universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Measuring Plot in the Age of Distant Reading.
A study in the search for the narrative arc within the stories of Sherlock Holmes.

verfasst von / submitted by

David Siegl, BA BA MA

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Arts (MA)

Wien 2024 / Vienna 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: UA 066 647

Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: Digital Humanities

Betreut von / Supervisor: Univ.-Prof. Mag. Dr. Tara Andrews

1. Introduction

The present master's thesis explores the application of natural language processing (NLP) techniques to unveil and analyse the narrative arcs within the collected stories of Sherlock Holmes written by Arthur Conan Doyle. Often considered as one of the most iconic representatives of the genre of detective fiction, *Sherlock Holmes* has been the subject of numerous literary studies revolving around its structuralist and formulaic qualities as well as its relation to the genre in general. While said research endeavours have already been able to produce interesting results and hypotheses, they have to this day remained for the most part merely anecdotal and non-exhaustive when it comes to their ability to consider the whole canon of the adventures of Sherlock Holmes and compare certain features across the different texts accordingly. Furthermore, due to an ongoing scepticism regarding the introduction of quantitative and/or computational methods into the sphere of literary analysis and interpretation, the possibilities of evaluating large corpora of similar texts without the need for disproportionate amounts of manual labour have remained largely unexplored and neglected in favour of the traditional, highly subjective and anecdotal approaches.

Thus, the research laid out in this master's thesis instead aims to employ concepts of distant reading and computational literary studies in order to identify and comprehend the underlying narrative structures that drive the adventures of Sherlock Holmes, shedding light on their inherent literary qualities and narrative dynamics in the process. In essence, the concept of distant reading – most notably brought forward by scholars such as Franco Moretti, Matthew Jockers and Ted Underwood – vouches for a reliance on quantitative measures and statistical models to extract basic, overarching features which can be observed across numerous texts and are representative of their narrative structure. The origins of this basic axiom of Moretti et al. can even be traced back to the early 20th century, where Russian formalists such as Boris Yarkho began to use metrics like word counts and computed them over large stretches of different texts with the aim of finally arriving at a more positivist and objective proof for notions of literary form as well as historical change. While the scope of these early endeavours in the realm of a distant reading avant la lettre remains admirable from today's perspective – especially given the fact that most of these calculations have been achieved without the processing power of a computer – these roots should even more so serve as a testament for the still unrealised potential of quantitative literary studies in an age where digital archives of literary corpora, statistical frameworks and the resources for computational aptness lie (figuratively speaking) at a digital humanist's fingertips.

With that being said, the present master's thesis not only tries to incorporate the rich theoretical background of formalism, structuralism and more recent developments of distant reading into its own research, but also makes use of various state of the art NLP frameworks which can easily be deployed

via the popular programming language *Python* and offer a much more streamlined and efficient way for the analysis of large amounts of literary texts than the aforementioned traditional approaches. Through common NLP tasks such as named entity recognition (NER) or part of speech (POS) tagging, one can then engineer a wide array of different features suitable for further statistical enquiries and computations. By integrating NLP methodologies into the study of literary works, this research paves the way for future investigations in narrative analysis, enabling interdisciplinary explorations of diverse literary works through a computational lens. The results obtained from this research contribute to the development of NLP pipelines capable of automatically identifying narrative patterns and comparing features across different texts of a given genre.

The general outline of the master's thesis entails firstly a definition of the structural features of crime/detective fiction and its historical implications, secondly a detailed description of the current state of distant reading as well as its connections to formalism/structuralism which will also function both as a methodological basis and a framework for the interpretation of the quantitative NLP analysis discussed in the third and last part.

2. The emerging patterns of *Crime/Detective Fiction* and its relevance for Sherlock Holmes

The history of *crime fiction* – at least when it comes to its infancy state – can be traced back to the depiction of criminal violence and also its subsequent punishment within the ancient as well as the biblical myths during some of the earliest days of literature. It is here where basic plot devices and narrative structures such as the motif of the murder, functioning as a starting point for the narrative trajectory or the final uncovering of the killer demarcating some form of cathartic resolution, begin to evolve and shape the future schematic qualities of the genre.[1] Given its long tradition, wide adaptation and reception but also broad definition – i. e. the mostly physical presence of criminality serving as the central theme – crime fiction has over time produced various subgenres, which started developing and portraying certain differentiating tropes respectively within the wide spectrum of suspenseful storytelling.

One of the most prominent examples of these subgenres can then be found in the realm of *detective fiction* with Arthur Conan Doyle's infamous private detective Sherlock Holmes and his adventures, as they are told through the perspective of his partner John Watson, at the helm. In these adventures, which were written between the late 19th and the early 20th century, the role of the clever, rationalistic protagonist in the form of the detective becomes crucial for the overall arc of the narrative, thereby also moving away from the more primordial, graphical descriptions of crime and pivoting instead towards a more subtle and intellectually stimulating style of representation. This new emerging form originated mainly from the American author Edgar Allan Poe's short stories about the private detective C. Auguste Dupin, whose mental capabilities to solve mysterious crimes – akin to the faculties of his British successor – take center stage and guide the reader through a web of hints and clues before finally revealing the solution to the riddle via deductive reasoning.[2] Subsequently, by building upon the formula of his American colleague and refining it over the course of 60 distinct stories, Doyle managed to establish a sort of capital haven of detective fiction within the far reaching literary landscape of crime fiction, thus also laying out several formulaic principles for generations of detective stories to come in the process.[3]

Whereas the traditional texts of crime fiction oftentimes focused more on the narration of the crime through the eyes of the actual perpetrator, the works of detective fiction instead enter the scene when the crime itself has already happened. Thus, the detective – and with him also the reader – experience the act of violence mainly through retrospective imaginations or – in a narratological sense – analeptically

[1] Cf. Scaggs, John (2005): Crime Fiction. The New Critical Idiom. London: Taylor & Francis Routledge, p. 1 ff.

[2] Cf. Abbott, Randy L. (2008): Roots of Mystery and Detective Fiction. In: Critical Survey of Mystery and Detective Fiction, ed. by Carl Rollyson. New York: Salem, p. 1891–1892.

[3] Cf. Rzepka, Charles J. (2010): Introduction: What Is Crime Fiction? In: A Companion to Crime Fiction, ed. by Charles Rzepka and Lee Horsley. New Jersey: Blackwell, p. 4.

in order to slowly put the pieces of the puzzle together and reach a solution. The term 'analepsis', which has in the context of narrative studies mainly been coined by the structuralist Gérard Genette and may in more colloquial terms also be known under the term 'flashback', denotes, generally speaking, an achronological form of story telling where events that have already happened previously are only told at a latter stage within the order of a given narrative. This temporal shift is furthermore also accompanied by a change of perspective – i. e. the observation that "[t]he criminal is now no longer the subject of the narrative but the object of the detective's pursuit, and the fact that the detective is on the side of the law makes reading about crime respectable as well as suggesting its containment."[4] From the unfiltered, gory and vile explorations of urban back-alleys and criminal thoughts back to the more philistine and civil spaces[5] such as Holmes' apartment at 221B Baker Street as well as his imaginative mind palace:[6] the evolution of crime fiction in general and the emergence of its new subgenre in particular bring with it not only a topological difference but even more importantly also a restructuring of its formal narrative essence. Defining said essence – especially in regards to the overarching aesthetic choices for a given genre – has consequently been the subject of a multitude of works within the field of literary studies surrounding detective fiction. While recent advancements within the Digital Humanities and especially its field of *distant reading* and *computational literary studies* have found new ways of working with large corpora of genre literature in order to for example extract and validate underlying structures and patterns within the texts over large quantities of data – an aspect which will be explored in much more detail throughout the following chapters –, the prominence of formalist and structuralist methods in conjunction with genre literature as a whole, as well as detective fiction in particular displays a considerable interest in operationalising and quantifying recurring narratological structures even before the advent of computers into the sphere of the humanities.

For example Tzvetan Todorov in his essay *The typology of detective fiction* characterises detective fiction as a general case of the whodunit which is then mainly transported via consistent narrative segments such as the exposition, the reconstruction of the murder and finally also its solution. Through this narrative structure the respective text then also reveals the basic, underlying duality of *fable* (story) and *subject* (plot), which suggests an analeptical retelling of a crime that has already happened but still has to be solved by the detective.[7] The chain of events slowly but surely unfolds in a relatively stable, logical succession where the repetition of recurring motifs and devices invites the reader's involvement

[4] Worthington, Heather (2010): From The Newgate Calendar to Sherlock Holmes. In: A Companion to Crime Fiction, ed. by Charles Rzepka and Lee Horsley. New Jersey: Blackwell, p. 21.

[5] Cf. for example Cawelti, John G. (1997): Canonization, Modern Literature and the Detective Story. In: Theory and Practice of Classic Detective Fiction, ed. by Jerome H. Delamater and Ruth Prigozy. London: Greenwood, p. 10 ff.

[6] Cf. also Ascari, Maurizio (2020): Counterhistories and Prehistories. In: The Routledge Companion to Crime Fiction, ed. by Janice Allan et al. London & New York: Taylor & Francis, p. 27.

[7] Cf. Todorov, Tzvetan (1977): The poetics of prose. New Jersey: Blackwell, p. 140.

in the dramatic turns and twists of the heroic investigator. Similarly Viktor Shklovsky shows on the basis of a few selected passages and stories from the complete works of Doyle's private detective how a general pattern of story telling emerges from which Shklovsky then extrapolates the existence of a general narrative schema defined by a finite number of recurring plot instances which together are constituting a distinct arc of dramatic progression. The initial exposition of the problem/the case as well as the introduction of the obligatory client, who consults the services of Holmes, is usually followed by subsequent events such as the continuous dropping of hints and clues, which offers an overarching hook along one or more red herrings, MacGuffins and other dramatic devices until finally culminating in the solving of the respective crime.[8] Along those lines William Nelles and Linda Williams in their paper about narrative order in detective fiction offer a – albeit still manual – way of breaking down Sherlock Holme's stories as well as similar representatives of the genre into smaller parts in order to then formally assess the basic underlying structure of said texts. Nelles and Williams go on to argue that detective fiction stories such as Poe's *Murders in the Rue Morgue or* Doyle's *The Speckled Band* offer a recurring narrative structure consisting of at least three key moments, which are the problem, the solution and the analysis of the problem, which are also corresponding to an analeptical shift between them, since the solution usually is revealed before the complete reconstruction of the case and the detective's observations.[9] The graphs, which are thereby produced, furthermore suggest a consistent rise of narrative tension throughout the temporal progression of a given text until its climax right at the end when the criminal gets identified and caught. Last but not least the rule-based, repetitive structure of detective fiction has also been pointed out by some of the genre's authors themselves. Here two of the most frequently cited essays are Ronald Knox's *Ten Commandments for Detective Fiction* as well as Austin Freeman's *The Art of the Detective Story*, wherein both Knox and Freeman formulate a number of basic patterns which constitute the narrative arc of a successful detective story.[10] [11]

One must furthermore not neglect "[t]he fact that the detective story is a product of the emergent magazine culture and the economy it promoted supports Martin Priestman's emphasis on the Holmesian detective story as a series. In contrast to the serial publication of long novels, here each tale is self-contained, the detective's solution providing full narrative satisfaction, but so managed as to stimulate an appetite for another, similar story – so much so that, notoriously, popular demand and apparently irresistible commercial pressures made it impossible for Doyle to kill Holmes off as he wished in

[8] Cf. Shklovsky, Viktor (1990): Theory of Prose. Elmwood Park: Dalkey Archive, pp. 115–116.

[9] Cf. Nelles, William and Williams, Linda (2021): Doing Hard Time: Narrative Order in Detective Fiction. In: Style, 55, 2, pp. 196–198.

[10] Cf. Freeman, Austin R. (1924): The Art of the Detective Story. In: http://gadetection.pbworks.com/w/page/7931646/The%20Art%20of%20the%20Detective%20Story [3.4.23].

[11] Cf. Knox, Ronald (1929): Ten Commandments for Detective Fiction. In: http://gadetection.pbworks.com/w/page/7931441/Ronald%20Knox%27s%20Ten%20Commandments%20for%20Detective%20Fiction [3.4.23].

1893."[12] According to Martin Kayman, the wide appeal of the Sherlock adventures are then firstly brought forward by a certain relatively stable form built upon carefully selected and arranged individual elements, that are secondly trying to satisfy certain demands and expectations set mainly by the economical context of the late 19th century and its rising popularity of literary magazines. This then again also ties back to the postulation of a general stability and repetitiveness of the different products of the genre of crime/detective fiction, which even still holds true when viewed on a grander evolutionary scale, where variation in the end seems to be serving a larger overarching essence of content. In this light the killing of the beloved protagonist in *The Reichenbach Falls* remains only temporary before the demand for the continuation of the series becomes too overwhelming to resist, so that it is then not only the resurrection by the creator himself that fuels the legacy[13] but even more so the still ongoing receptions, adaptations and retellings of the already well known topics and patterns:

> Holmes has been significant to the history of detective fiction not only in the extent of his influence upon the crime writers that came afterwards but also in the manner in which his figure has cast a shadow over both his precursors and peers as well. He has become a seemingly permanent fixture in the landscape not only of crime writing, but also of television and film adaptation, fan fiction and graphic novels.[14]

Thus, the general consensus regarding (classical) research around the theory and history of crime related genre fiction with Doyle's infamous protagonist at the forefront already suggests a heavy reliance on formulaic and structuralist approaches, especially emphasising the (re-)creations of underlying abstract patterns as an adequate representation for the inherent narrative features of a given detective story in the process.[15] At the same time most of the (traditionally) carried out research still predominantly relies on a few handpicked and manually evaluated texts in order to use them as individual examples or subsets of a larger, organic whole – such as the extensive corpus of 60 Sherlock stories or even genre literature produced by different authors for a given time period. Despite the mostly convincing arguments brought forward by various papers in the realm of the (methodologically) orthodox literary studies and genre research, a large part of said papers then also appears to be hampered by its own reluctance or inability to extend the object of its inquiries on a *quantitative* level, given that a sample size of at best a couple of texts can still be called out as anecdotal, regardless of the actual *quality* of the work being done. With that being said it then seems all the more plausible to vouch for a more quantitatively centred method, which could enable a much more expansive foray into the depths of for example the collected adventures

---

[12] Kayman, Martin A. (2003): The short story from Poe to Chesterton. In: The Cambridge Companion to Crime Fiction, ed. by Martin Priestman. Cambridge: Cambridge University, p. 43.

[13] Cf. Glazzard, Andrew (2018): The Case of Sherlock Holmes. Secrets and Lies in Conan Doyle's Detective Fiction. Edinburgh: Edinburgh University, p. 229 ff.

[14] Burrow, Merrick: Holmes and the History of Detective Fiction. In: The Cambridge Companion to Crime Fiction, ed. by Martin Priestman. Cambridge: Cambridge University, p. 27.

[15] Cf. also Nicol, Bran (2019): Holmes and Literary Theory. In: The Cambridge Companion to Sherlock Holmes, ed. by Janice M. Allan and Christopher Pittard. Cambridge: Cambridge University, p. 191 ff.

of Sherlock Holmes in order to build upon and verify previously formulated notions by the literary scholars of detective fiction. This endeavour could be achieved – as this thesis will aim to prove – with the aid of the aforementioned advancements of computational literary studies and here especially the concepts of distant reading, whose origins and main goals will be explained going forward.

3. *Formalism, Structuralism & Distant Reading*: The origins of Distant Reading and its epistemological implications for (computational) literary studies moving forward

In order to fully comprehend the scientific roots and intentions of the school of *distant reading* as well as its (future) potential when it comes to bringing new insights and results into a data driven field of digital humanities, one first has to travel back in time for about a 100 years – namely to the early twentieth century where a considerable surge in positivism and fact centred methodology brought about the school of *Russian formalism*, a group of Slavonic philologists and linguists which devoted themselves to formal studies in poetic/literary language. Generally speaking, members of this school vouched for a 'scientific' and 'objective' method of extracting the particular qualities of artistic language use, which in turn meant that context reliant approaches, such as sociocultural studies, were mostly abandoned in favour of a strictly text based practice, which focused on finding reoccurring patterns and features that could then be declared as pertaining to the specific qualities of a given object of study and as existing independent of subjective judgement. As Boris Eichenbaum, one of the founding fathers of Russian formalism, for example writes in one of his essays regarding the fundamental goals of a 'science of literature':

> [O]ur Formalist movement was characterized by a new passion for scientific positivism – a rejection of philosophical assumptions, of psychological and aesthetic interpretations, etc. Art, considered apart from philosophical aesthetics and ideological theories, dictated its own position on things. [...] The basis of our position was and is that the object of literary science, as such, must be the study of those specifics which distinguish it from any other material.[16]

Given such presumptions, the early proponents of formalism subsequently often turned to more quantitative approaches in order to generate knowledge about their respective research interests while also drawing concepts from the natural sciences, such as biology, or imitating the formal representations of mathematics for describing important relations across larger corpora of texts. In this regard the meaning of a given text is first and foremost understood as being expressed through its form, while form itself is constituted by the sum of its different elements and their logical connections.[17] The basic epistemological paradigm of formalism thus relies on the notion of a systematic model, where one can split a given object into smaller subsegments in order to then analyse and compare them before finally arriving again at a better understanding of the initial object, which gets now represented as the sum of its different parts. It becomes then the main task for the formalist to firstly look for proper demarcations in regard to an adequate operationalisation of the relevant parts of a literary text and secondly also to assign fitting labels as well as categories to individual features as well as texts or even groups of texts that share similar features.

---

[16] Eichenbaum, Boris (2012): The Theory of the "Formal Method". In: Russian Formalist Criticism. Four Essays, 2nd Edition, ed. by Lee T. Lemon and Marion J. Reis. Lincoln and London: University of Nebraska, pp. 107–108.

[17] Cf. Yarkho, Boris (2016): The elementary foundations of formal analysis. In: Studia Metrica et Poetica 3, 2, pp. 151– 152.

For example Boris Yarkho, who was also part of the Russian formalists and is nowadays considered one of the most idiosyncratic literary scholars to have worked during the early 20th century, as he was mainly known for designing a methodology of exactitude that integrated metrics like word and other frequency counts into his arsenal – with the ultimate goal being their computation over large stretches of different texts via the usage of statistical formulas.[18] Thus, in his central essay on the foundations of formal analysis Yarkho goes on to divide the study of literary form into three distinct categories which are *metrics*, *stylistics* and *iconology*. Each of these three categories is then furthermore split into numerous subcategories that equally contribute to the overall meaning of their respective parent label. While the first category of *metrics* has been a frequent research interest for a lot of formalists right from the start – given especially the highly restrictive and formalised structure of poetry and verse with its repetitive combinations of syllables, sounds and similar qualities that can often be seen as an obvious candidate for quantitative measures –[19] Yarkho on the other hand, with this concept of *iconology*, also argues for a formal exploration of the 'images' that take up larger stretches of (narrative) texts and can be considered conducive to the description of motifs and ultimately also plot.[20] In addition Yarkho points to the fact that these formal elements are strongly connected to the *aesthetics of number and time* by which he basically means the exploration of frequencies in relation to a longer period consisting of several points in time with each point displaying a distinct frequency value.[21] Or to put it more formally: one could instead imagine this relation as a line graph, where the x-axis represents a time line and the y-axis the frequency of a given feature at a given point in time, whereas the overall result then describes a curve of the relative change of a given value over time. Consequently, according to Yarkho the coordination of the elements pertaining to literary form is heavily based on the relation of number and arrangement which finally confronts the researcher with problems that "can be solved only by means of *quantitative analysis*."[22]

In order to provide a sound ontological framework for his aim of distilling exact and reproducible results from literary analysis Yarkho furthermore asserts that literary scholars should also engage in the determination of relevant features that can be considered relevant for the expression of the essence not only of single texts but even across a whole category of texts and in distinction to other categories of dissimilar texts.

