So, R. J., & Bamman, D. (2021, November). Narrative Theory for Computational Narrative Understanding. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 298–311). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.26>
- Poe, E. A. (1848). *Eureka: A prose poem*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/32037>
- Portelance, E., & Piper, A. (1997). Understanding Narrative: Computational Approaches to Detecting Narrative Frames. *Computational Linguistics*, 23. <https://dh2017.adho.org/abstracts/172/172.pdf>
- Porteous, J., Charles, F., & Cavazza, M. (2016). Plan-based narrative generation with coordinated subplots. *Frontiers in Artificial Intelligence and Applications*, 285, 846–854. <https://doi.org/10.3233/978-1-61499-672-9-846>
- Porteous, J., & Cavazza, M. (2009). Controlling Narrative Generation with Planning Trajectories: The Role of Constraints. In I. A. Iurgel, N. Zagalo, & P. Petta (Eds.), *Interactive storytelling* (pp. 234–245). Springer Berlin Heidelberg.
- Porteous, J., Cavazza, M., & Charles, F. (2010). Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Trans. Intell. Syst. Technol.*, 1(2). <https://doi.org/10.1145/1869397.1869399>
- Prince, G. (1973). *A Grammar of Stories: An Introduction* (Vol. 13). De Gruyter Mouton. <https://doi.org/10.1515/9783110815900>
- Prince, G. (1982). Narrative Analysis and Narratology. *New Literary History*, 13(2), 179–188. <https://doi.org/10.2307/468908>
- Propp, V. I. (1968). *Morphology of the folktale* (L. Scott, Trans.). University of Texas Press.
- Purver, M., Körding, K. P., Griffiths, T. L., & Tenenbaum, J. B. (2011). Unsupervised Topic Modelling for Multi-Party Spoken Discourse. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 17–24. <https://doi.org/10.3115/1220175.1220178>
- Reese, B., Denis, P., Asher, N., Baldridge, J., & Hunter, J. (2007). *Reference Manual for the Analysis and Annotation of Rhetorical Structure (Version 1.0)* (tech. rep.). University of Texas, Departments of Linguistics and Philosophy. Austin. http://timeml.org/jamesp/annotation_manual.pdf

- Reiter, N. (2015). Towards Annotating Narrative Segments. *Annual Meeting of the Association for Computational Linguistics, 2015-text*, 34–38.
- Reiter, N., Marcus, W., & Gius, E. (2019). A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks. *Journal of Cultural Analytics*, 4(3). <https://doi.org/https://doi.org/10.22148/16.048>
- Ren, P., Wang, Y., & Zhao, F. (2023). Re-understanding of data storytelling tools from a narrative perspective. *Visual Intelligence*, 1. <https://doi.org/10.1007/s44267-023-00011-0>
- Reynar, J. C. (1994). An automatic Method of Finding Topic Boundaries. *Association for Computational Linguistics, 32nd Annual Meeting of the Association for Computational Linguistics*, 331–333. <https://doi.org/10.3115/981732.981783>
- Richardson, A., & Hay, D. (2022). A dynamic topography for visualizing time and space in fictional literary texts. *Information Design Journal*, 27(1), 5–20. <https://doi.org/https://doi.org/10.1075/idj.22010.ric>
- Rodosthenous, C. T., & Micheal, L. (2019). Web-STAR: A Visual Web-based IDE for a Story Comprehension System. *Theory and Practice of Logic Programming*, 19(2), 317–359. <https://doi.org/10.1017/S1471068418000443>
- Rogers, A., Karpinska, M., Gupta, A., Lialin, V., Smelkov, G., & Rumshisky, A. (2024). NarrativeTime: Dense Temporal Annotation on a Timeline. *LREC-COLING 2024*, 12053–12073.
- Rosenbaum, R. (2019). Toward a Renewed Theory of the Narreme. *The American Journal of Semiotics*, 35, 187–215. <https://doi.org/10.5840/ajs201982255>
- Ryan, M. (1992). Possible Worlds, Artificial Intelligence, and Narrative Theory. *Special Issue: Between Science & Literature*, 23(2), 135–139. <https://doi.org/10.2307/3685081>
- Sandeep, S., Sihra, A., Evans, E., Wilkens, M., & Bamman, D. (2023). Grounding Characters and Places in Narrative Texts. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1, 11723–11736.
- Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. some problems and solutions. *Quality & Quantity*, 54, 1095–1108.
- Schmidt, B. M. (2015). Plot arceology: A vector-space model of narrative structure. *2015 IEEE International Conference on Big Data (Big Data)*, 1667–1672. <https://doi.org/10.1109/BigData.2015.7363937>
- Schneider, F., Barz, B., & Denzler, J. (2021). Detecting Scenes in Fiction Using the Embedding Delta Signal. *STSS@KONVENS*. <https://api.semanticscholar.org/CorpusID:243990361>
- Schwan, H., Jacke, J., Kleymann, R., Stange, J.-E., & Dörk, M. (2019). Narrelations - Visualizing Narrative Levels and their Correlations with Temporal Phenomena. *Digit. Humanit. Q.*, 13. <https://api.semanticscholar.org/CorpusID:210166164>
- Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, 1–27. <https://www.cfilt.iitb.ac.in/~cfiltnew/resources/surveys/rahul-ner-survey.pdf>