> First of all: what is an individual feature? In theory, the analytical powers of the human mind know no bounds, and any feature can be divided into new features. This problem can however be resolved in a much simpler way for the purposes of this study. After all, we don't need all the features, only

[18] Cf. Gasparov, Mikhail (2016): Boris Yarkho's works on literary theory. In: Studia Metrica et Poetica 3, 2, pp. 130 ff.
[19] Cf. also Eichenbaum (2012): p. 111.
[20] Cf. Yarkho (2016): pp. 155 ff.
[21] Cf. Yarkho (2016): pp. 160 ff.
[22] Yarkho (2016): p. 167.

> the distinguishing features. Thus, the analysis will keep going until a distinction is found with sufficient clarity. Let us explain using an example: In two entities – A and B – there are stylistic devices; in this they are similar. There are also semantic devices among these stylistic devices; again similar. However, there are similes in A, but not in B. That is it. Let's suppose that there are similes in both entities, but in A they constitute 2% of devices, and in B – 75% (with more or less the same number of devices); here again we can stop. However, suppose that the similes are represented in equal proportions; then we would have to conduct further analysis; suppose that in A similes with »like« were most prevalent (»flies like an arrow«), and in B – with »as … as« (»flies as fast as an arrow«); then, the analysis is complete at this stage. [...]
>
> Another difficulty is much more problematic and is as follows: can we be sure that we have found all the distinguishing features? Of course not. Yet it is clear that the more we find, the closer we will be to resolving the problem. The problem is essentially solved as follows: we find a particular set of features in entity A (i. e.forms, their proportions and connections), which are not found in any part of entity B; this set can be considered stage one of the solution, the »sufficiency« stage. Having achieved this, we can add to the set obtained until the capabilities of our intellect have been exhausted. [...]
>
> It is clear that the ambition of any synthesis is to reduce the largest possible number of distinguishing characteristics to the smallest possible number of distinguishing concepts.[23]

Again one finds here the aforementioned formalist notion that a given literary text can be broken down into smaller, distinctive parts, which in turn results in a more concrete and fundamental examination of the underlying structure as well as a thorough comparison of these distinctive parts when it comes to their contribution to for example two different categories of genre. In this light genre A could then for instance be represented by a combination of n elements that were particularly often found in texts associated with the same label, while genre B on the other hand could be represented by another combination of n elements or in fact also by a clear lack of the feature set indicative of its counterpart A. Echoing the common statistical method of A/B testing, where two groups of varying feature sets are first defined and subsequently compared on the basis of their respective calculated statistical significances, one can herein see another clear indicator of Yarkho's mathematical aptness that in a way also anticipated certain approaches which were recognised in full only decades later.

One noteworthy example for this method of feature generation and testing can then be found in Yarkho's study on the speech distribution in five-act tragedies within the periods of Classicism and Romanticism. Herein he proposes, for a proper dissection of its research objects, to segment the texts to be analysed into their respective acts and then calculate for each segment metrics such as the ratio of the number of characters speaking versus the number of characters present on stage.[24] As a result a given drama finally gets represented by a sequence of numbers, denoting the relative change of said ratio or similar quantitative measures over the succession of the different segments. Not only does this approach, as a sort of distant reading avant la lettre display early signs of normalisation attempts, for the different texts

[23] Yarkho, Boris (2019): Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism). In: JLT 13, 1, pp. 15–17.
[24] Cf. Yarkho (2019): p. 18 ff.

get all put into the same basic form, but it also mimics the now common NLP methods of vectorisation, where human language in general is transformed into arrays of numbers trying to capture the latent meaning as accurately as possible while at the same time offering a standardised way of evaluating the relations between these different arrays. Concluding his study on the differences between the classicist and romanticist dramas, Yarkho then turns to statistical formulas such as the correlation coefficient or the standard deviation to further explore the relationship between the already derived features. Arriving at the result of being able to state a significant difference in the overall narrative structure between the two classes of periods, he also manages to back up his claims by extensive numerical data that does not merely stick to anecdotal observations of single texts. Thus, despite Yarkho's limited tool set, when it comes to actually carrying out his numerical and statistical analysis, the scope of his research endeavours can be described as nothing short of admirable – especially given the fact that these endeavours and considerations still had been achieved without the processing power of a computer – making these early roots appear even more so as a testament for the still unrealised potential of today's quantitative literary studies.

Another important contemporary of Yarkho can be found in Vladimir Propp, who is in particular known for his morphological treatises on the Russian folktale. The basic theory, which is most extensively laid out in Propp's monography *Morphology of the Folktale*, revolves around the idea that the different variants of traditional folktales can be represented in a treelike hierarchy, where the complete pool of the available different texts descends from one or more archetypes that display overarching and recurring features when it comes to their combinations and orderings of distinct plot elements. Thus, Propp also builds upon the basic formalist principle that literary texts can only be examined properly when being broken down into smaller subsections or segments:

> For the sake of comparison we shall separate the component parts of fairy tales by special methods; and then, we shall make a comparison of tales according to their components. The result will be a morphology (i.e., a description of the tale according to its component parts and the relationship of these components to each other and to the whole).[25]

Again in Propp's methodology a given text is therefore represented by a sum of smaller parts that can subsequently be easily compared across the different samples of a given genre and/or similar categories. In order to provide a proper nomenclature for the narrative elements to be extracted, Propp introduces the term *function* into his repertoire, which denotes in its general definition "an act of a character, defined from the point of view of its significance for the course of the action."[26] With that being said, a given tale can then be described as the succession of multiple functions that produce a logical causation of narrative events, which display in most cases the progression from an initial 'lack' of something or

[25] Propp, Vladimir (2009): Morphology of the Folktale. Texas: University of Texas, p. 19.
[26] Propp (2009): p. 21.

someone towards its gradual (re-)attainment and the resulting relief from a particular insufficiency that for example ails the hero, who consequently sets out on a quest full of challenges and other obstacles.[27] Such an abstract description of the general structure of the folktale makes it already possible to arrive at the core – or to pick up again on the biological terminology brought forward by Propp and other formalists – the essential *root* which constitutes the gist of the studied texts pertaining to a given category. According to the hereby resulting, underlying epistemological framework the different incarnations of existing folktales then appear as the children of a corresponding parent structure, which share the same elements or functions, with the only difference being their respective permutative (re-)arrangement of said functions. The actual individual gestalt is mostly secondary and not so much indicative of a completely alternative expression, given that a certain type of narrative can be represented by one out of several possible concrete options. According to this principle and akin to the popular storytelling trope of the *MacGuffin*, it therefore does not matter so much whether for example the agent in a crime thriller has to chase a suitcase full of money or a ticking time bomb as long as it keeps the general flow of the plot going. In the case of the folktales the hero may then for instance be granted a magical gift to aid him on his quest while the actual appearance of the gift in a given tale can vary depending on the setting of the story and its remaining elements. Or as Propp puts it: "We observe that the substitution of certain aspects by others, within the confines of each type, is practiced on a large scale."[28] Nevertheless the complete number of all possible functions is, according to Propp's extensive investigations, found to be quite limited, which in turn means that it is not so much a large scope of variation that defines the morphological model of the Russian folktale but rather different chains of a finite amount of permutations that mostly generate meaning through their relative positions to one another.[29] These 'chains of permutations' are last but not least very similar to the concepts already encountered in the works of Yarkho, for they are also generally represented by Propp as sequences or arrays of formal signs and numbers that stand in for the general narrative trajectory of a given tale in an abstract and normalised way. In essence Propp formulates the thesis that literary texts – at least texts that can initially be grouped under the category of a specific genre and/or similar labels – adhere to general formulaic principles and rules that get repeated throughout the historical persistence of such phenomena and can therefore also be subject to formalistic analysis.[30]

Speaking of historical persistence and evolution: another important proponent of Russian formalism, who for the most part focused on the change and development of certain literary structures and forms

---

[27] Cf. Propp (2009): pp. 34–35.
[28] Propp (2009): p. 49.
[29] Cf. Propp (2009): p. 64.
[30] Cf. also Propp, Vladimir (2014): Fairy Tale Transformations. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, p. 50 ff.

over time, can be found in Yuri Tynianov. Corresponding to the general affinity for botanical metaphors and concepts present in the formalist tradition, Tynianov for example postulated that a given literary genre is not merely a constantly stable system but rather displays changes over time, which can be first and foremost perceived through the examination of specific features at different points in time.

> In this way, the genre as system can vacillate. It arises (from anomalies and embryonic ideas in other systems) and declines, turning into the rudiments of other systems. The genre function of a device is not fixed. It is impossible to conceive of a genre as a static system because the very awareness of a genre emerges as the result of its clash with a traditional genre (i.e. the sense of a shift—at least partial—from the traditional genre to a "new" one taking its place). The whole point here is that the new phenomenon supplants the old one, takes its place and, without being a "development" of the old one, nevertheless acts as its substitute. When this "substitution" does not occur, the genre as such disappears, falls apart.[31]

In this light the observation of certain parts as a representation of a larger whole is not merely applicable to single texts or smaller groups of them but can also be used in the context of larger historical analysis. When employing these methods of historical analysis, as Tynianov further suggests, evolutionary patterns then begin to emerge that reveal a trajectory in constant flux, entailing phases of high points as well as decay when viewed over a long enough time frame. In concordance to the formal examination of individual narratives, regarding for example their plot structures, one can also produce cumulative graphs of the same features across multiple instances corresponding to the same genre and/or similar categories in order to arrive at a visualisation of these evolutionary paradigms. At its core there lies therefore a postulation of the agreement between a singular literary work as a system and literature in general that is finally mirrored in an albeit abstract congruence of the same elements such as plot devices. Akin to Propp, Tynianov refers to these elements as distinct functions that constitute their full meaning first and foremost through the interdependence of the different parts, their relative positions to one another which conclusively make up the texts' structure.[32] These structures are in turn defined as *series* by Tynianov, which can be seen as a grouping of common literary texts at a given point in time that are then superseded by other series as time evolves.[33] For example the evolution of crime fiction laid out in the previous chapter already displayed similar historical patterns, so much so that the genre in question constantly produced new and different forms (even subgenres) as its influence and recognition widened or rather as its maturity grew. In essence the literary evolution according to Tynianov is then understood as an ongoing succession of shifts that render older formal qualities obsolete and promote newer ones instead.

> In conclusion: studying literary evolution is only possible when literature is treated as a series, a system interrelated with other series and systems and conditioned by them. This investigation

[31] Tynianov, Yuri (2019): Permanent Evolution. Selected Essays on Literature, Theory and Film. Boston: transcript, pp. 155–156.

[32] Cf. Tynianov (2019): p. 269 ff.

[33] Cf. Tynianov (2019): p. 276.

> should proceed from the constructive function to the literary function, from the literary function to the speech function. It should clarify the evolutionary interaction of forms and functions. The study of evolution should proceed from the literary series to the closest interrelated series, not to more distant ones, though the latter may be of major significance. This does not deny the overarching significance of major social factors; in fact, this significance becomes clearest when considering the evolution of literature. The unmediated diagnosis of the "influence" of major social factors, meanwhile, substitutes the study of the modification and deformation of literary works for the study of the evolution of literature.[34]

Last but not least one also has to refer to Viktor Shklovsky's contributions to the methodology of Russian formalism. Generally speaking Shklovsky is known for his works surrounding the formalistic qualities of (genre) prose – a fact which gets exemplified by his aforementioned essay on Sherlock Holmes and his relation to detective fiction. In his popular essay *Art as Technique*, a treatise on the style of Tolstoy's prose texts, Shklovsky for instance vouches for an algebraic method for the representation of literature via specific symbols to account for the inevitable abstraction the human mind undertakes when confronted with stimuli such as art. Thus, a given perceived object is in fact only characterised by its most noteworthy features and not remembered or even cognitively apprehended in its entirety. The algebraic method is therefore warranted by a process of *algebrization* that lies, according to Shklovsky, at the core of human perception and especially its economic reduction of unnecessary details.[35] Through the repetitive encountering of such recognisable and abstract algebraic elements one finally also arrives at the possibility of formulating general principles in terms of the (dis-)similarity of different texts pertaining to the same categories such as genre.

In his monography *Theory of Prose* Shklovsky goes on to extensively focus on the structural features of fiction. First and foremost he argues that plot can be defined as a progressive accumulation of motifs which are causally connected to one another. As a result a given narrative's plot or rather the sum of its different motifs and similar elements is characterised by a sense of resolution that draws towards an end as the problem initially formulated is also resolved. Throughout a given plot's trajectory one may therefore often encounter antagonistic elements that produce tension in for example the form of certain segments where the protagonist goes off on tangents and is confronted with obstacles before finally arriving at his original goal.

> But what precisely does a story need in order to be understood as something truly complete? It is easy to see that, in addition to a progressive development, there exists in a story also a structure analogous to a ring or, rather, a loop. The description of happy lovers does not in and of itself create a story. What a story needs is love hindered by obstacles (i.e., it needs happy lovers perceived against the traditional background of love hindered by obstacles. For example, A loves B, but B doesn't love A. By the time B falls in love with A, A has ceased to love B. [...]

---

[34] Tynianov (2019): p. 282.

[35] Cf. Shklovsky, Viktor (2012): Art as Technique. In: Russian Formalist Criticism. Four Essays, 2nd Edition, ed. by Lee T. Lemon and Marion J. Reis. Lincoln and London: University of Nebraska, pp. 33 ff.

> Thus, to be a true "story," it must have not only action but counteraction as well (i.e., some kind of incongruity).[36]

The principal formula of narrative then consists of a conducive notion of change and progression that establishes its varying qualities throughout its different segments and motifs. It is (in an antithetical sense) precisely that shifting procedure, the alternation of dissimilar elements that in the end constitutes the movement of prose as a whole. With this discovery Shklovsky not only contributed a considerable amount of new insights to the formalistic research of literature in general but especially in regard to the exploration of plot structures. As Eichenbaum therefore argues, Shklovsky's work made it possible to differentiate the compositional form from the thematic content by introducing the dichotomy between plot and story, which originated from the Russian terms *syuzhet* and *fabula* and describes a basic but important distinction that was later echoed and refined in the works of the structuralists to which this thesis will turn shortly.[37]

To conclude this short excursion on Russian formalism, one can clearly see from the different proponents and their ideas discussed above that these early attempts of exploring and describing the language of literature via quantitative and statistical means at the beginning of the 20th century already offered a lot of sound approaches for extracting the underlying form of the different literary types such as verse, drama and most importantly prose. To turn once again to Eichenbaum and his already cited essay about the evolution of the 'formal method', one can then also infer a general progression towards a systematic development of a formalistic vocabulary that started at its onset with the highly formalised language of poetics and slowly but surely also spread out its principles to other text forms before arriving probably at the most complex and elaborate literary structures such as the (modern) novel.[38]

The fact that such a systematic and formalistic theory still has not lost its academic merit is for example testified by the existence of *structuralism*, another school of thought that rose to prominence mostly during the second half of the 20th century and at least in part originated directly from Russian formalism. As Jurij Striedter recounts in his monograph on the evolution of formalism and structuralism, the school of (Czech) structuralism descended directly from the later works of the already mentioned Tynianov and his linguist colleague Roman Jakobson, which were primarily written towards the end of the 1920s when the surrounding intelligentsia shifted its centre of its academic fruition – in part due to political upheaval – from Russian territory to the confinements of Prague. Thus, the origins of Czech structuralism were first and foremost fuelled by an intent to connect more closely the systematic studies of human language in general and of literary language in particular via the postulation of a common methodology that would

[36] Shklovsky (1990): p. 52–53.
[37] Cf. Eichenbaum (2012): p. 115.
[38] Cf. Eichenbaum (2012): p. 133.

be able to equally account for both phenomena and their respective structural qualities.[39] On a side note one can here already draw a parallel to the much more recent incorporation of certain NLP methods and frameworks that originally were mainly intended for their applications on linguistic corpora but have consequently also been adapted for the purposes of literary studies in accordance to the principles of distant reading, on which the present thesis bases much of its own considerations as well and which will be discussed in more detail in the following chapters. On another side note one also should be aware of the fact that the school of structuralism has undergone a lot of changes since its initial inauguration around the 1930s, spreading its influence to numerous other disciplines such as for example sociology or even psychoanalysis and finally also (at least in part) culminating in the more radical currents of post-structuralism. Recounting and explaining these different tangents of such a historical genesis in its entirety would warrant at least another thesis on its own, and since the present thesis has already established its methodological roots more firmly on the basis of formalism, it will instead mainly focus on supplementing said roots with a more selective description of certain structuralist concepts that were mainly brought forward by Tzvetan Todorov and Gérard Genette.

First and foremost, the Bulgarian structuralist Todorov, who has also already been briefly mentioned in the previous chapter in regard to his contributions to the study of crime fiction, is generally considered as one of the figureheads when it comes to the study of genre in the realm of structuralism, as well as his expansions on the concepts laid out by Propp's formal studies. Along those lines Todorov asserts that most if not all fictional texts correspond to a specific genre which are in their most basic sense classes of texts. A given text can then also 'disobey' its genre but it still relates to it – albeit ex negativo. As a result one is therefore via the method of formal analysis able to extract common properties from a collection of texts which are pertaining to the same genre in order to put them together into the same class.[40] Building upon the notion of plot or syuzhet laid out by Shklovsky, Todorov furthermore argues that one of the most decisive properties or features of a particular genre, at least when it comes to prose texts, can be achieved through the study of the structuring of the respective plot elements, whereas for example the classic detective story differs fundamentally from a romance novel.[41] Probably the most extensive contribution within Todorov's oeuvre to the study of genre can finally be found in his monograph *The Fantastic*, where the structuralist tries to provide in essence a definition of the core elements contriving the genre of paranormal literature. In the introductory chapter of said study Todorov formulates the basis of his epistemological approach as follows:

---

[39] Cf. Striedter, Jurij (1989): Literary Structure, Evolution, and Value. Russian Formalism and Czech Structuralism Reconsidered. Cambridge and London: Harvard University, p. 83 ff.

[40] Cf. Todorov, Tzvetan (2014): The Origin of Genres. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, p. 196 ff.

[41] Cf. Todorov (2014): p. 199.

> The notion of genre immediately raises several questions; fortunately, some of these vanish once we have formulated them explicitly. The first question is: are we entitled to discuss a genre without having studied (or at least read) all the works which constitute it? The graduate student who asks this question might add that a catalogue of the fantastic would include thousands of titles. Whence it is only a step to the image of the diligent student buried under books he must read at the rate of three a day, obsessed by the idea that new ones keep being written and that he will doubtless never manage the fantastic to absorb them all. But one of the first characteristics of scientific method is that it does not require us to observe every instance of a phenomenon in order to describe it; scientific method proceeds rather by deduction. We actually deal with a relatively limited number of cases, from them we deduce a general hypothesis, and we verify this hypothesis by other cases, correcting (or rejecting) it as need be. Whatever the number of phenomena (of literary works, in this case) studied, we are never justified in extrapolating universal laws from them; it is not the quantity of observations, but the logical coherence of a theory that finally matters.[42]

Here the impetus regarding the classification of genre and its respective texts clearly lies on a scientific, objective approach that critically derives the logically most plausible judgement from the data at hand. Akin to the quantitative and statistical attempts of the formalists, Todorov therefore also vouches for the postulation of a probabilistic hypothesis that represents merely the best approximation of truth possible within the present research state. With a sizeable amount of relevant instances representing a given text genre or category there may consequently arise distinct patterns on which one can then base an inference of rules and similar devices that are conducive to the constitution of said genre/category. Note also that Todorov still provides here the example of the lone scholar sitting in his chamber and manually sifting through pages upon pages of relevant literature in order to formulate his conclusions; while more recent advancements in the quantitative studies of literature such as distant reading on the other hand render it much more convenient to accomplish similar deductions via the help of computational methods and without much need for such magnitudes of manual effort, this observation can then again serve as a further legitimisation of the present thesis' methodological approaches. Referring back to Propp's concept of morphology, Todorov also relies on the relational, hierarchical tree structure, in which the dependencies between the different literary 'specimen' can be conceived. The sum of these dependencies and relations underlies the same 'root' or subgroup of literature (i. e. the specific genre), so that the original formalistic notion regarding the dissection of an (abstract) phenomenon into its smaller parts in order to properly assess it is again echoed in the structuralist's mantra "that every literary study must participate in a double movement: from the particular work to literature generally (or genre), and from literature generally (from genre) to the particular work."[43] In this light a proper study and subsequent definition of genre must entail the extraction of the relevant and observable properties as well as laws that together make up the literary structure. Such properties can, according to Todorov, be split up into three different aspects of the literary work which are the *verbal* (concrete sentences), the *syntactical*

---

[42] Todorov, Tzvetan (1973): The Fantastic. A structural approach to a literary genre. Cleveland and London: Case Western Reserve University, pp. 3–4.

[43] Todorov (1973): p. 7.

(logical, temporal and spatial relations) and the *semantic* (themes).[44] For the purposes of the present master's thesis and its ambitions of analysing the plot structures of genre texts the latter two aspects seem to be of particular importance and also strongly match with the more recent theories of distant reading when it comes to the domain of measuring the narrative trajectory where for example the extraction of temporal and/or spatial relations can easily be accomplished via POS tagging or NER.

The second important representative of structuralism, Gérard Genette, is mainly known for his formulation of a precise vocabulary for the study of the different elements of narrative texts, which is most commonly referred to as *narratology*. Besides these widely accepted contributions to the formal study of narrative, Genette furthermore also attended to questions of genre, to which his essay *The Architext* is probably one of the most important testaments in this regard, wherein he emphasises the usefulness of a proper classification of different groups of texts, that might then for example be characterised by a label like 'detective novel' and similar sub branches, in the context of a systematic literary study.[45] When it comes to the epistemological foundation of Genette's narratology one can clearly see the influence of the formalistic tradition – especially regarding the reliance on the basic notion of separating the plot/syuzhet (form) from its story/fabula (content) as well as the splitting up of a given text into smaller parts that can be examined independently. It is therefore only the narrative text itself with all its different structural features that is defined as the object of a narratologist's study, while its content or its act of narration are intentionally left out of the equation.[46] Without wanting to provide an exhaustive recount of all the narratological concepts laid out by Genette at this point – for such an undertaking would clearly go above and beyond the scope of this thesis – one concept, that is of particular relevance for the study of detective fiction and has also already briefly been mentioned in the previous chapter, warrants a bit more attention in the context of the present research endeavour: This is namely the *temporal ordering* of narrative time that can, according to Genette, generally be defined as the relation between the order of sequences as they are causal-logically encountered in the narrative material on the one hand and the (re-)ordering of said sequences as they are actually laid out in the present narrative text on the other hand. This in turn means that the creative freedom of fictional narration also allows for example for the achronological restructuring of its different events with the end result being another recombination that deviates from its source. For this achronology Genette provides the two specific terms *prolepse* and *analepse* that denote the earlier telling of a later event and the later telling of an earlier event respectively.[47] In the context of detective fiction, the analeptical chronology is thus

[44] Cf. Todorov (1973): p. 14 ff.
[45] Cf. Genette, Gérard (2014): The Architext. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, pp. 210–218.
[46] Cf. Genette, Gérard (2010): Die Erzählung. 3rd Edition. Paderborn: Wilhelm Fink, p. 12.
[47] Cf. Genette (2010): p. 17 ff.

of particular importance, since it is a recurring motif within the narration of criminal cases and their subsequent solution to (fully) reconstruct the crime in question at a later point in time and only through the perspective of the detective, who has finally arrived at the end of his deductions regarding the identification of victim as well as perpetrator.

Closing on the current state of the art within the field of formal literary study and its more recent evolutions into the digital sphere through theoretical frameworks such as distant reading, one might still ask if and how the roots of Russian formalism, which date back almost a 100 years, can maintain their relevance in the light of today's advancements. Fortunately there exists a quite sizeable amount of papers which have in recent years tried to show the fundamental correspondences between the origin of formal literary study and its digital counterpart that has mostly come to fruition only in the last decade. For example Basil Lvoff suggests in his paper on the connections between Russian formalism and distant reading a general affinity that can already be gathered from their shared goal of formulating the fundamental principles that constitute artistic language use as well as their similar approaches of retrieving them.

> The study of change is, in effect, both a challenge and a tentative solution in RF and DH. As literary features alter from one work to another, new genres emerge; the system of literature never stabilizes, sustaining a boundless universe of facts. Yet, it is this chaos that triggers the theorist's search for structure and meaning. Thus, the countless facts of the past – disjointed, i.e., meaningless by themselves – impelled RF to organize them into a system of historical patterns. Likewise, Moretti's bewilderment by "the great unread" (as Margaret Cohen first called it) made him and his counterparts seek order in graphs, maps, and trees (as the title of his book (2007) has it.[48]

Therefore both schools in essence try to understand literary phenomena as a collection of recurring patterns and structures that can be described through general terms. On a similar note Marek Debnár also tries to reinforce the ties between Russian formalism and distant reading by providing several examples of (implicit) references to already discussed thinkers such as Shklovsky or Propp in the works of their more distant (pun intended) colleagues of the digital age.[49] Furthermore Ilya Kliger recognises the resurfacing of numerous basic formalist notions such as the distinction between story/fabula and plot/syuzhet as well as the general idea that texts are constituted of even smaller entities that represent an adequate object of study within the recent developments of distant reading.[50] Last but not least Fabio Ciotti mentions that the rich tradition and flexible abstractions of formalistic and/or structuralist theory still offer a lot of even untapped potential for the future discipline of digital literary studies that has not

[48] Lvoff, Basil (2021): Distant Reading in Russian Formalism and Russian Formalism in Distant Reading. In: Russian Literature 122–123, p. 32.

[49] Cf. Debnár, Marek (2018): Formalism and Digital Research of Literature. In: Digital Age in Semiotics & Communication 1, 1, pp. 113–120.

[50] Cf. Kliger, Ilya (2021): Dynamic Archeology or Distant Reading. Literary Study between two formalisms. In: Russian Literature 122–123, p. 7 ff.

yet been fully realised by for example the proponents of distant reading, rendering it all the more plausible to explore these traces and paths further.[51]

With all that being said, it is now time to turn much more closely to the school of distant reading and its most important representatives as well as their theories in order to not only comprehend these epistemological evolutions from the historical perspective of their predecessors but rather from the specific intricacies within their own system. Thus, the following pages will serve as an account for that task, while at the same time providing more specification on the corresponding research question that this master's thesis will try to answer with the aid of a suitable theoretical background as well as the practical methods derived from it.

[51] Ciotti, Fabio (2016): Toward a Formal Ontology for Narrative. In: Matlit 4.1, p. 31.

4. *Distant Reading* and *Sherlock Holmes*

While *distant reading* has been – at least as a concept – around a lot longer in the realm of literary studies than its name and general methodological orientation might suggest, the latest attempts of integrating programming skills such as NLP and statistical knowledge as well as quantitative methods in general have opened up a multitude of new possibilities and research interests. This phenomenon can often be directly linked to a stronger focus on interdisciplinary aspirations and especially the emergence of the discipline of digital humanities as a distinct subject at universities around the world in recent years. In its broadest terms *distant reading* refers to a quantitative interpretation of literature, which tries to cover vast amounts of different texts and their basic similarities at once rather than focusing on a few preselected works and analysing them intensively via for example traditional hermeneutic principles. This basic practice of distant reading can then be traced back to as early as the 19th century, where the underlying dichotomy between microscopic traditional methods of literary studies on the one hand and the inclination towards a macroscopic scale of the latter developments on the other hand slowly started to come into fruition.[52]

One of the most – if not *the* most – influential thinkers on the topic of distant reading in recent times can be found in Franco Moretti. Despite the fact that Moretti himself does not posses the technical/mathematical hard skills necessary to produce new practical solutions in the realm of quantitative text analysis, he has written several monographic treatises offering on a theoretical level multiple interesting pathways in the area of computerised distant reading through the means of for example programming and statistical modelling. Therefore Moretti argues for example for the incorporation of abstract forms such as graphs, maps and trees into the existing repertoire of a modern literary scholar in order to visualise and explain structural relations between the different instances within a larger overarching corpus.[53] In the more efficient and faster, simultaneous evaluation of considerably larger amounts of literary texts Moretti sees then also an inductive upward trend which moves "[f]rom individual cases to series; from series to cycles, and then to genres as their morphological embodiment."[54] It is therefore not so much the individual text with its distinct qualities anymore which lies at the centre of the scholar's attention but its role in verifying or falsifying a given claim about the more general adherence to for example principles of genre consistently spanning over a larger array of multiple texts and/or a given time frame. Along those lines Moretti also focuses in his monograph *Distant Reading* more heavily on the genre of detective fiction and especially its prominent representative of Sherlock Holmes. With the help of binary decision trees Moretti then goes on to show

[52] Cf. Underwood, Ted (2017): A Genealogy of distant reading. In: Digital Humanities Quarterly, 11, 2, pp. 1–2.
[53] Cf. Moretti, Franco (2005): Graphs, Maps, Trees. Abstract Models for Literary History. London & New York: Verso, pp. 1–2.
[54] Moretti (2005): p. 17.

that the premise of a (successful) text for a given genre first and foremost hinges on basic narrative operations or devices that can be observed and described as smaller units within the formal structure. In the case of crime stories Moretti furthermore argues that the respective presence or lack of clues within a given narrative arc is conducive to a reader's engagement with the story, thereby also trying to find a differentiating feature of Holmes's famous adventures in contrast to those of his lesser known colleagues.[55]

Building upon these theoretical concepts of Moretti and focusing on the notion that the paradigmatic construction of genre literature proves to be a promising example on which to establish as well as prove general principles of distant reading, Ted Underwood went on to show with the help of predictive modelling that large corpora of genres such as detective fiction, gothic and science fiction display a clear textual coherence across a time frame of about 200 years. Opposite to Moretti's thesis that genres constantly evolve and only stay consistent for roughly 25 years[56] Underwood on the other hand concludes that for example "Poe's stories already display many of the same features that distinguish twentieth-century crime fiction from other genres"[57], thereby going against the notion of cyclical periods of genre development in favour of a more straight forward and stable progression. The main reason why Underwood then chooses to bring computers into his evaluative methods is not so much the sheer scalability regarding quantitative processing but more so the extraction of underlying, implicit structures within the texts that could finally be used as a verification of vague beliefs with the help of formal terms and/or mathematical language. The statistic, probabilistic approach of Underwood's research allows then also for a denial of simple binary decisions and instead provides the respective likelihood of a given text to correspond to a given genre. This furthermore means that one could define certain textual archetypes for a specific genre which might be the ones displaying the highest probability of adhering to the class of a genre.[58] In his monograph *Distant Horizons* Underwood describes the epistemological implications of his methodology in more detail: For example he points out that the formulation of an interpretive hypothesis has to lie at the core when using distant reading in the interest of literary scholarship, so that the measuring, pattern extraction and feature engineering becomes not a goal in itself but rather another way of looking for discerning textual qualities and finally also making a convincing case for them.[59]

---

[55] Cf. Moretti, Franco (2013): Distant Reading. London & New York: Vers, p. 71 ff.
[56] Cf. Moretti (2005): p. 18.
[57] Underwood, Ted (2016): The Life Cycles of Genres. In: Cultural Analytics, May 23, p. 5.
[58] Cf. Underwood (2016): p. 11 ff.
[59] Cf. Underwood, Ted (2019): Distant Horizons. Digital Evidence and Literary Change. London: The University of Chicago, p. 17 ff.

While distant reading and other forms of computational literary studies have also often been criticised for relying too much on numbers and broad categorisations, thereby loosing the individual text and its respective ambiguous sentiment as a direct cause of subsequent readings, out of sight, Underwood provides a convincing argument why quantitative advancements within the humanities do not necessarily have to be a bad thing and could even improve existing methods as well as extract more subtlety in the process:

> In fact, the imprecision of the human world is part of the reason why numbers are so useful in social science: they allow researchers to describe continua instead of sorting everything into discrete categories. Our model of genre will similarly be continuous. We will have to train it on a sample of volumes that librarians have sorted into two discrete groups (fiction and biography) because the human beings who wander through libraries tend to prefer a space organized by clear boundaries. But the model itself doesn't have to be discrete; it predicts the probability that a volume will be perceived as fiction. We can use that probability to test the model in yes/no fashion, treating anything above 50% probability as a yes. But we can also use probabilities to place volumes on a continuum where nothing is purely fictional or nonfictional.[60]

In this light, methods such as distant reading should not merely be seen as an evil nuisance threatening the integrity of the field's traditions but rather as an opportunity to better formulate and prove already existing notions (on a larger scale). Macroscopic and microscopic approaches are therefore not mutually exclusive and may also coexist or even be combined within one and the same methodological framework, as it is pointed out for example by the concept of *Literary Pattern Recognition*, propagated by Long Hoyt and Jean So Richard. The idea of *Literary Pattern Recognition* basically boils down to a blending of both humanistic and computational approaches or the synthesis of close and distant reading.[61] Not only do Hoyt and Richard in their paper subsequently make use of the more modern statistical machine learning models but they also give credit to the more conventional ways of studying individual texts by for example double checking the patterns and results, found by the computer, via human intuition, which becomes especially helpful when looking at the specific texts, which were misclassified by the algorithm. Hoyt and Richard then conclude that the results produced by the computer should be understood as only one half of a larger whole where the technical possibilities – similar to Underwood's philosophy of distant reading – build upon already existing notions of human made assumptions and knowledge but at the same time add more formal nuance to them.[62] Similarly Katherine Bode in her essay *The Equivalence of "Close" and "Distant" Reading* argues for a more interdependent conception of literary studies, with the main goal being the enrichment of both the quantitative and the qualitative side of the methodology.[63] With that being said, Underwood on the other

---

[60] Underwood (2019): p. 20.

[61] Cf. Hoyt, Long and Richard, Jean So (2016): Literary Pattern Recognition. Modernism between Close Reading and Machine Learning. In: Critical Inquiry, 42, 2, pp. 236–237.

[62] Cf. Hoyt and Richard (2016): p. 267.

[63] Bode, Katherine (2017): The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object of Data-Rich Literary History. In: Modern Language Quarterly, 78, 1, p. 94.

hand poignantly voices the main objections, which still seem to run wild on the general premises of the humanities when it comes to the idea of their methodological apparatus suddenly going digital:

> Like the scientists of Jurassic Park, the first four chapters of this book have been "so preoccupied with whether or not they could" use numbers to learn something about literary history, that "they didn't stop to think if they should." In this final chapter, I hope to remedy the oversight by considering the risks of the approach I have taken. Generally, historical research is less risky than cloning dinosaurs. But applying numbers to the literary past, in particular, remains controversial enough that an analogy to Jurassic Park is not absurd. Objections to new methods of analysis run deeper than I have so far acknowledged.[64]

Namely, despite the more positive outlook on the possibilities and promises of distant reading one has to be aware of the fact that there exist still numerous objections against the wider acceptance of such methods within the field of literary studies. This can of course be partly attributed to the hurdle of acquiring the obligatory technical and mathematical aptness for skilfully applying said methods but furthermore it also equally appears to be a symptom of stronger negative preconceptions, which are rather content in sticking to their traditionally narrow confinements of close reading where the extraction of unique, singular observations already suffices. Nevertheless – as the preceding paragraphs have also shown – the persistent introduction of computational methods into the humanities does not bear the same threats equal to the reproduction and subsequent enabling of free roaming dinosaurs. As a more fitting metaphor one might then for example suggest the picture of a big puzzle representing the field of literary studies as a whole where every participant can add pieces at their own scale and pace slowly uncovering the overarching picture in the process.

As already pointed out, the examination of detective fiction and Doyle's stories in particular through the lens of distant reading is in general not a very novel idea. However, past research has been mostly focusing on the direct relation between the stories of Sherlock Holmes and the characteristics of genre, while aspects like narrative structure and other structural consistencies *within* the Holmes series have at best only being hinted at so far. But given the aforementioned structuralist approaches of operationalising and describing the plot trajectory of the detective's adventures as well as the concept of Moretti interpreting these stories as a chain of clues, the liaison of distant reading and Sherlock Holmes still seems to offer much unexplored territory for subsequent inquiries. Along those lines Jernej Habjan's essay *The Bestseller as the Black Box of Distant Reading* builds upon Moretti's clue formula and suggests treating the adventures of Holmes as a sequence of elements which are constituting the overall structure of the narrative. Habjan exemplifies his claims mainly through the Sherlock case *The Adventure of the Speckled Band*.[65] Similarly Lauren Goodlad shows with the help of the story *A Study*

---

[64] Underwood (2019): p. 143.

[65] Cf. Habjan, Jernej (2012): The Bestseller as the Black Box of distant reading: The Case of Sherlock Holmes. In: Primerjalna književnost, 35, 1, pp. 91–105.

in Scarlet that a closer inspection of the narratological formula of Doyle's works reveals a "connection between parts"[66] and may therefore be regarded as a promising subject for further investigations via the methods of distant reading on a larger scale. While these papers and especially their hypotheses add another dimension to the discussion of genre literature and its future role in distant reading or computational literary studies as a whole their practical executions remain again mostly tentative or anecdotal. Thus, the trend that the investigations of detective fiction and/or the stories of Sherlock Holmes appear to be either too macro- or microscopic in their endeavours without providing any midway in between – seems to stay present, even in the face of today's vast possibilities of modern distant reading approaches. This fact then renders it all the more plausible why this thesis will aim to bridge the still present gap within the current state of research by on the one hand 'widening' the viewpoint on the complete works of Sherlock Holmes from a distant reading perspective but then again also 'zooming' in from the other end of the spectrum represented by the likes of close reading. Through exploring this middle ground the methodological approach of this thesis may in addition move closer towards the suggested and already discussed concept of literary pattern recognition, where the blending of both distant and close reading as well as the approximation of macro- and microscopic viewpoints becomes all the more prominent. Since past research has both shown the structural consistency on the level of individual texts and the overarching level of genre, the underlying hypothesis of this treatise can then be formulated as follows: Regarding the narrative arc of Doyle's detective fiction this work presupposes a consistent pattern which can be traced throughout the adventures of Sherlock Holmes via computational methods of distant reading such as the extraction of key plot elements and their subsequent (statistical) quantitative measuring as well as their visualisations as curves. With all that being said it becomes furthermore also necessary to clarify in detail how the abstract concept of plot can actually be dissected into smaller units of homogeneous elements and subsequently measured via computerised methods. This opens up room for another important question, which has already been discussed extensively by numerous proponents of distant reading and its adjacent fields, therefore warranting a separate chapter for not only probing said arguments but also sketching out a possible pathway for answering the laid out research question.

[66] Goodlad, Lauren M. E. (2020): A Study in distant reading: Genre and the Longue Durée in the Age of AI. In: Modern Language Quarterly 81, 4, p. 504.