- Shelley, M. W. (1818). *Frankenstein; or, the modern prometheus*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/84>
- Sims, M., Park, J., & Bamman, D. (2019). Literary event detection, 3623–3634. <https://doi.org/10.18653/v1/P19-1353>
- Skorochoďko, E. F. (1971). Adaptive Method of Automatic Abstracting and Indexing. *IFIP Congress*.
- Steele, T. (1981). The structure of the detective story: Classical or modern? *Modern Fiction Studies Lafayette, Ind.*
- Stolarek, J. (2009). A crime story or metafictional game? a definition and redefinition of the status of the detective novel in martin amis's london fields and tzvetan todorov's "the typology of detective fiction". *Anglica Wratislaviensia, Edited by Anna Michoska-Stadnik, 3155*, 27–39.
- Sui, P., Wang, L., Hamilton, S., Ries, T., Wong, K., & Wong, S. (2023, July). Mrs. Dalloway Said She Would Segment the Chapters Herself. In N. Akoury, E. Clark, M. Iyyer, S. Chaturvedi, F. Brahman, & K. Chandu (Eds.), *Proceedings of the the 5th workshop on narrative understanding* (pp. 92–105). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.wnu-1.15>
- Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. <https://doi.org/10.1109/DSAA.2017.61>
- Taivalkoski-Shilov, K. (2019, August). Free indirect discourse: An insurmountable challenge for literary MT systems? In J. Hadley, M. Popovi, H. Afli, & A. Way (Eds.), *Proceedings of the qualities of literary machine translation* (pp. 35–39). European Association for Machine Translation. <https://aclanthology.org/W19-7305>
- Todorov, T. (1971). The 2 Principles of Narrative. *Diacritics, 1*(1), 37. <https://doi.org/10.2307/464558>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., . . . Straus, S. E. (2018, October). *PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation*. <https://doi.org/10.7326/M18-0850>
- Trollope, A. (1857). *Barchester towers*. Retrieved November 12, 2025, from <https://www.gutenberg.org/ebooks/3409>
- Uglanova, I., & Gius, E. (2020). The order of things. a study on topic modelling of literary texts. *Workshop on Computational Humanities Research*. <https://api.semanticscholar.org/CorpusID:227913848>
- Underwood, T. (2018). Why Literary Time is Measured in Minutes. *ELH, 85*, 341–365. <https://doi.org/10.1353/elh.2018.0013>
- van Zundert, J., Koolen, M., Neugarten, J., Boot, P., van Hage, W. R., & Mussmann, O. (2022). What do we talk about when we talk about topic? *Workshop on Computational Humanities Research*. <https://api.semanticscholar.org/CorpusID:254045843>
- Vonnegut, K. (2005). *A man without a country*. New York: Seven Stories Press.