5. Measuring Plot in the Age of Distant Reading

In a lecture held by Kurt Vonnegut titled *The Shapes of Stories* the American author points out – in an albeit roundabout, humorous manner – that most if not all plots of literary works can be described via drawing curves, representing thereby the basic trajectories of a given narrative such as the emotional states of its protagonists.[67] While Vonnegut's explanations in this instance remain mainly exemplary – especially given the author's reliance on nothing but a chalkboard to bring his point across – this rudimentary theory issued by a writer of fiction himself nevertheless already suggests that there exists a certain kind of merit in abstracting and consequently also modelling literary plots, insofar as one can in addition provide the obligatory framework in order to properly assess the important features of a literary text. This is exactly where the developments of recent years within the field of computational literary studies and distant reading come into play: Not only do these newer approaches provide the necessary terminological pungency, which will be laid out in the following paragraphs, they even enable the researcher interested in the patterns of plot to measure much larger amounts of texts simultaneously (and without the employment of a chalkboard as a rather limited intermediary).

As Moretti suggests in one of his pamphlets of the *Stanford Literary Lab*, a research group founded on the premise of leading the progress of distant reading within the humanities, every research question that tries to incorporate the quantitative approach of distant reading into its literary studies requires first and foremost a clarification of how and why one can actually adequately transfer the essence of a literary text – or at least parts of it – into the computational/digital sphere. Therefore Moretti introduces – following the tradition of Russian formalism and structuralism – the already referenced term of *operationalising*, which basically denotes the transformation of a concept into a series of operations with the intention of making these operations then quantifiable and also processable by a machine.[68] In the case of the present master's thesis this would mean taking the narratological concept of the plot and breaking it down into smaller, similar units that express the meaning of a given narrative tale when linked together. To put it more concretely and place it in the context of recent advancements in the field of distant reading: The computational examinations of plot mostly followed the aforementioned structuralist dichotomy between plot and story, and while the latter concept has been receiving the most attention across the different papers and treatises published so far, this thesis will instead focus more on the first one in order to find a suitable way of modelling the arc of a given text. Andrew Piper in his monograph *Enumerations* notes that the most established method of measuring as well as visualising these kinds of narrative operations has been found in emotion analysis where certain hotspots of distinct

[67] Cf. Vonnegut, Kurt: Kurt Vonnegut on the Shapes of Stories. In: https://www.youtube.com/watch?v=oP3c1h8v2ZQ [10.4.23].
[68] Cf. Moretti, Franco (2013): Pamphlet 6. "Operationalizing": or, the function of measurement in modern literary theory. In: https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf [9.4.23].

emotional intensity are identified and subsequently used as indicators of a story's trajectory.[69] This notion also corresponds to the findings of cognitive science in the realm of formalising narrative summarisations where the role of emotional reactions are conducive to the remembrance of specific events and thereby also crucial for their subsequent retelling by a given speaker. The resulting smaller operations that constitute a plot in this model are in this instance referred to as *affect states* which are marking clear, relative changes in mood.[70] Jodie Archer's and Matthew Jockers' study on a large collection of commercially successful bestseller novels also shows on the basis of emotion analysis that their plot trajectories can be grouped into seven distinct shapes. The upwards and downwards turn experienced by the characters within a given novel are therefore then extracted and used as a basis of reconstructing the plot with the aid of curves[71]– a quantitative, implementation which in essence echoes again the principles of Vonnegut's aforementioned, conceptual approach.

Furthermore Piper introduces a second approach of modelling plot via a more general evaluation of *linguistic drift*. By taking then the vocabulary used as a whole into account and examining its shifts as the narrative time progresses, Piper finds for example statistical significance in the usage of proper nouns as well as in the density of individual words used within a given sequence. In this instance the change within a narrative text is therefore mainly captured on a semantic level whereas one looks for sudden and distinct variations within the average lexical word pool.[72] In this regard Piper develops, based on Augustine's *Confessions*, a model of distant reading which aims to split a given novel into two parts of roughly equal size in order to then extract a sense of 'turning around' or 'before and after' on a linguistic level as the reader flips the pages. According to Piper this *idea of conversion* is conducive to the modern novel establishing itself for the most part in the 19th century and can be measured via firstly transforming a text into multiple subsegments which are then represented by vectors which can finally be compared according to the occurrences and frequency of similar words. The higher the similarities between two selected vectors the closer they also appear within a coordinate space suggesting therefore the existence of a distinct cluster. Along those lines Piper goes on to prove that the binary pattern of two distinct clusters holds for the most part true across a range of more than 400 different novels of modern literature whereas the chapters of the first half of a narrative appear much more tightly connected than the chapters of the second half, suggesting therefore also a general expansion of linguistic style throughout the progression of a plot.[73] Another important point that Piper makes in regard to his methodological

---

[69] Cf. Piper, Andrew (2018): Enumerations. Data and Literary Study. Chicago & London: The University of Chicago, p. 43.
[70] Cf. Lehnert, Wendy G. (1981): Plot Units and Narrative Summarization. In: Cognitive Science, 4, p. 294.
[71] Cf. Archer, Jodie and Jockers, Matthew L. (2016): The Bestseller Code. Anatomy of the Blockbuster Novel. New York: St. Martin's, p. 82 ff.
[72] Cf. Piper (2018): p. 45 ff.
[73] Cf. Piper, Andrew (2015): Novel Devotions. Conversional Reading, Computational modelling, and the Modern Novel. In: New Literary History, 46, 1, p. 65 ff.

framework is that he sees the quantitative approaches of distant reading as an enabler for further, more qualitative hypotheses about individual characteristics of smaller subsets of the initial corpus. Thus, he vouches – similar to the approaches already discussed in chapter 4 – for an interconnected combination of distant and close reading in order to arrive at a satisfying model of *computational hermeneutics*:

> My aim in this essay is to offer a methodological polemic against the either/or camps of close versus distant reading or shallow versus deep that have metastasized within our critical discourse today. I want us to see how impossible it is not to move between these poles when trying to construct literary arguments that operate at a certain level of scale (although when this shift occurs remains unclear). In particular, I want us to see the necessary integration of qualitative and quantitative reasoning, which, as I will try to show, has a fundamentally circular and therefore hermeneutic nature. As we move out from a small sample of texts toward larger, more representative populations and back into small, but now crucially *different* samples, such circularity serves as the condition of new knowledge, of insight per se. It puts into practice a form of conversional reading, one whose *telos* is not a single, radical insight, but instead an iterative and circular process that can serve as a vehicle for conceptual change.[74]

Piper's linguistic approach of measuring the shifts and turns of narrative texts can furthermore also be perceived in the attempts of Ryan Boyd et al., where presegmented texts are compared in regard to their occurrences of certain *function words* such as prepositions, articles, pronouns and auxiliary verbs as well as *cognitive* words marking a given story's progression. Boyd et al. then go on to define three distinct phases within a given plot which consist of the initial *staging*, the *plot regression* or its progression and finally the *cognitive tension* denoting the climax of the later parts of the story. In order to quantify these patterns through a corpus of over 2500 novels the rates of the function and cognitive words are then calculated for each subsegment of a given text and set in relation to the total number of words in said text. Given the quantitative evidence retrieved via their methods, Boyd et al. similarly conclude that the pattern of story telling observed across their large corpus remains fairly consistent, thereby suggesting a schematic progression or narrative arc that displays high rates of certain words at certain points in narrative time across a wide array of different texts.[75]

A third approach of measuring plot via the means of distant reading has been mainly brought forward by the already mentioned Matthew Jockers and revolves around the concept of topic modelling. Basically speaking, a given topic is defined by a cluster of words frequently recurring in conjunction and can be retrieved by scanning large corpora consisting of different documents and subsequently assigning high probabilities to word clusters that often co-occur within a given corpus. The most straight forward approach of applying this method to a literary text would be to extract *one* topic per text or document which corresponds to the most frequently recurring cluster of specific words and comparing these most prominent topics then across the different texts within the corpus in regard to its distribution

---

[74] Piper (2015): p. 69.

[75] Cf. Boyd, Ryan L. et al. (2020): The narrative arc: Revealing core narrative structures through text analysis. In: ScienceAdvances 32, 6, pp. 1–9.

and similar measures. While this method can already grant some interesting insights – especially when it comes to the research of literary history and its development across multiple periods – Jockers furthermore provides a more fine grained solution which looks closer into the distribution of *multiple* topics on the individual text level: Here a given text is again split into multiple subsegments, whereas each subsegment gets fed separately into the topic modelling algorithm in order to find for example distinctive points across the progression of novelistic time where a given topic occurs the most relative to other segments.[76] Building on this concept of using topic modelling as an enabler for extracting plot structures from narrative texts, Benjamin Schmidt also shows with the help of about 80.000 movie and television show transcripts that there exists a strong correlation between certain topics and their point in time when they appear within a given movie or show. In addition to that Schmidt manages to extract several distinctive plot arcs over the respective genres contained within his dataset, therefore suggesting a correlation between archetypical narrative structures and specific genre conventions. Last but not least the study also provides further, more formulaic evidence to the general notion that the contents of a narrative (story/fabula) have to be first and foremost transported through interconnected operations forming together the figure of an arc (plot/syuzhet) in order to establish a sense of overarching progression which can then be processed by the reader/viewer.[77]

Thus, the state of the art within the current research regarding the measurement of plot through the means of distant reading reveals three distinct branches, which generally speaking involve the use of emotion analysis, linguistic change and topic modelling respectively. Since the approach of topic modelling seems to be for the most part better suited for larger corpora and studies situated at a grander scale, this master's thesis will instead primarily rely on the first two approaches in order to retrieve meaningful results from its object of study. Nevertheless the fact that crime fiction has also often been characterised by its reliance on repetitive successions of distinct events or actions by its main characters, warrants (at least in part) the additional exploration of some topic or event extraction (EE) methods which aim to uncover some general trends and commonalities in the realm of content. How a particular implementation of these concepts can be achieved exactly, will be described in detail in the following two chapters.

[76] Cf. Jockers, Matthew L. (2013): Macroanalysis. Digital Methods and Literary History. Urbana, Chicago and Springfield: University of Illinois, p. 124 ff.

[77] Cf. Schmidt, Benjamin M. (2015): Plot Arceology: a vector-space model of narrative structure. In: 2015 IEEE International Conference on Big Data, ed. by Howard Ho et al., pp. 1667–1672.

6. Dataset and Methodology

First and foremost the collected stories of Sherlock Holmes, written by Conan Doyle, consist of 60 individual texts which are also considered as 'the Canon'. These texts were published over a period of around forty years, ranging from 1887 to 1927.[78] The four novels and 56 short stories are held to this day in high regards by its large fan community which is a testament to the fact that Doyle's stories display a fairly consistent formula with engaging narrative structures and recurring, recognisable plot elements throughout their publishing lifespan.[79] This huge influence can furthermore be seen for example in the wide array of different adaptations, fan fictions and even unofficial continuations long after the detective's official retirement,[80] and last but not least also in the (playful) blending of fictitious and non-literary reality when it comes to the legacy of the infamous cult surrounding Holmes and his sidekick Watson.[81] With that being said, this master's thesis does not take such tangents and deviations from its original source material into account – especially given the fact that such an endeavour would just not be feasible within the scope of this work – and instead focuses solely on the canon, which was written by Doyle. Thus, the dataset, on which further methods of distant reading and especially the analysis of plot are based on, consists of all the sixty official stories, which were downloaded from a Sherlock Holmes internet archive as separate text files.[82]

As already suggested the methodological framework for extracting the plot structures of the corpus at its core entails a mixture of both the tracing of general linguistic change akin to the approach laid out by Piper and the analysis of emotional change throughout a given narrative as it has mostly been brought forward by Jockers. The computational part of the analysis has been done in the programming language *Python*, given its general flexibility and ease of use as well as its wide array of additional libraries when it comes to text processing or NLP modelling. The first stage of the pipeline, after the data had been loaded, revolved around the various necessary preprocessing steps that are required in order to apply further evaluations on the corpus. Each individual text file has then been considered as an instance of a document representing first and foremost the whole narrative content as a single string respectively. Subsequently some standard transformations have been applied to these strings such as punctuation removal, tokenisation and lemmatisation. Since word categories like articles or conjunctions are central for measuring linguistic change throughout a given narrative – as mentioned already in the context of Boyd et al. – the removal of stopwords were left out intentionally for some specific tasks. Last but not

[78] Cf. Redmond, Christopher (1993): A Sherlock Holmes Handbook. Quebec: Simon & Pierre, p. 9 ff.
[79] Cf. also Campbell, Mark (2007): Sherlock Holmes. Somerset: Pocket Essentials, p. 13 ff.
[80] Cf. for example D'haen, Theo (2017): Sherlock's Queen Bee. In: Crime Fiction as World Literature, ed. by Louise Nilsson et al. New York: Bloomsbury, p. 233.
[81] Cf. Tobin, Vera (2006): Ways of reading Sherlock Holmes: the entrenchment of discourse blends. In: Language and Literature 15, 1, pp. 73–90.
[82] Cf. https://sherlock-holm.es/ [21.4.23].

least the individual documents were furthermore split into a consistent number of equal-sized segments which allows for a better representation and/or visualisation of a plot's trajectory over narrative time. Here Boyd et al. suggest the choice of five segments which granted them the best outcomes and also correspond to the five traditional stages of a narrative dating back to the ancient poetics of Aristotle.[83] In addition, this segmentation may finally result in a vectorisation of a given document and its computed scores/features allowing for easier means of comparison with other documents through for example cosine similarity. The general pipeline has been mainly implemented with the help of libraries such as *NLTK*[84] and *spaCy*[85], which provide numerous functions for common NLP tasks such as POS tagging or NER and even more sophisticated methods like EE, *scikit-learn*[86] and *SciPy*[87], which are extensive machine learning frameworks that include a large variety of widely used statistical algorithms, and finally also the common data processing tools of *pandas*[88], *NumPy*[89] as well as *matplotlib*[90] and *seaborn*[91], which expand upon *Python*'s basic capabilities of representing and manipulating data via more refined object types and visualisation options.

For extracting and analysing the structure of linguistic change the underlying operation includes the creation of lists containing certain keywords of interest which are then matched against the segments of a given document. For each segment the frequency of the keyword(s) is calculated relative to all the words over the same document in order to also account for the varying length of different documents within the corpus. For the measurement of the basic narrative arc of the documents the necessary keywords lists have been derived from a dictionary provided by Boyd et al. which matches function words as well as cognitive tension words to their respective role of staging, plot progression or the production of narrative tension.[92] Since the aforementioned literature surrounding the narrative features of detective fiction also considers the particular, often times analeptical usage of temporal structures as an important feature of the genre,[93] the linguistic analysis furthermore entails some inquiries into the temporal references and relations within the stories of Sherlock Holmes. As Lindsay Ross notes in her study regarding the patterns of story arcs within palliative care conversations, these temporal structures can best be observed by using a POS tagging approach in order to retrieve the respective tense expressed

[83] Cf. Boyd et al. (2017): p. 3.
[84] Cf. https://www.nltk.org/ [18.4.23]
[85] Cf. https://spacy.io/ [15.10.23].
[86] Cf. https://scikit-learn.org/stable/ [15.10.23].
[87] Cf. https://scipy.org/ [15.10.23].
[88] Cf. https://pandas.pydata.org/ [15.10.23].
[89] Cf. https://numpy.org/ [15.10.23].
[90] Cf. https://matplotlib.org/ [15.10.23].
[91] Cf. https://seaborn.pydata.org/ [15.10.23].
[92] Cf. https://osf.io/wpcx8 [17.4.23].
[93] Cf. again Nelles and Williams (2021): pp. 191–192.

by the inflection of a given verb.[94] While this approach of mainly counting word frequencies and subsequently comparing their distributions is certainly not the most complex or 'sophisticated' method within the possibilities of distant reading, it can nevertheless "provide a good starting point for much research"[95] and even produce meaningful results on its own already.[96]

Furthermore the different documents within the corpus have also been analysed according to their emotional changes throughout their progression. While Jockers has developed his own popular package for the programming language R called *Syuzhet,*[97] which basically consists of a number of functions for handling the emotion analysis of (literary) textual data and its subsequent visualisation as a narrative arc, said package has also been subject to a number of critiques in more recent debates.[98] For example Hoyeol Kim in her article about the limits and progress of *Syuzhet* points out that there has arisen a constant accumulation of various qualms within the research community of digital humanists, often times pointing out the production of erroneous results and similar inaccuracies by the tool. These qualms have so far only been partly addressed by Jockers and his research lab while major fixes still have not been provided.[99] Given the current state of *Syuzhet,* Kim also suggests using a different, more recent sentiment analysis tool called *VADER* which is available both for *R* and *Python* and appears to be more transparent as well as much less limited in the different methods or functions it provides out of the box.[100] Therefore instead of using the already established but more error-prone implementation by Jockers, the subsequent analysis rather relied on *VADER*, thereby providing on the one hand more assurance in regards to future (scientific) scrutiny whilst at the same time offering a new perspective to the existing research within the field of emotion analysis and distant reading.

Generally speaking, extracting the emotional qualities of a given input text requires the calculation of a valence score which is based on a scale assigning very negative words with very low negative values and very positive words with very high positive values accordingly, while zero stands for neutral sentiment. These individual sentiment scores of words are then provided through a large dictionary which can be matched and compared with a given input sequence. Most commonly the valence score is

---

[94] Cf. Ross, Lindsay May (2019): Exploration of Story Arcs in Palliative Care Conversations Using Natural Language Processing. In: UVM Honors College Senior Theses 293, pp. 11–12.

[95] Archer, Dawn (2016): Data Mining and Word Frequency Analysis. In: Research Methods for Reading Digital Data in the Digital Humanities, ed. by Gabriele Griffin and Matt Hayler. Edinburgh: Edinburgh University, p. 73.

[96] Cf. also McClure, David (2017): Distributions of words across narrative time in 27,266 novels. In: https://litlab.stanford.edu/distributions-of-words-27k-novels/ [17.4.23].

[97] Cf. Jockers, Matthew L. (2020): Introduction to the Syuzhet Package. In: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html [17.4.23].

[98] Cf. Swafford, Annie (2015): Problems with the Syuzhet Package. In: https://annieswafford.wordpress.com/2015/03/02/syuzhet/ [17.4.23].

[99] Cf. Kim, Hoyeol (2022): Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons. In: http://www.digitalhumanities.org/dhq/vol/16/2/000612/000612.html#p10 [17.4.23].

[100] Cf. https://github.com/cjhutto/vaderSentiment [17.4.23].

computed on a sentence level through a so-called compound score where the scores of the individual words within the emotional dictionary are summed up and finally normalised in order to provide a consistent metric for the comparison between sentences/sequences of different length.

In the case of *VADER* the valence scale ranges from -4 to +4 and is provided via a *Wisdom of the Crowd* (*WotC*) lexicon which consists of manual labels from human participants. In contrast to more traditional emotion analysis tools *VADER* furthermore provides five additional heuristics which incorporate the detection of word-order sensitive relationships as well. These heuristics entail the inclusion of punctuation such as exclamation marks or capitalisation to emphasise certain emotions, the consideration of degree modifiers and their function of increasing or decreasing emotional intensity, the role of contrastive conjunctions such as 'but' which can mark a sudden shift in emotion and finally also the leveraging of trigrams in order to detect negations that actually flip the meaning of an accompanied adjective (e. g. 'The food here isn't really all that great.'). The benchmarks provided by C. J. Hutto and Eric Gilbert in their paper about the development of *VADER* therefore also showcase a considerable increase in accuracy score when compared to eleven other highly regarded sentiment analysis tools. Said improvements in accuracy are furthermore not only limited to the domain of social media texts but can be observed across different tasks and types of texts, portraying a promising ability to generalise well throughout.[101] Finally, the fact that *VADER* is built on top of *NLTK* and can be called within the same library also allows for a more seamless integration of the emotional analysis task into the general NLP pipeline.

While it has already been pointed out in the previous chapter that, besides the more pronounced approaches of linguistic change and emotion analysis, the third option of topical analysis might also be able to provide additional insights into the narrative structure of the chosen texts, it is still necessary to clarify how exactly a quantitative study situated in the realm of detective fiction might profit from such a distillation of more abstract devices. Fortunately, the preceding research has already been able to identify a certain linkage between story structures of crime-related plots and the extraction of underlying and common actions or events through the means of computational measures. For example Aditya Motwani et al. show in their paper titled *Extracting Evidence Summaries from Detective Novels* that the often highly formulaic basic structure of detective stories appears to be a quite suitable candidate for NLP tasks such as extracting the basic abstract concepts that constitute the narrative and even subsequently predicting the culprit of a given plot via probabilistic models.[102] In a similar vein Louis and Engelbrecht argue in their paper on the unsupervised identification of textual relations that criminal

[101] Cf. Hutto, C. J. and Gilbert, Eric (2014): VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Proceedings of the International AAAI Conference on Web and Social Media 8, 1, pp. 216–225.

[102] Cf. Motwani, Aditya et al. (2019): Extracting Evidence Summaries from Detective Novels. In: Proceedings of the Text2StoryIR'19 Workshop, ed. By A. Jorge et al.. Germany: Cologne, pp. 1–7.

novels such as *The Mysterious Affair at Styles* by Agatha Christie display a high interconnectedness of different parts throughout their story progression and represent therefore a fitting example on which to base analytical methods such as operationalisation and the extraction of thereby resulting relations between different story elements.[103] Since the aforementioned *Python* library *spaCy* already provides an easy way of formulating a pipeline for the task of EE through the inclusion of the native entity type EVENT, going forward it is then easy to construct a function for that purpose which not only readily provides one with all the relevant entities of interest but which may also be refined via the introduction of additional rulesets in order to then take for instance co-occurring persons and temporospatial expressions into account as well.

A second common approach for retrieving the basic underlying content of a given document can be found in latent semantic indexing (LSI). In general LSI denotes an NLP technique for analysing the relationships between terms and documents based on their semantic meaning, aiming to overcome the limitations of traditional keyword-based approaches by considering the contextual meaning of words. LSI has several advantages: For example it can capture said underlying semantic relationships between words and documents, even if they do not share the exact same terms and helps in accounting for more nuanced semantic qualities of words such as synonymy and polysemy. The basic process behind LSI revolves around breaking the given texts of interest down into three distinct matrices which capture the relation between documents and latent concepts as well as their importance. The dimensionality of these matrices is then reduced through mathematical transformations so that a given document is finally represented as a vector allowing for easy comparability between the occurring topics within different texts.[104] Including LSI into the here laid out pipeline should consequently enable the examination of topic frequencies over narrative time and see for example whether there exist specific segments that on average display a higher density of topics across the whole corpus.

Last but not least the third major part of the NLP pipeline consists of evaluating the previously computed features and insights via statistical modelling. As scholars within the discipline of distant reading such as Underwood have already noted, the deployment of common probabilistic machine learning algorithms in conjunction with more basic methods of feature engineering and statistical tests can act as a sound numerical verification method (or falsification for that matter) which grants clear decision boundaries to already postulated notions by for example the traditional literary scholar making more subjective claims and/or the digital humanist exploring his datasets while identifying perceived patterns. Thus, the additional usage of unsupervised approaches, such as clustering, or supervised classification, such as

[103] Cf. Louis, A.L. and Engelbrecht, A.P. (2011): Unsupervised discovery of relations for analysis of textual data. In: Digital Investigation 7, p. 156.
[104] Cf. also Deerwester, Scott et al. (1990): Indexing by Latent Semantic Analysis. In: http://wordvec.colorado.edu/papers/Deerwester_1990.pdf [15.10.23].

logistic regression and support vector machines, grants one the ability to not only assign specific probabilities and similar metrics to the given claims but might also reveal even more (high-dimensional) correlations along the way. Said correlations may finally result from the interactions between the different parts, which have been retrieved beforehand via the more narrow feature engineering steps within the overall pipeline, which will be laid out and explained in much more detail in the subsequent chapter.

7. Results and Discussion

As pointed out before, the general research goal of the present master's thesis lay in identifying certain key elements conducive to the constitution of narrative plots, which can then be compared with each other across the whole corpus and may finally help in the verification or falsification of the claim that there exists a certain recurrence of similar patterns throughout the genre literature of detective fiction. With the help of distant reading this main task has therefore been laid out and explored extensively within a *Jupyter Notebook* in the programming language *Python*. While most of the code contained in the notebook as well as the resulting visualisations and similar outputs will be discussed in the following paragraphs of this chapter in considerable detail, the full source code along with its explanatory documentation can also be retrieved independently via GitHub.[105] The following discussion therefore provides a pipeline for the stated research endeavour, which includes subtasks such as the initial preprocessing of the text corpus, exploratory data analysis, the computation of the necessary individual features for plot analysis and finally also the (statistical) evaluation of said features. Although the pipeline in this case only sticks to the application on the stories of Sherlock Holmes, moving forward it may also be of considerable use for further research endeavours such as analysing other specimens of detective fiction and (albeit with some further adjustments) even other genres of literature and their respective narrative arc(s).

The first stage of the pipeline mainly revolved around common and mandatory preprocessing steps with the end goal of bringing the different texts into a consistent and standardised format. While the complete works of Sherlock Holmes were initially retrieved as individual text files, the files were initially put through a sequential loop that read each file line by line in order to extract its respective title, remove some unnecessary lines and finally build a dictionary with the keys corresponding to the respective titles and the values to each text accordingly. The created dictionary was finally converted into a *pandas* dataframe for providing more flexibility when it comes to the computation of additional values and/or introducing supplementary data granularity.

[105] Cf. https://github.com/DavidSiegl/Sherlock-Holmes-Narrative-Arc [17.10.23].

```
df_dict = defaultdict(str)
title = None

for f in os.listdir('./data/sherlock/'):
    with open('./data/sherlock/' + f, 'r') as text:
        lines = text.readlines()
        text.seek(0)
        ls = list()
        for number, line in enumerate(lines):
            if number == 4: # extracting the title of each story as a key
                title = line
                title = title.strip()
            elif number not in [4, 6] and line != "\n": # skipping
unnecessary lines
                ls.append(line.strip(" "))

        conc_text = "".join(ls)
        conc_text = re.sub(r'[-]{10}\n([\S\s]+)', '', conc_text) #
removing additional lines at the bottom of the text files
        conc_text = conc_text.strip()
        df_dict[title] = conc_text

df_sherlock = pd.DataFrame.from_dict(df_dict, orient='index',
columns=['text'])
df_sherlock.index.name = 'title'
```

Figure 1: Preprocessing

Speaking of more data granularity, a new column was added to the created dataframe containing for each title the year of its first publication in *The Strand*. Since this information was not provided in the original text files, the individual years were scraped manually from the web and put together into an array in correct order. Besides that, standard preprocessing steps such as the removal of punctuation, stopwords and numerical values as well as tokenisation and lemmatisation were applied to the textual data. Note that the different steps of feature engineering, which will be discussed further down below, often required different formats of input. For example the *VADER* package for sentiment analysis also takes punctuation into account and works on a sentence level while the measuring of narrative coherence looks specifically for the distribution of stopwords in a given text sequence. This aspect is then reflected in the fact that the preprocessing pipeline branches off into different subtasks, whereas some branches also include the removal of stopwords and punctuation while others do not and therefore stay closer to the raw text format. Since describing and providing every single code block of these ‘branches’ would introduce a lot of unnecessary redundancy into this chapter, they have been left out intentionally, so that one should instead consult the underlying notebook if he wants to look further into all the details. Nevertheless, remarks about certain decisions, which have been deemed especially important and/or

noteworthy to the understanding of the overall thought process behind the construction of the pipeline, will be mentioned in the corresponding sections.

As a next crucial step within the preprocessing of the corpus it was mandatory to segment each text into five parts of equal length in order to compute local values for certain points in narrative time, which can then be compared according to the notion of a narrative arc. In this case it proved to be most beneficial to opt for a value of five segments per text because it firstly corresponds to the traditional notion of the dramatic climactic progression of rise and fall and is secondly also supported by the relevant literature which has been consulted beforehand.[106] To back up this decision further, a small sample size from the complete pool of the instances was hand-picked and segmented manually in order to align the general notion of text segmentation with the more subjective judgement of the literary scholar via cross-validation. In essence a segment number of five with the individual stages being the initial exposition, the set-up of the main plot and its motifs, characters, etc., the climax of the action, the peripetia of the story and finally the solving of the case made the most sense, when being compared to some other scenarios of segmentation with differing number sequences. For example one of the most popular stories within the corpus, *The Hound of the Baskervilles,* could best be described through the employment of heuristics pertaining to sequences of five stages with the general labels being the introduction to the client and his case alongside the arrival at Baskerville Hall, the subsequent investigation, the uncovering of clues, the confrontation, and finally the conclusion of the case. Similarly, another popular text, *The Adventure of the Speckled Band*, displays in general an almost identical schema, which revolves around the visit of the client, the examination of the crime scene, the unfolding of the mysterious events, the confrontation and finally the conclusion of the case in that particular order. Last but not least the texts *The Adventure of the Blue Carbuncle*, *The Adventure of the Engineer's Thumb*, *The Adventure of the Musgrave Ritual*, which complete the selection of the drawn samples, turned out to follow in essence the same formula.

The following function therefore takes as an argument a given string of text and returns the same text segmented into five equal parts. When applied to a dataframe and combined with the native *pandas explode* function the same function generates then a new dataframe, in which each row corresponds to a new segment. Additionally another column is created which contains a number from 1 to 5 in order to denote the respective segment number for each row entry in the dataframe. This makes it then easy to for example group the data according to the different segments and discern certain trends amongst and/or between different segment numbers within the corpus.

[106] Cf. Boyd et al. (2020): pp. 2–3. See also for example the classical Aristotelian model of dramatic progression.

```
def segment_text (text):
    segment_length = len(text) // 5
    segments = [text[i:i+segment_length] for i in range(0, len(text),
segment_length)]
    return segments[:5]

df_sherlock['segments_narrative'] =
df_sherlock['text_prepro_narrative'].apply(segment_text)

df_sherlock_segments_narrative = df_sherlock.explode('segments_narrative')

values = list(range(1, 6)) * ((len(df_sherlock_segments_narrative) // 5) +
1)
df_sherlock_segments_narrative['segment_num'] =
values[:len(df_sherlock_segments_narrative)]
```

Figure 2: Segmenting text into equal parts

Before actually diving into the main and most extensive part of the pipeline – i. e. the computation of the different features conducive to the constitution of the genre literature of detective fiction – it is still wise to retrieve additional understanding on the data at hand at a more basic level. Thus, after the described preprocessing steps have been applied to the (originally) raw text files, the following step turns first and foremost to the methods of exploratory data analysis (EDA), with the main intention being to not only create some initial visualisations but also gather better insights on for example the distributions of the individual instances. This in turn then might enable one to make better informed decisions regarding further normalisations or optimisations of the precise input values, which will eventually be fed into statistical algorithms and similar evaluation methods.

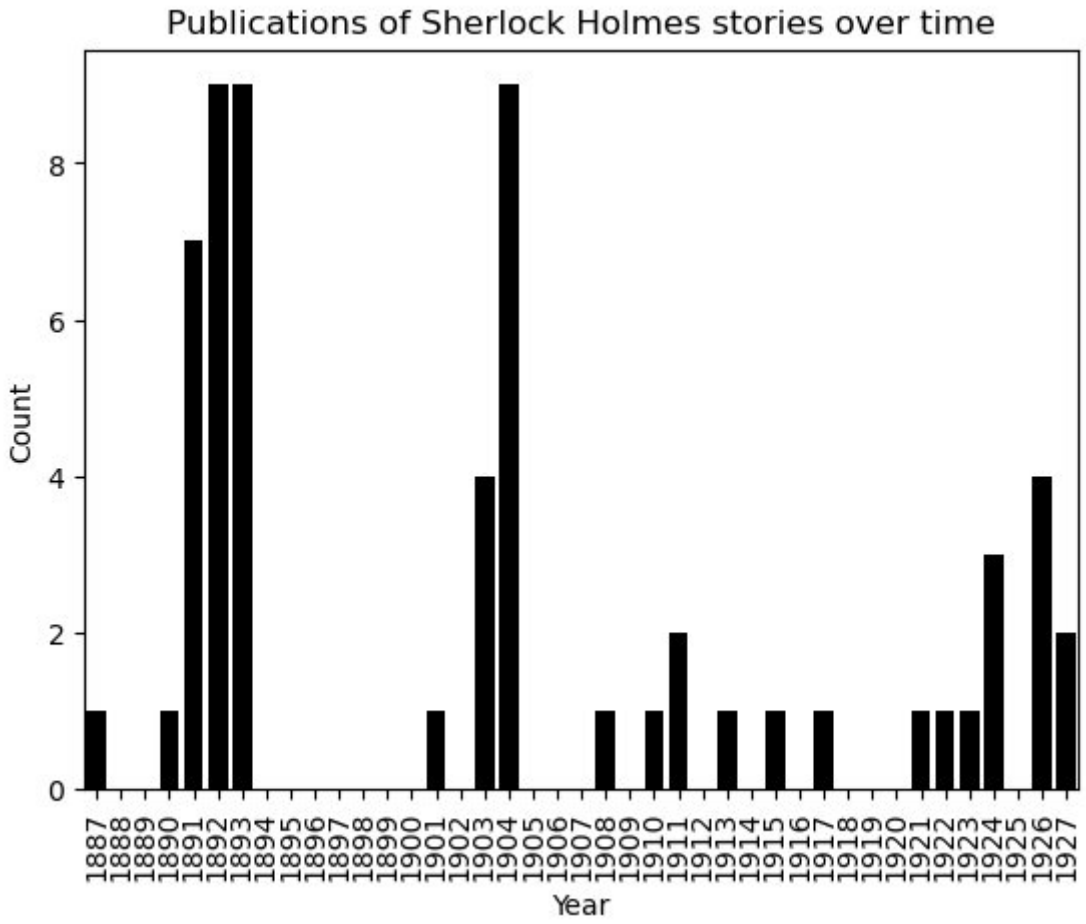

Figure 3: Distribution of stories over time

Firstly, the examination of the publication counts of Sherlock Holmes stories per year displays a significant peak around the early 1890s, which also corresponds to the steady rise in popularity of Doyle's hero. After 1893 the count of publications suddenly plummets and stays that way until the turn of the 20th century. This phenomenon can be explained by Doyle's, as already explained, (temporary) decision to 'kill off' his protagonist. The continuation and subsequent rise of publications then corresponds to the literary 'resurrection' of Holmes, which reaches as far as into the late 1920s until the detective's last case and his official retirement. With that being said, a first lead can already be derived from this trend, which makes it appear sensible to also include the question whether or not there exists a certain change in the plot structuring between these two 'periods' of Holmes' publication life cycle into the subsequent analysis. Along this notion one might for example expect to find noticeable changes in the numerical features of the textual data as the narrative style of Doyle matures (and as he also grows slightly sick of his subject for that matter). This echoes lastly one of the main goals of distant reading, which has also been addressed beforehand – namely the combination of traditional literary beliefs with more empirical and objective findings in order to back up long before made claims and (maybe) even expand upon them via more nuanced elaborations expressed within the realm of numbers and quantifications.

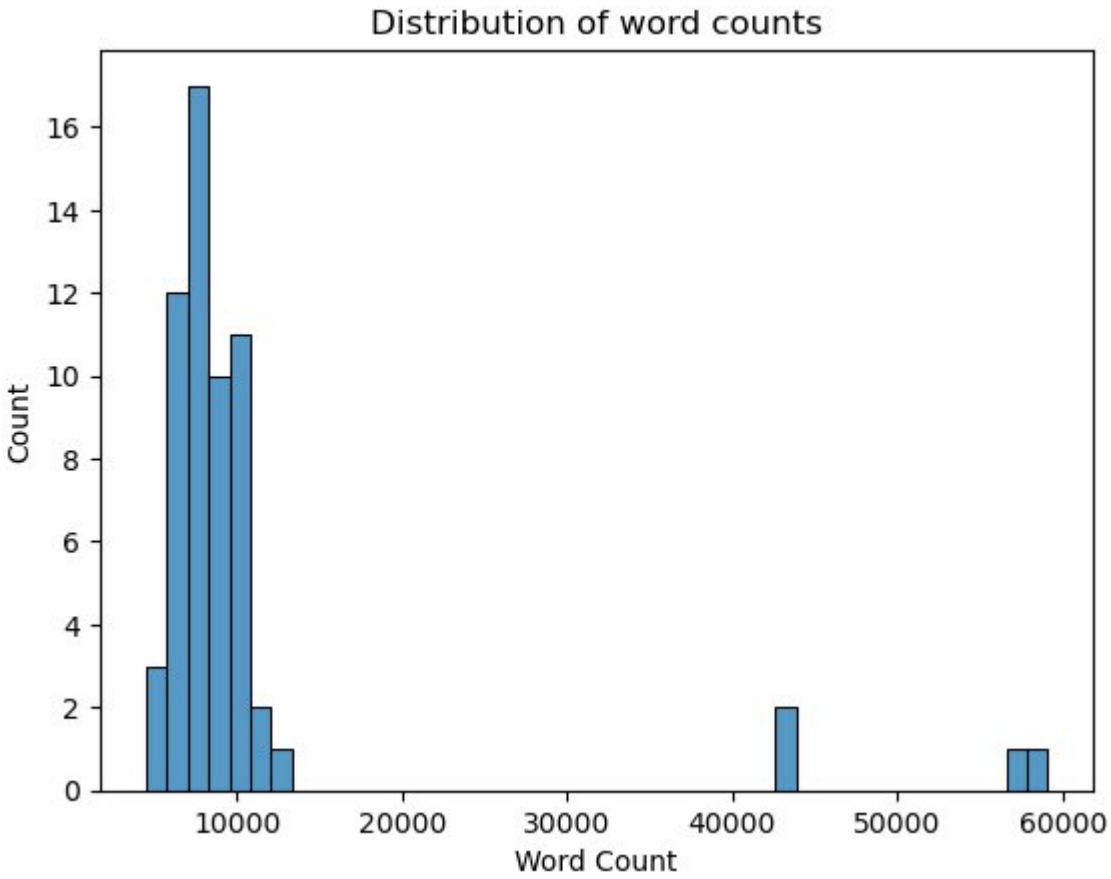

Figure 4: Distribution of word counts

Speaking of latent distributions within the dataset, another important observation can be made when looking at the distribution of word counts across the different texts: This basic histogram already displays a considerable skew of the data towards shorter texts of around 10000 words, which can be mainly explained by the fact that the original Sherlock Holmes stories were for the most part published as a series of short stories within the literature magazine *The Strand*. One can also already identify a few outliers with a word count considerably higher than the majority of the samples. Unsurprisingly, these outliers regarding word count are the four famous short novels, which were published respectively as serialised novels across several issues of *The Strand*. This means in turn that these instances have to be excluded from the rest of the dataset within the further analysis steps. From a quantitative standpoint this grants one more assurance that the feature engineering computations and especially their subsequent scalings do not get distorted by a few select instances. Secondly, from the retrieved domain knowledge one can also assume that the novels might display a more complex and elaborate structure compared to their short stories counterparts when it comes to their respective plot elements – given that they were published not as a whole but in several segments and that they are often ranked amongst the most popular works of Doyle. Thus, it can be concluded for now that these outliers should be analysed independently

and in a bit more detail, which last but not least also does the initially laid out claim for a combination of both distant and close reading all the more justice.