- Wake, P. (2006). Narrative and narratology. *The Routledge Companion to Critical and Cultural Theory*, 23–36.
- Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. <https://doi.org/10.1145/1553374.1553515>
- Wells, D. (2010, February 13). Dan Wells on story structure.
- Wirén, M., & Ek, A. (2021). Annotation Guideline No. 7 (Revised): Guidelines for Annotation of Narrative Structure. *Journal of Cultural Analytics*, 6. <https://doi.org/10.22148/001c.30703>
- Woods, D. (2022). *The Emotional Arcs of Horror: a distant reading of Stephen King's novels* [Master's thesis, UVM Patrick Leahy Honors College Senior Theses]. <https://scholarworks.uvm.edu/hcoltheses/509>
- Yauney, G., Underwood, T., & Mimno, D. Computational Prediction of Elapsed Narrative Time. In: *Workshop on narrative understanding at naacl*. Minneapolis: NAACL, 2019. <https://gyauney.github.io/papers/elapsed-narrative-time.pdf>
- Zehe, A., Konle, L., Dümpelmann, L. K., Gius, E., Hotho, A., Jannidis, F., Kaufmann, L., Krug, M., Puppe, F., Reiter, N., Schreiber, A., & Wiedmer, N. (2021, April). Detecting Scenes in Fiction: A new Segmentation Task. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 3167–3177). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.276>
- Zehe, A., Konle, L., Guhr, S., Dümpelmann, L. K., Gius, E., Hotho, A., Jannidis, F., Kaufmann, L., Krug, M., Puppe, F., Reiter, N., & Schreiber, A. (2021). Shared Task on Scene Segmentation @ KONVENS 2021. *STSS@KONVENS*. <https://api.semanticscholar.org/CorpusID:243990426>
- Zhang, Y. (2020). *Temporal Dependency Structure Modeling* [Doctoral dissertation, Brandeis University, Graduate School of Arts and Sciences].
- Zhang, Y., & Xue, N. (2018, October). Neural Ranking Models for Temporal Dependency Structure Parsing. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3339–3349). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1371>
- Zhou, D., Xu, H., Dai, X.-Y., & He, Y. (2016). Unsupervised storyline extraction from news articles. *IJCAI*, 3014–3021.

Appendix A

Final tagset for narrative discourse

Table A.1: Tags used during the annotation process.

Tagset	Tags	Properties	Values
Diegetic level*	Narrator	Extradiegetic*	0*, meta
		Intradiegetic	1a*, 1b, 1c, ...
		Metadiegetic	2a*, 2b, ..., 3a, 3b, ...
Segment	Scene	Time*	e.g., evening
		Place*	e.g., crime scene
		Character_constellation*	characters in scene
	Non-scene	Summary	
		Scenic_Passage	
		DESCRIPTIVE_PASSAGE	
Anachronisms	Ellipsis		
	Analepsis	Embedded	
		Interruptive	
Perspective*	Prolepsis	Embedded	
		Interruptive	
		Detective_thought†	
	Voice*	Homodiegetic	
		Hetrodiegetic	
Misc†	Clue	Identified	e.g., murder weapon
		Referred	e.g., murder weapon
Acts†	Introduction		
	Investigation		
	Confrontation		
	Confession		
	Reveal		
	Aftermath		

Tagsets and the DETECTIVE_THOUGHT tag marked with a dagger (†) are modular and specific to whodunit texts. Tagsets marked with an asterisk (*) are compulsory for each

text. Properties marked with an asterisk (*) are compulsory if the related tag was chosen and values marked with an asterisk (*) are compulsory if the related property was chosen.

Appendix B

Distribution of tags in narrative discourse annotated text

This appendix shows an overview of the number of tags annotated for each of the initial four selected stories in Chapter 5.