title	word_counts
THE HOUND OF THE BASKERVILLES	59108
THE VALLEY OF FEAR	57551
A STUDY IN SCARLET	43379
THE SIGN OF THE FOUR	42998
THE NAVAL TREATY	12603

Figure 5: Outliers (word count)

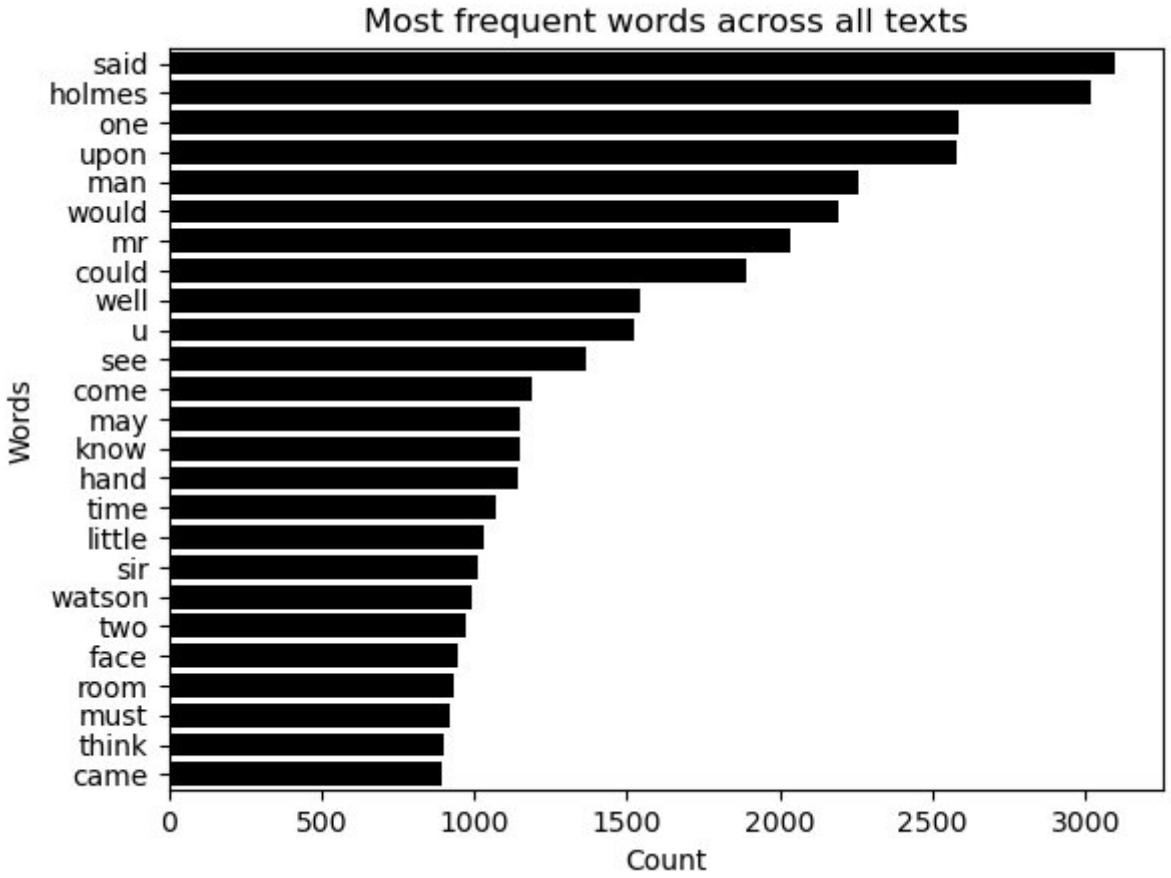

Figure 6: Most frequent words

To round the first exploratory probing of the dataset up, the most frequently occurring words across the whole corpus are also computed. Though the extraction of these word counts does not provide too much of a surprise at a first glance, it may still point towards another clue as to where further explorations in the following steps could lead: The frequent mentioning of the two protagonists names (Holmes and

Watson) are to be expected. Furthermore the frequent usage of basic verbs (such as 'know' or 'think') might already suggest the establishing of a certain narrative tension throughout a given story via the usage of action verbs that underline the intellectual processing of the detective and his sidekick. Last but not least the high occurrence of temporospatial nouns such as 'room' or 'time' point towards a reliance on topological and temporal relations to drive the narrative forward, which might especially be of importance when looking at NER and/or EE.

After the data has been wrangled accordingly and some initial NLP analysis on the texts has been accomplished, the next step (and probably also most crucial) is all about the computation of different, more complex features from the preprocessed data, which correspond directly to the main goal of measuring the narrative arc or trajectory of the adventures of Sherlock Holmes. In most cases this has been done by calculating certain frequencies (except for the task of emotion analysis) of words, sequences and other entities which are of interest for the purposes of this master's thesis' research focus and can be retrieved from a given instance. As previously mentioned, given that the EDA has shown that the lengths of the texts in the dataset differ at least in some cases, this fact has to be taken into account as well when computing the respective frequencies. Fortunately this problem can easily be solved with the help of TF-IDF – a common normalisation approach in NLP. Generally speaking, TF-IDF consists of three steps:

- The calculation of the term frequency (TF) which is defined by the frequency of a given term divided by the length of the text it occurs in.
- The calculation of the inverse document frequency (IDF) which is defined by the logarithm of the number of texts divided by the number of texts the given term occurs in.
- The multiplication of the two values which then results in the TF-IDF score.[107]

For this purpose the following function is defined which implements the TF-IDF formula on a given dataframe and then also normalises the results of TF-IDF to a value in the range of 0 and 1.

[107] Cf. also Aizawa, Akiko (2003): An information-theoretic perspective of tf-idf measures. In: Information Processing and Management 39, pp. 45–65.

```
def min_max_scale_column(df, new_column_name, col_to_norm, col_length):
    """this function first computes tf-idf on a dataframe and then scales
the retrieved values to a range between 0 and 1"""
    df['tf'] = df[col_to_norm] / len(df[col_length])
    df['idf'] = math.log(len(df.index)) / (df[col_to_norm] !=
0).value_counts()[True]
    df[new_column_name] = df['tf'] * df['idf']
    scaler = MinMaxScaler()
    df[new_column_name] = scaler.fit_transform(df[[new_column_name]])
    return df
```

Figure 7: Function for computing TF-IDF

The first task of feature engineering was inspired by Boyd et al. and their approach of defining and extracting three basic stages that constitute a given narrative text. Boyd et al. use a dictionary containing different keywords where each keyword is matched to one of the three stages to denote its respective contribution to the establishing of a certain plot element within the overarching trajectory of narrative coherence. Meaning that a given text segment, in which a higher occurrence of the keywords denoting the three stages is perceived displays also a higher value of narrative coherence, which in essence translates to a more rigid structure in its compositional form, while a text segement with fewer matches across the different groups of keywords in turn can be defined as 'less coherent' – or at least more 'liberal' in its sequential gestalt relative to the normative heuristics employed by Boyd et al. Applying this framework on the stories of Sherlock Holmes should then provide a first – albeit rough – probing of the general predictability that has often been associated with the recognisable patterns of Doyle's prominent works of detective fiction.

The aforementioned dictionary containing all the necessary keywords is publicly available and will therefore also be used for the purposes of this pipeline. As a result, three separate lists are created from the dictionary, each of them containing certain keywords that can then be matched against the segment of the texts of Sherlock Holmes within the dataframe. Each match for each of the three narrative stages is then summed up per segment and computed in a separate column. Note that the preprocessing step for this feature engineering task has resulted in the inclusion of stopwords, since the keyword lists consist in part also of these type of words, and the tokenisation has been applied on a word level, so that said lists can easily be matched with the text segments.

After the four outliers, which were previously defined, and the rest of the stories have been separated into two distinct dataframes, the numerical values for the feature of narrative coherence can finally be calculated. This means that the whole process involves firstly the computation of the raw frequency word counts, which are secondly normalised via the scaling function provided above (see figure 7). After this processing of the necessary values for the whole dataset has been done, it is possible to run some

basic statistical tests and also visualise their results. First only the mean and the standard deviation for each segment are calculated and secondly some ANOVA tests are run – a method which basically looks for significant deviations between the mean values of two or more categorical groups through the calculation of f-scores and p-values.[108]

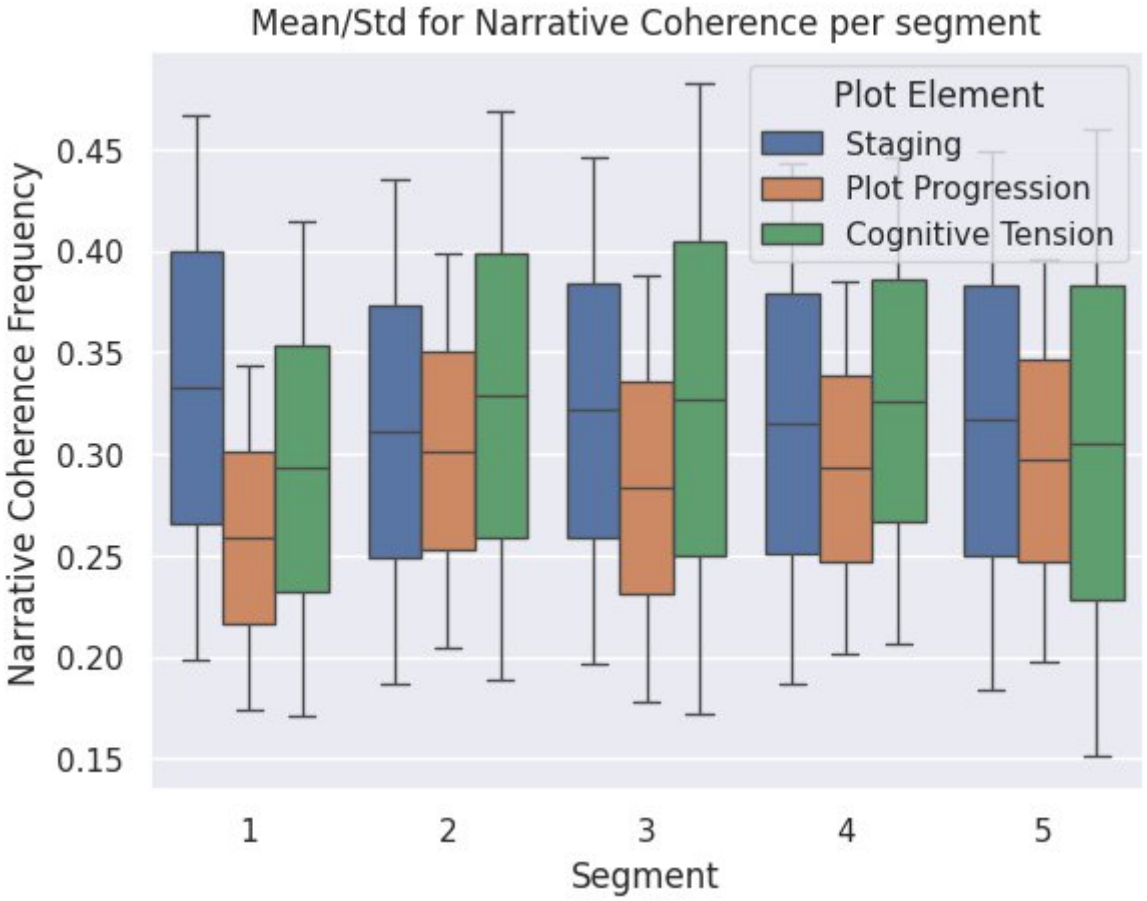

Figure 8: Mean/Std for narrative coherence across the different segments

Overall the plots as well as the ANOVA tests display a slight to moderate change of narrative coherence between the different segments. Nevertheless, the ANOVA metrics show that there is still a very high chance that the perceived stability of narrative coherence throughout the plots of Sherlock Holmes' stories are not due to random chance but very likely to also occur in similar populations. Therefore it can be concluded so far that the narrative coherence within the text corpus is mostly defined by a pattern of relative stability throughout the narrative progression. To explore this trace further and follow this notion it also makes sense to compare the different texts (through the combination of their respective scores for each of their sequences) with one another based on the features that were just created beforehand. In order to achieve this task some additional wrangling of the data is necessary: For this basically each individual text gets represented via a vector consisting of the different values for all of its five sequences

[108] Cf. also Kim, Tae Kyun (2017): Understanding one-way ANOVA using conceptual figures. In: Korean Journal of Anesthesiology 70, 1, pp. 22–26.

– i. e. every vector contains 5x3 values. By subsequently computing the cosine similarity of these vectors and finally also running a hierarchical clustering algorithm on them grants another, more elaborate opportunity to look whether or not there exists plausible evidence for the postulation of one or more archetypical narrative arc(s).

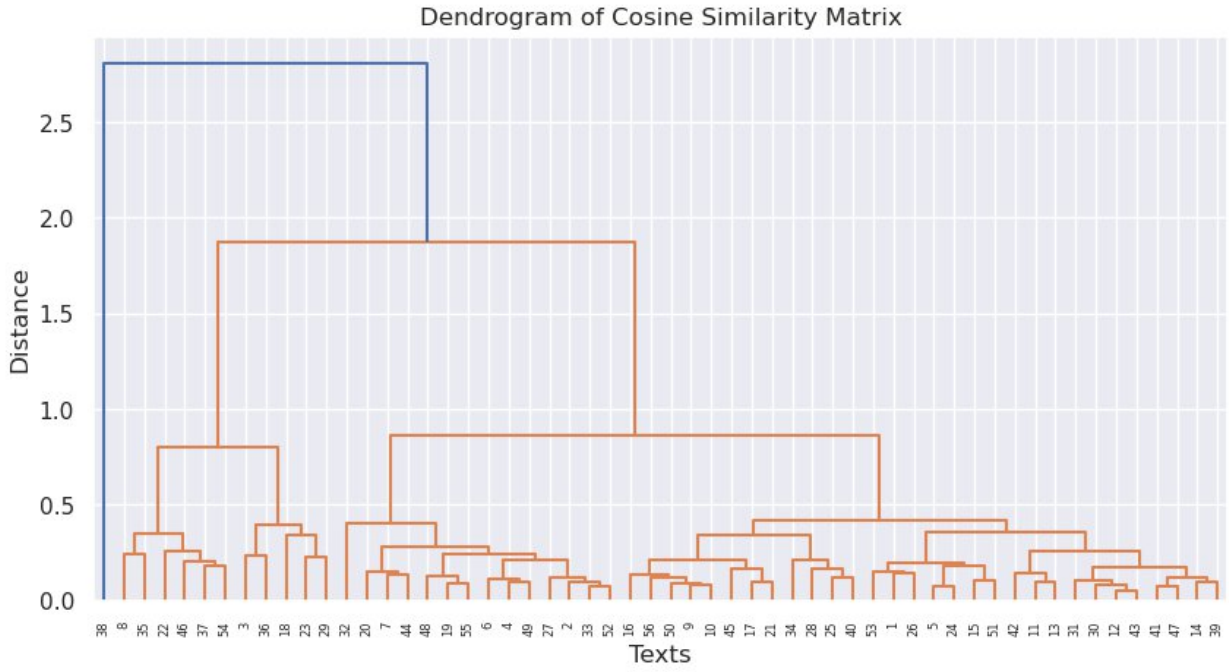

Figure 9: Clustering results for narrative coherence

While the results of the cosine similarity already suggest a high intrasimilarity across all the texts in the corpus – with only a few instances remaining which do not conform to the overall pattern – the resulting dendrogram from applying hierarchical clustering to the data emphasises the observation of a general narrative consistency throughout the different texts even more. Thus, all of the texts are put into one cluster (indicated by the orange colour), except for only one single text, which is put into its own cluster (blue), and can therefore be considered as an outlier.

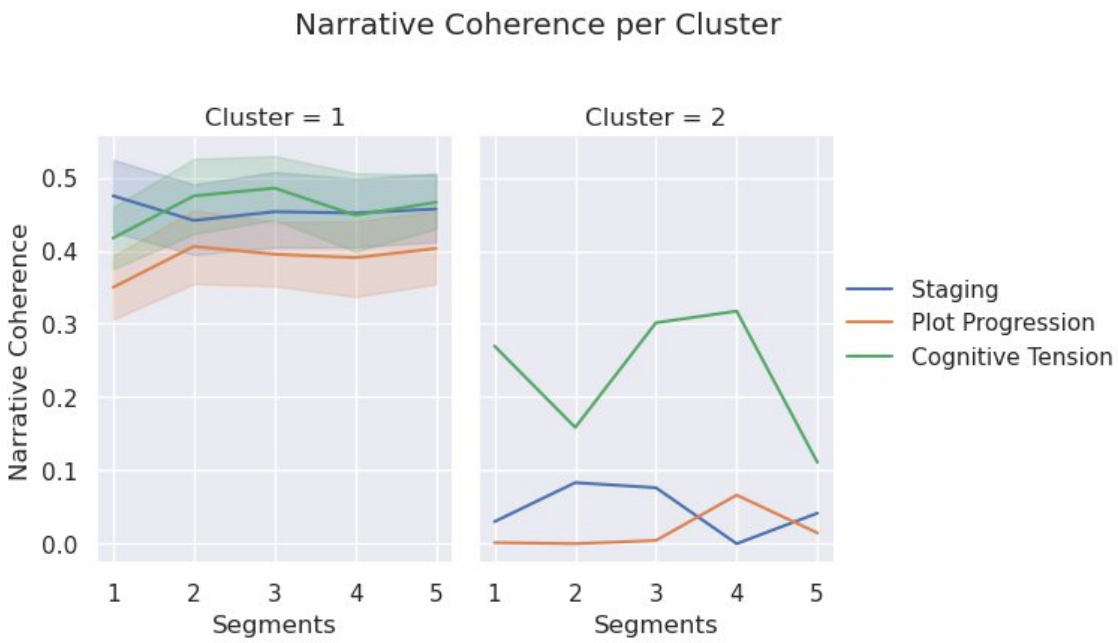

Figure 10: Development of narrative coherence per cluster

The visualisation of the narrative arc of the different clusters again confirms the already formulated notion that the different clusters do not only display a general stability in their respective narrative trajectory for the three different features of narrative coherence, but also show a general overarching similarity, pointing towards a genre-specific (or at least Sherlock Holmes specific) trend beyond the boundaries of individual texts. The results of the clustering algorithm reveal that the short story *The Adventure of the Veiled Lodger* differs considerably from the rest when it comes to narrative coherence. One should therefore note here that this text was published in 1927, which makes it one of the last publications within the corpus. Thus, this finding may already point towards a certain difference between Doyle's earlier and later writings. Of course this divergence may simply be anecdotal and therefore further investigations into this possible correlation are needed.

Last but not least the outliers are also examined closer by creating some separate plots of their narrative structure. From these plots one can already see at a first glance that the texts *The Hound of the Baskervilles* and *The Valley of Fear* display more similar patterns of narrative coherence, while on the other hand *A Study in Scarlet* and *The Sign of Four* constitute another couple of their own respectively. Looking in addition into their respective years of publication, it is also revealed that *The Hound of the Baskervilles* and *The Valley of Fear* were published during the second period of Doyle's writing (after 1893), while the other two outliers were published before that hiatus. Again it can be inferred from that finding that there might already emerge a certain trend regarding a distinguishable difference of writing in the stories of Sherlock Holmes as time progressed.