Property	C1	C2	D1	D2
Diagetic level				
Extradiegetic level – 0	1 [‡]	1 [‡]	1 [‡]	1 [‡]
Extradiegetic level – Meta	1	0	2	1
Intradiegetic level	2	3	19	10
Metadiegetic level	1	2	18	9
Segment				
Scene	11	12	5	10
Non-Scene – Summary	3	5	5	1
Non-Scene – Description	0	1	1	0
Non-Scene – Scenic passage	0	0	0	0
Ellipsis	3	2	1	3
Anachronism				
Analepsis – Embedded	1	4	3	5
Perspective				
Voice – Homodiegetic	1 [‡]	1 [‡]	1 [‡]	1 [‡]
Embedded focalisation	1	4	18	9
Detective thoughts	4	10	4	9
Acts				
Introduction	1	1	1	1
Investigation	1	1	1	1
Confrontation	0	1	1	1
Confession	0	0	0	1
Reveal	1	1	1	1
Aftermath	1	1	1	0
Misc				
Clues	8	9	14	18

APPENDIX B. DISTRIBUTION OF TAGS IN NARRATIVE DISCOURSE ANNOTATED TEXT¹⁵⁷

Entries marked with a double dagger (‡) are assigned to the entire text. C1 is *The adventures of the Italian nobleman*, C2 is *The jewel robbery at the grand metropolitan*, D1 is *A case of identity*, and D2 is *The red-headed league*.

Appendix C

Ollama prompt template

The following prompt template and examples are used in Chapter 8 for annotating literary texts based on the guidelines provided in Chapter 5. We used “A Case of Identity” row 8 of the CSV file by the same name (made available in Chapter 8¹), “A Case of Identity” row 5 of the CSV file by the same name and “The Adventures of the Italian Nobleman” row 27 of the CSV file by the same name as examples as they represented a diverse combinations of tags from the data. The resulting annotations is used as the baseline in Chapter 8

```
prompt_template = f"""
```

```
You are an expert in narratology, annotating literary texts based on the
    guidelines. For each text segment, output a CSV row with these columns:
    SANTA_ID;Text;Scene;Summary;Descriptive_passage;Analepsis
    ;Prolepsis;Extradiegetic;Intradiegetic;Metadiegetic;Focalisation
    ;Voice_homodiegetic;Voice_heterodiegetic.
```

Here are the definitions for each tag to guide your annotation:

- Scene: A segment of narrative discourse that presents the `\emph{histoire}` (story), typically involving a coherent sequence of events with specific characters, time, and place, annotated using the SCENE tag.
- Summary: A non-scene where events are condensed or narrated briefly, often as a sub-scene within a broader scene, assigned as a property of NON-SCENE.
- Descriptive_passage: A non-scene focused on description rather than events, providing detailed pauses in the narrative for setting or character details, assigned as a property of NON-SCENE.
- Analepsis: A flashback that shifts the narrative time from the present to the past, tagged as ANALEPSIS, which can be embedded or interruptive.
- Prolepsis: A flashforward occurring in forms like visions, prophecies, or foreshadowing, shifting narrative time to the future, tagged as PROLEPSIS, which can be embedded or interruptive.
- Extradiegetic: The level of the narrator or implied author outside the story world, annotated with NARRATOR and value 0, potentially including metatextuality (meta).
- Intradiegetic: The diegetic level of characters and events within the story, annotated with value 1 (and letters like 1a, 1b for sequential arrangement).

¹<https://github.com/Nuette/Automatic-narrative-annotation/tree/main/MD>

- Metadiegetic: A secondary narrative embedded within the primary diegetic level, such as stories told by characters, annotated with value 2 (and letters for arrangement).
- Focalisation: The perspective from which the narrative is seen, indicating the narrator's access to information; can be zero/unrestricted (omniscient, knows more than characters), internal (limited to a character's knowledge), or external (observes without internal access); tag with FOCALISATION and properties like EMBEDDED or INTERRUPTIVE.
- Voice_homodiegetic: When the narrator appears in the story as a character, usually referring to themselves in the first person.
- Voice_heterodiegetic: When the narrator does not appear in the story, with narration mostly in the third person.

Use 1 for 'yes' and 0 for 'no' in binary columns (for example, Scene).

Examples:

{examples_str}

Now annotate this new segment:

ID: {{id}}

Text: {{text}}

Output only the CSV row (no extra text).

"""