Another important feature for measuring narrative progression can be found in the usage of temporal expressions and/or relations. As already discussed, the traditional structuralist narratological research of detective fiction suggests that there often exists a certain achronological or analeptical structuring within the narration of crime stories such as the adventures of Sherlock Holmes where the exposition mainly revolves around the telling of the crime itself while the subsequent segments are then committed to the retelling or rather reconstruction of what had already happened, mainly signified by the investigations of the protagonist. This observation renders it all the more plausible to apply some POS tagging and then compute the frequencies of past, present and future tense usage across the different segments of a given text. To accomplish this task a function is defined which then returns the counts of past, present and future tense verbs based on their respective tags as three distinct columns.

```
def tense_counts(pos_tags):
    """this function takes as an input an array of pos tags and computes
the counts for past, present and future tense usage by grouping them into
three distinct categories according to their respective tag"""
    past_count = 0
    present_count = 0
    future_count = 0

    for word, tag in pos_tags[1::2]:
        if tag.startswith('VBD'):
            past_count += 1
        elif tag.startswith('VB'):
            present_count += 1
        elif tag.startswith('MD'):
            future_count += 1

    counts = {'past_count': past_count,
              'present_count': present_count,
              'future_count': future_count}

    return pd.Series(counts)
```

Figure 11: Function for retrieving frequencies of tense usage

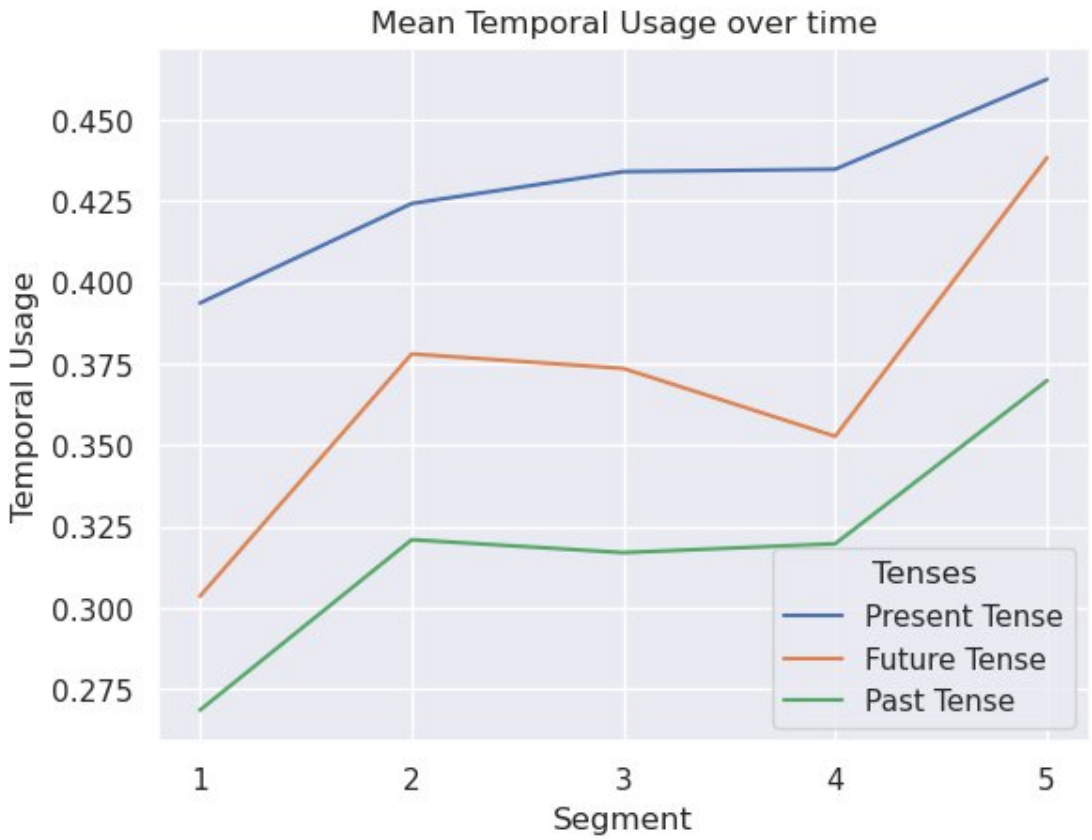

Figure 12: Average development of temporal usage over time

After the function has computed the relative frequencies over all the segments and the numerical results have again been normalised, the summary statistics display quite noteworthy results: For example one can clearly see that on average the usage of future and past tense verbs rises throughout the progression of a given narrative text. Along those lines the ANOVA tests point towards a considerable change in the usage of past and future tense across the different segments as well, as the p-values for both groups are clearly below 5%. In addition the clustering results reveal a similar consistency, distributing the different instances over two groups that both display increases in achronology as their plots move towards their conclusions. Thus, there seems to emerge a certain trend akin to the postulated prevalence of analeptical narration mentioned in the consulted literature which is indicated by a considerable increase of past (and future) tense usage as the plot progresses. From a quantitative point of view this increase can also be declared as statistically relevant enough to warrant a confirmation of this argument, and allows for the rejection of the nullhypothesis. This observed correspondence between the theories of traditional literary scholarship and these present findings can first and foremost then be seen as a legitimation of the here laid out modelling approach of distant reading and may furthermore also function as a more sound and systematic backing of former research and its rather anecdotal analysis of single exemplary texts.

As a third part of the feature engineering pipeline the research turns to emotion analysis. As previously pointed out, the current research regarding the fields of distant reading or plot analysis defines the

studying of emotional change throughout a given text's progression as one of its key components within existing methodological frameworks. This approach is furthermore also backed up by more recent findings within psychological or cognitive studies, which point towards a strong correlation between emotional intensity and the remembrance of pivotal moments in a given narration – be it for example a written text or also a verbal retelling of everyday experiences. In the specific case of the genre of detective fiction, one can certainly argue – since most detective stories (at least in the traditional sense) revolve around the procedural solving of a case and the general conflict of the good and the bad – that the alternations of positive emotions such as contentment, hope or relief on the one side and negative sentiments like grief, fear and anger on the other are also conducive to the convincing portrayal of a struggle between crime and justice. For computing the sentiment scores the pipelines makes use of the emotion analysis package *VADER* which is included in *NLTK* and works on a sentence level. Since *VADER* takes also additional information into account for its calculations such as punctuation, upper cases or ngrams, the preprocessing of the segments has been kept to a minimum for this task and mainly sticks to splitting them up into individual sentences.

```
def get_sentiment_scores(sentence_list):
    scores = []
    for sentence in sentence_list:
        score = vader.polarity_scores(sentence)
        scores.append(score['compound'])
    return scores
```

Figure 13: Function for retrieving emotional compound scores

After the sentiment scores for every sentence of every segment in the dataset have been retrieved the average compound score for each segment is extracted, which then amounts to the relative prevalence of positive or negative sentiment in a given segment. For this purpose a separate function is again defined, which firstly applies the *VADER* polarity method to each sentence and then extracts from the resulting array the compound score for each row in the dataframe. The compound score can take up a value between -1 and 1, where -1 denotes very negative sentiment and 1 very positive sentiment, a score between -0.05 and 0.05 is considered neutral.

Figure 14: Average emotional change over time

Generally speaking the average trajectory of the compound score reveals a quite clear difference between the segments – namely in the form of a perceivable movement towards a quite neutral territory of sentiment. Accordingly the ANOVA test reports a p-value < 5% which suggests that there is a statistically significant difference between the mean values of the individual segments. Thus, the evolution of sentiment seems – at least at a first glance – to be distributed quite unevenly throughout the narrative progression of the Sherlock Holmes stories.

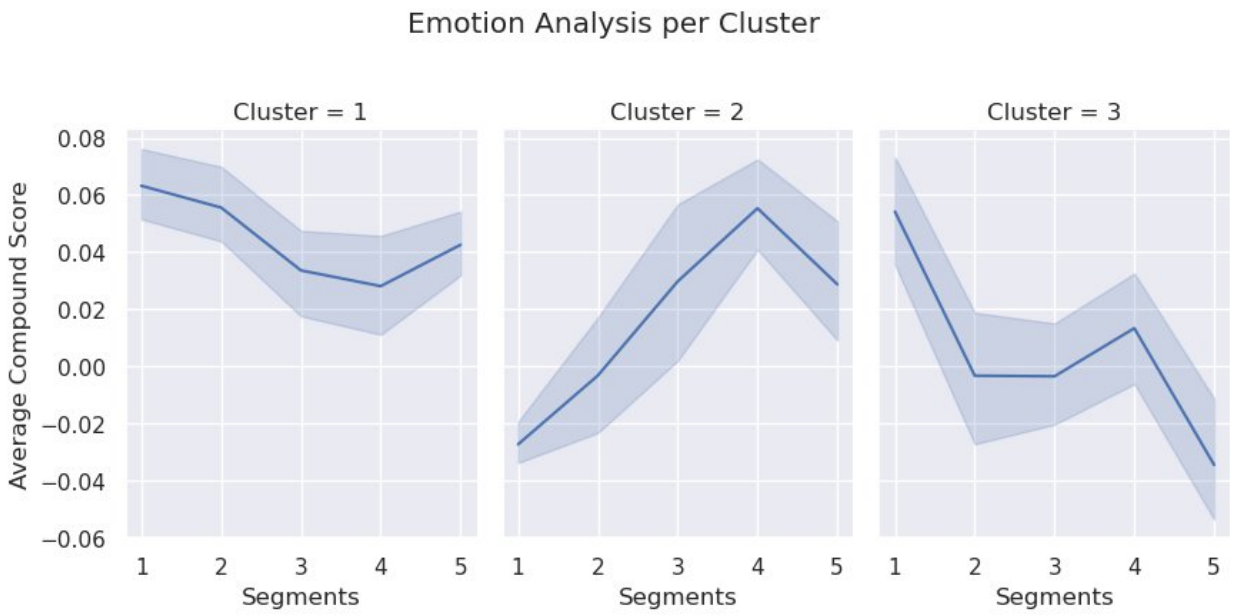

Figure 15: Clustering results for emotion analysis

The visualisation of the different clusters this time displays much more distinct plot variances – both between the clusters and its respective segments. What comes as a bit of a surprise though is the fact that not all stories seem to move towards a positive resolution with their conclusion. Since the general plot trajectory of the cases of Sherlock Holmes for the most part ends with the solving of the crime and the persecuting of the perpetrator, one would then also expect a more positive evolution of sentiment across all of the texts. On the other hand the plots also reveal that the general emotional trajectory across the three clusters seems to stay for the most part in a neutral territory which could for example be attributed to the more 'rationalistic' language and style of detective fiction, where the protagonist (here in the form of Holmes) tries to arrive at his conclusion mainly by analytical and logical reasoning. Speaking of the here perceived occurrence of *three* distinct clusters (compared to the prevalent number of *two* clusters for the other tasks) one can infer from this finding first and foremost a higher importance for the feature of emotional qualities. On the formal side this then leads to a more noticeable dispersion of emotional ranges, pointing towards constant attempts of engaging and surprising the reader anew, while also selling numerous iterations of publications, which have to combine both reoccurring structures (e. g. familiar characters, places and themes) and novel excitations, to a fairly large audience in the process. Such a reasoning finally appears especially convincing when one is reminded of the origins of detective fiction with its strong ties to suspenseful story telling and its reliance on invoking excitement throughout its trajectory. Plotting each story individually as a facet grid then even strengthens the notion of a notable variance of emotional plot developments across the whole corpus all the more.

Figure 16: Emotion analysis results per title

When it comes to the emotional qualities of the outliers one can furthermore see that three out of four stories (with the exception being *The Sign of the Four*) display a rather erratic dispersion of the emotional contents throughout and closing in towards a neutral territory. On the other hand *The Sign of the Four* ends on a more uplifting note, which could be explained by the fact that it also features a romantic subplot between Watson and his love interest Mary Morstan and in general revolves around the idea of redemption through justice and moral righteousness. On the other hand the rest of the stories focuses again for the most part on more radical sentiments such as mystery, suspense as well as revenge.

Last but not least the feature engineering part of the pipeline also looks into the distribution of persons and locations across the different segments of the texts. This approach of measuring plot has been mainly brought forward by Andrew Piper's research in the field of distant reading and is often referred to as

linguistic drift. The basic idea behind this concept is that for example the expansion of different characters and places, sceneries and similar loci throughout the progression of a given narrative text carries significant meaning for the constitution of the narrative arc. In order to extract this information from the texts it is necessary to first apply NER to the dataset and then count the frequencies of both recurring PERSON and LOCATION, GPE and FACILITY entities across the different sequences.

```
def extract_entities(text):
    pos_tags = nltk.pos_tag(text)
    tree = nltk.ne_chunk(pos_tags)
    return tree


def count_entities(parse_tree, *labels):
    """this function takes as input a parsed ner tree and one or more
labels and then computes the frequencies for the given label(s)"""
    tree = parse_tree
    count = 0
    for subtree in tree.subtrees():
        if subtree.label() in labels:
            count += 1
    return count


def most_common_entities(row, labels):
    """helper function for extracting the most common entities across all
texts"""
    tree = row
    entities = []
    for label in labels:
        entities.extend([subtree.leaves() for subtree in
tree.subtrees(lambda t: t.label() == label)])
    flattened_entities = [item for sublist in entities for item in
sublist]
    if flattened_entities:
        most_common_entity = max(set(flattened_entities),
key=flattened_entities.count)
        return most_common_entity
    else:
        return ''
```

Figure 17: Functions for extracting entities and computing their frequencies

For this purpose these three helper functions provided here are constructed, which first apply POS tagging and create a tree structure containing the recognised entities. Said tree can then be parsed via the two latter functions which compute the frequencies for given labels of interest and furthermore also extract the most common entries for a given label. Applying these three functions to the data then in turn

grants one not only the ability the look into the most prominently recurring characters and locations within the texts, but enables also further analysis in the realm of linguistic drift.

entities	
(Holmes, NNP)	137
(Watson, NNP)	21
(Lestrade, NNP)	5
(Godfrey, NNP)	4
(Brunton, NNP)	4
(Ferguson, NNP)	4
(Robert, NNP)	4
(Bork, NNP)	3
(Hudson, NNP)	3
(Barclay, NNP)	3

entities	
(London, NNP)	56
(Street, NNP)	14
(England, NNP)	8
(Holmes, NNP)	7
	7
(Indian, JJ)	6
(American, JJ)	5
(English, NNP)	5
(French, JJ)	5
(Greek, JJ)	4

Figure 18: Most common persons and locations

The most common characters and locations within the texts do not provide much of a surprise: With the two protagonists Holmes and Watson being by far the characters with the most mentions while the local entities mainly revolve around the confinements of England and London respectively.

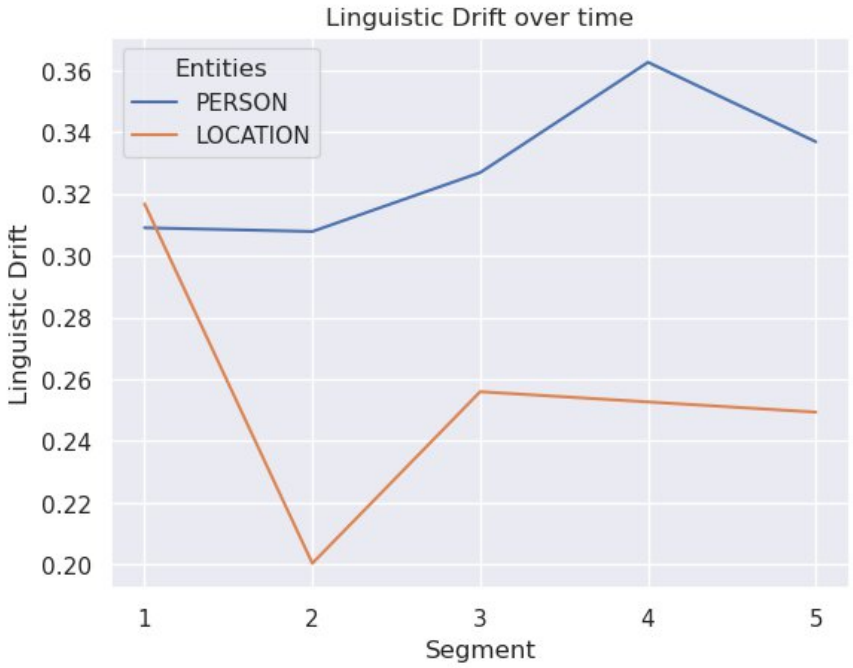

Figure 19: Average development of linguistic drift over time

Here the statistical tests produce mixed results for the postulation of clear differences between the segments regarding the feature of linguistic drift: Firstly, as one can see from the plots and initial

probing, there seems to exist on average a moderate increase in the occurrence of person mentions throughout the narrative progression, which could for example correspond to a wider network of characters as a given detective case with all its victims and perpetrators unfolds. However this pattern is still not quite statistically significant enough according to the ANOVA tests. On the other hand the mentions of specific locations appears to be at its highest right at the beginning of the narrative, then takes a sudden dip within the second segment and afterwards stays quite stable over the remaining segments. This could be directly related to the fact that the first segment is mainly responsible for the exposition and the introduction of the given setting. and also that most of the cases of Sherlock Holmes take place in the vicinity of London, thereby not really evoking a considerable expansion of different scenery as the narrative progresses. This pattern seems also significant enough, given that it produces a p-value of 0.014. Similarly the clustering of the NER results again separates the instances into two groups. Both clusters also display a certain expansion of characters while the frequency of locations remains for the most part stable.

The arcs of the outliers furthermore again paint a similar picture already discovered earlier and therefore also relate to the domain knowledge: While the settings of *The Hound of the Baskervilles* and *The Valley of Fear* both take place in rural or isolated locations (i. e. in Dartmoor and in a mining community in the English countryside), the settings of the other two stories include London as a primary location. The latter two stories, which were published earlier than their counterparts, furthermore serve as introductions to the characters of Sherlock Holmes and Dr. Watson, while the former pair builds upon an already existing character relationship and expands the general character network.

This concludes the first section of the feature engineering part of the pipeline revolving around the examinations of the plot or formal attributes. Therefore the remaining analysis of relevant features will now instead focus on the actual content of the stories – or to be more concrete on the topical aspects that are recurring throughout the narratives of Sherlock Holmes. For this task two common approaches of topic extraction are used which are event extraction (EE) and latent semantic indexing (LSI).

Beginning with EE, the decision to employ this kind of method here is twofold: First of all the relevant literature – as it has been discussed in more detail in the last chapter – suggests a sensible connection between EE and the computerised analysis of detective fiction which has so far been mainly employed for building models for automatically detecting the culprit in a given crime story. Secondly, the NLP library *spaCy* already provides an easy to use pipeline for this endeavour out of the box which we will also use for our purposes in the function provided below.

```
def extract_events(text):
    """this function applies the basic spacy pipeline for event extraction
to the textual input and then extracts certain entities based on the
ruleset provided afterwards"""

    nlp = spacy.load('en_core_web_md')

    processed_text = ' '.join(text)
    doc = nlp(processed_text)

    events = []

    for ent in doc.ents:
        if ent.label_ == 'EVENT':
            event = {
                'event': ent.text,
                'start': ent.start_char,
                'end': ent.end_char,
                'context': [t.text for t in ent.sent],
            }
            events.append(event)

    for sent in doc.sents:
        persons = [ent for ent in sent.ents if ent.label_ == 'PERSON']
        locations = [ent for ent in sent.ents if ent.label_ in ['LOC',
'GPE', 'FAC', 'TIME']]

        for person in persons:
            for location in locations:
                event = {
                    'event': person.text + ' ' + location.text,
                    'start': min(person.start_char, location.start_char),
                    'end': max(person.end_char, location.end_char),
                    'context': [t.text for t in sent],
                }
                events.append(event)

    return events
```

Figure 20: Function for event extraction

The function *extract_events* therefore applies under the hood several common NLP steps to its textual input such as POS tagging, sentence parsing and NER in order to label specific words or longer phrases and measure their respective part in contributing to the content of a text. After the pipeline has been applied one can also add certain rules to the function in order to tell it for example which particular entities it should return. Note here that *spaCy* (unlike *NLTK*) even provides the entity type EVENT for a straightforward method of EE. Furthermore the function is also instructed to provide the co-occurring persons and temporospatial expressions in order to build upon the aforementioned NER task regarding

the notion of linguistic drift and also incorporate the initial findings of the EDA, especially regarding the frequently occurring words and ngrams. Of course this function could still be refined and expanded upon via the introduction of more detailed rules, but since the extraction of topics is not the main part of the analysis, a more basic ruleset has been chosen in order to focus more on the general trends which are latent in the data.

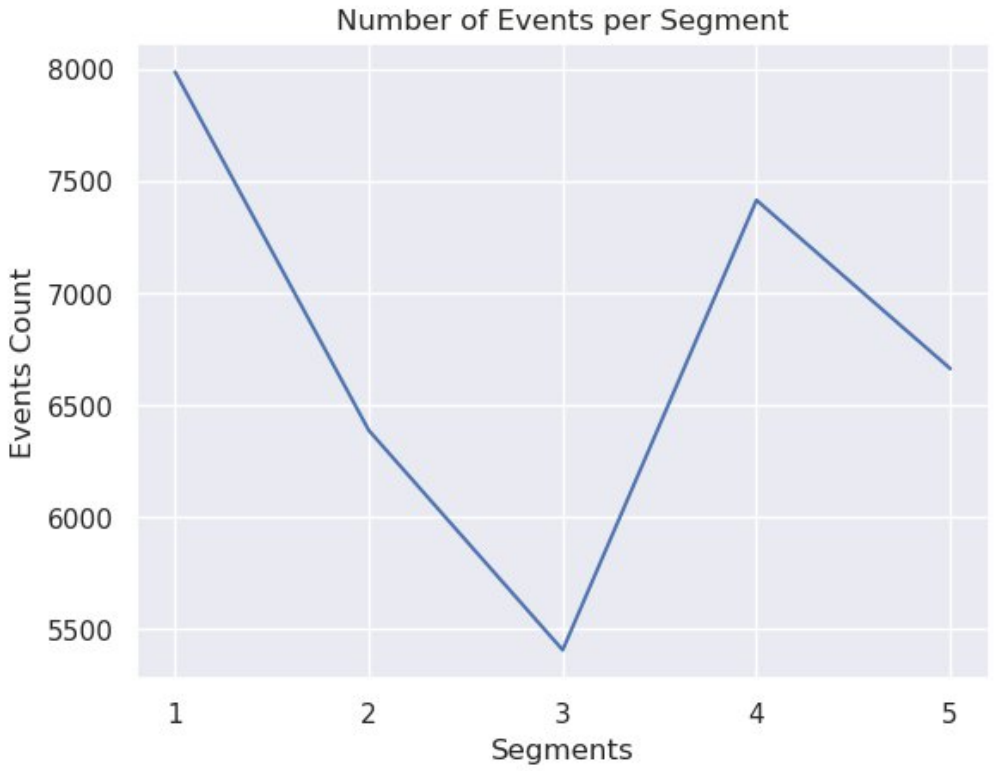

Figure 21: Development of events over time

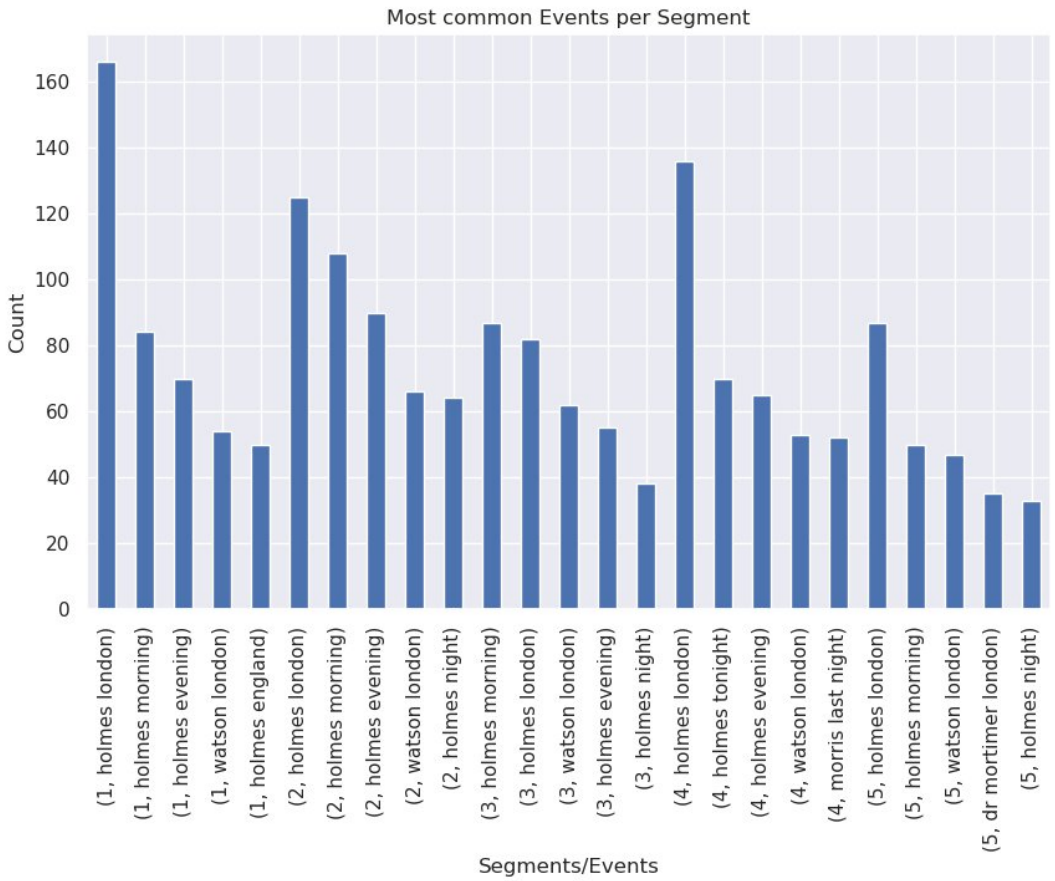

Figure 22: Most common events for the different segments

Visualising the results for EE one can first and foremost see a correspondence between the first and the fourth segment and the second, the third and the last segment respectively. This can be for example explained by the initial introduction of certain characters and locations which then is picked up again towards the end of the narrative, where the final conflict and resolution of the crime is imminent. The three remaining segments on the other hand mostly seem to expand upon the already existing arc established before and after them. Looking at the most common events per segment it does not really come to a surprise that the progression of the narrative is mainly dominated by the character movement of Holmes and Watson in conjunction with the area of London and/or England. Last but not least one can also see a certain prevalence of night time settings (represented by the entities ‘evening’ and ‘night’) which most likely plays a crucial role in expressing the more eery and mysterious qualities of the detective stories.

A second common method for topic extraction, which will also be employed here, can be found in LSI. The different steps of computing LSI are as follows:

- Document-Term Matrix: LSI starts by constructing a matrix where each row represents a document, and each column represents a term in the entire collection of documents. The entries in the matrix typically represent the frequency or importance of each term in each document.
- Singular Value Decomposition (SVD): SVD is applied to the document-term matrix to decompose it into three separate matrices: U, S, and V. U represents the relationship between documents and latent concepts, S captures the importance of each latent concept, and V represents the relationship between terms and latent concepts. The SVD decomposition allows for dimensionality reduction by truncating the matrices. By keeping only the most important singular values and associated columns from U, S, and V, we obtain a lower-dimensional representation of the original data.
- Semantic Space: The reduced matrices (U', S', V') represent a transformed semantic space, where the documents and terms are related based on their underlying semantic structure. Each document and term is now represented as a vector in this semantic space.
- Query and Similarity: To find relevant documents given a query, the query is transformed into the same semantic space using the matrix transformations obtained from SVD. Similarity between documents and the query can then be measured using techniques like cosine similarity, where closer vectors indicate higher similarity.[109]

For the purposes of this pipeline a function that computes LSI on our data and extracts for each segment the top 5 topics is defined. In addition, after extracting the most important topics for each segment, the pipeline also computes the relative change of topics over narrative time. For this task we first vectorise our different topics and then apply cosine similarity to them to retrieve their (dis-)similarity scores.

```
def compute_lsi_on_subsets(df, text_column, category_column):
    vectorizer = TfidfVectorizer()
    lsa = TruncatedSVD(n_components=30, random_state=19)
    results = {}

    unique_categories = df[category_column].unique()

    for category in unique_categories:
        subset_df = df[df[category_column] == category]
        dtm = vectorizer.fit_transform(subset_df[text_column].apply(lambda
x: ' '.join(x)))
        lsa.fit(dtm)
        lsa_vectors = lsa.transform(dtm)
        feature_names = vectorizer.get_feature_names_out()
        topics = []
```

[109] Cf. also Deerwester, Scott et al. (1990):

```
        for topic_idx, topic in enumerate(lsa.components_):
            top_features = [feature_names[i] for i in topic.argsort()[:-5
- 1:-1]]
            topics.append(top_features)

        results[category] = topics

    return results

def vectorize_lsi_results(lsi_results):
    vectorizer = TfidfVectorizer()
    category_vectors = {}

    for category, topics in lsi_results.items():
        features = [' '.join(topic) for topic in topics]
        category_vectors[category] = vectorizer.fit_transform(features)

    return category_vectors

def compute_difference(category_vectors):
    categories = list(category_vectors.keys())
    num_categories = len(categories)
    differences = np.zeros((num_categories, num_categories))

    svd = TruncatedSVD(n_components=min([vector.shape[1] for vector in
category_vectors.values()]))
    reduced_vectors = {category: svd.fit_transform(vector) for category,
vector in category_vectors.items()}

    for i in range(num_categories):
        vector_i = reduced_vectors[categories[i]]
        for j in range(i+1, num_categories):
            vector_j = reduced_vectors[categories[j]]
            similarity = cosine_similarity(vector_i, vector_j)[0][0]
            difference = 1 - similarity
            differences[i, j] = difference
            differences[j, i] = difference

    return differences
```

Figure 23: Functions for computing LSI and retrieving the most important topics via cosine similarity

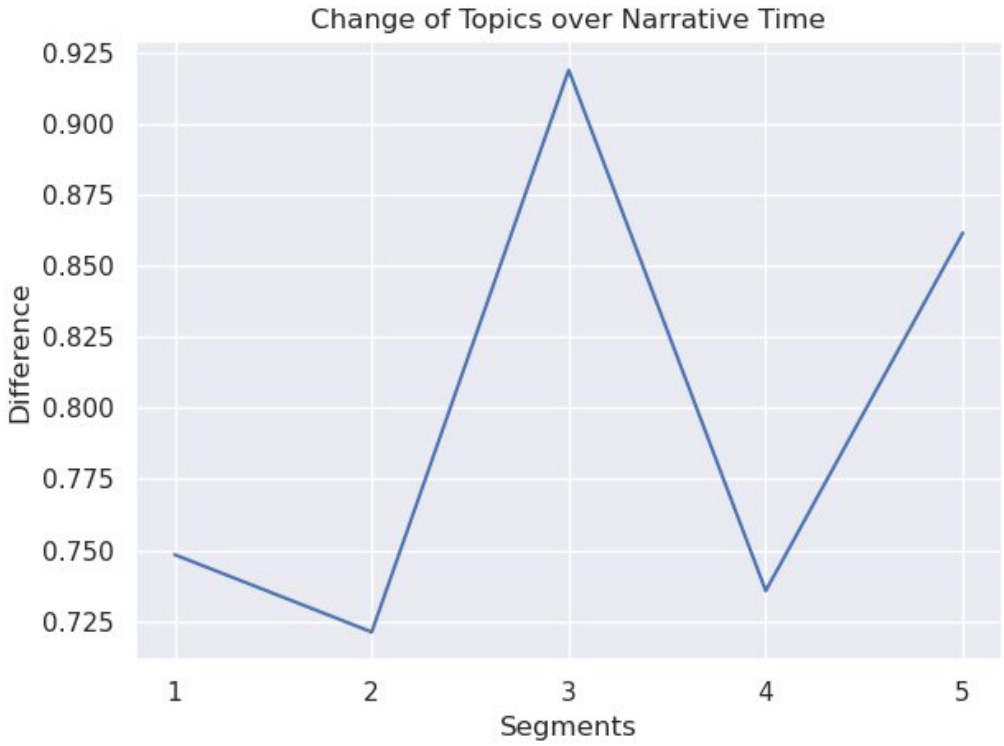

Figure 24: Development of topics over time

From plotting the average change of topics over time a pretty similar graph to the one produced above for the task of EE – at least when it comes to the rather erratic alternations of frequency patterns – is finally retrieved (see figure 21). In general the trajectory shows a higher similarity between the first, the second and the fourth, and the third and the last segment respectively. This observation can for example be explained by the fact that the confrontation of the crime as well as its solution appear in most cases right in the middle and ending of the story, thereby creating a certain (topical) link between these parts of the narrative while the remaining parts then build a kind of 'bridge' between these two main segments, specifying and expanding upon certain characters and settings without adding that much topological weight on their own.

After all the necessary features have successfully been computed, the final step of combining them and exploring their respective contributions to the overarching narrative arc as well as their interactions between each other still remains to be implemented. For this all the features are first and foremost put together into a single dataframe, which makes it then possible to apply some more complex methods of statistical modelling in order to further build on the notions and the results already derived from the more basic statistical evaluations discussed above. Since the initial EDA showed that the data – at least when it comes to the publication count – can roughly be split into two distinct time periods (the first one before the 'death' of Holmes and the second one after his 'resurrection') – an additional binary column is introduced into the dataframe which discriminates the instances according to these two periods.

The dendrograms of the initial clusterings attempts to the respective features for the most part already suggested the existence of two distinct clusters. Furthermore the hypothesis that there might exist a distinction between the two periods of Doyle's writing and the introduction of the additional period variable also warrants the attempt of trying to separate the instances with all the relevant features combined into two clusters. This notion can now be tested with the method of k-means clustering in order to see whether or not a value of 2 for the parameter of the number of clusters produces results that verify the previous claim. Since at this point it is not of any interest to examine each segment separately, the data is instead grouped by the title and then only the different texts are clustered. The silhouette score and the Dunn index are two common measures for testing the results of k-means or rather the inter- and intra-stability of the separating qualities of the clusters thereby produced. The silhouette score measures how well each sample in a cluster fits with the other samples in the same cluster compared to samples in other clusters. It ranges from -1 to 1, where a score close to 1 indicates that the samples are well-clustered, a score close to 0 indicates overlapping or ambiguous clusters and a score close to -1 indicates that samples may have been assigned to the wrong clusters. The Dunn index measures the compactness of clusters and the separation between different clusters. It aims to find clusters that are tight and well-separated. The Dunn index is calculated as the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. A higher Dunn index value indicates better clustering quality.[110] Computing these two values on the clustering results produces values of 0.6 and 2 respectively, which already suggest a quite good separation of the instances.

[110] Cf. also Gareth, James et al. (2021): An Introduction to Statistical Learning. With Applications in R, 2nd Edition. Berlin: Springer, pp. 517–520.

	title	cluster
0	HIS LAST BOW	0
1	THE ADVENTURE OF BLACK PETER	0
2	THE ADVENTURE OF CHARLES AUGUSTUS MILVERTON	0
3	THE ADVENTURE OF SHOSCOMBE OLD PLACE	0
4	THE ADVENTURE OF THE ABBEY GRANGE	0
5	THE ADVENTURE OF THE BRUCE-PARTINGTON PLANS	0
6	THE ADVENTURE OF THE CREEPING MAN	0
7	THE ADVENTURE OF THE DANCING MEN	0
8	THE ADVENTURE OF THE DEVIL'S FOOT	0
9	THE ADVENTURE OF THE DYING DETECTIVE	0
10	THE ADVENTURE OF THE EMPTY HOUSE	0
11	THE ADVENTURE OF THE GOLDEN PINCE-NEZ	0
12	THE ADVENTURE OF THE LION'S MANE	0
13	THE ADVENTURE OF THE MAZARIN STONE	0
14	THE ADVENTURE OF THE MISSING THREE-QUARTER	0
15	THE ADVENTURE OF THE NORWOOD BUILDER	0
16	THE ADVENTURE OF THE PRIORY SCHOOL	0
17	THE ADVENTURE OF THE RED CIRCLE	0
18	THE ADVENTURE OF THE RETIRED COLOURMAN	0
19	THE ADVENTURE OF THE SECOND STAIN	0
20	THE ADVENTURE OF THE SIX NAPOLEONS	0
21	THE ADVENTURE OF THE SOLITARY CYCLIST	0
22	THE ADVENTURE OF THE SUSSEX VAMPIRE	0
23	THE ADVENTURE OF THE THREE GABLES	0
24	THE ADVENTURE OF THE THREE GARRIDEBS	0
25	THE ADVENTURE OF THE THREE STUDENTS	0
26	THE ADVENTURE OF THE VEILED LODGER	0
27	THE BLANCHED SOLDIER	0
28	THE DISAPPEARANCE OF LADY FRANCES CARFAX	0
29	THE ILLUSTRIOUS CLIENT	0
30	THE PROBLEM OF THOR BRIDGE	0

	title	cluster
0	A CASE OF IDENTITY	1
1	A SCANDAL IN BOHEMIA	1
2	SILVER BLAZE	1
3	THE "GLORIA SCOTT"	1
4	THE ADVENTURE OF THE BERYL CORONET	1
5	THE ADVENTURE OF THE BLUE CARBUNCLE	1
6	THE ADVENTURE OF THE CARDBOARD BOX	1
7	THE ADVENTURE OF THE COPPER BEECHES	1
8	THE ADVENTURE OF THE ENGINEER'S THUMB	1
9	THE ADVENTURE OF THE NOBLE BACHELOR	1
10	THE ADVENTURE OF THE SPECKLED BAND	1
11	THE ADVENTURE OF WISTERIA LODGE	1
12	THE BOSCOMBE VALLEY MYSTERY	1
13	THE CROOKED MAN	1
14	THE FINAL PROBLEM	1
15	THE FIVE ORANGE PIPS	1
16	THE GREEK INTERPRETER	1
17	THE MAN WITH THE TWISTED LIP	1
18	THE MUSGRAVE RITUAL	1
19	THE NAVAL TREATY	1
20	THE RED-HEADED LEAGUE	1
21	THE REIGATE SQUIRES	1
22	THE RESIDENT PATIENT	1
23	THE STOCK-BROKER'S CLERK	1
24	THE YELLOW FACE	1

Figure 25: k-means clustering results with k=2

By printing all the of instances per cluster, one can also see that the sizes of the two clusters are quite even, which suggests a good distribution and therefore warrants it to explore the notion of two distinct phases in the writing of the stories of Sherlock Holmes further.

As a next step principal component analysis (PCA), a common method of dimensionality reduction, is applied to the features in order to better visualise the instances in relation to their attributes.[111] Again the segments are grouped by their corresponding titles in order to compare the different texts to one another. First the features are reduced to two dimensions and PCA is computed twice – once without the period variable and once with it being included.

[111] Cf. also Gareth, James et al. (2021): pp. 498–510.

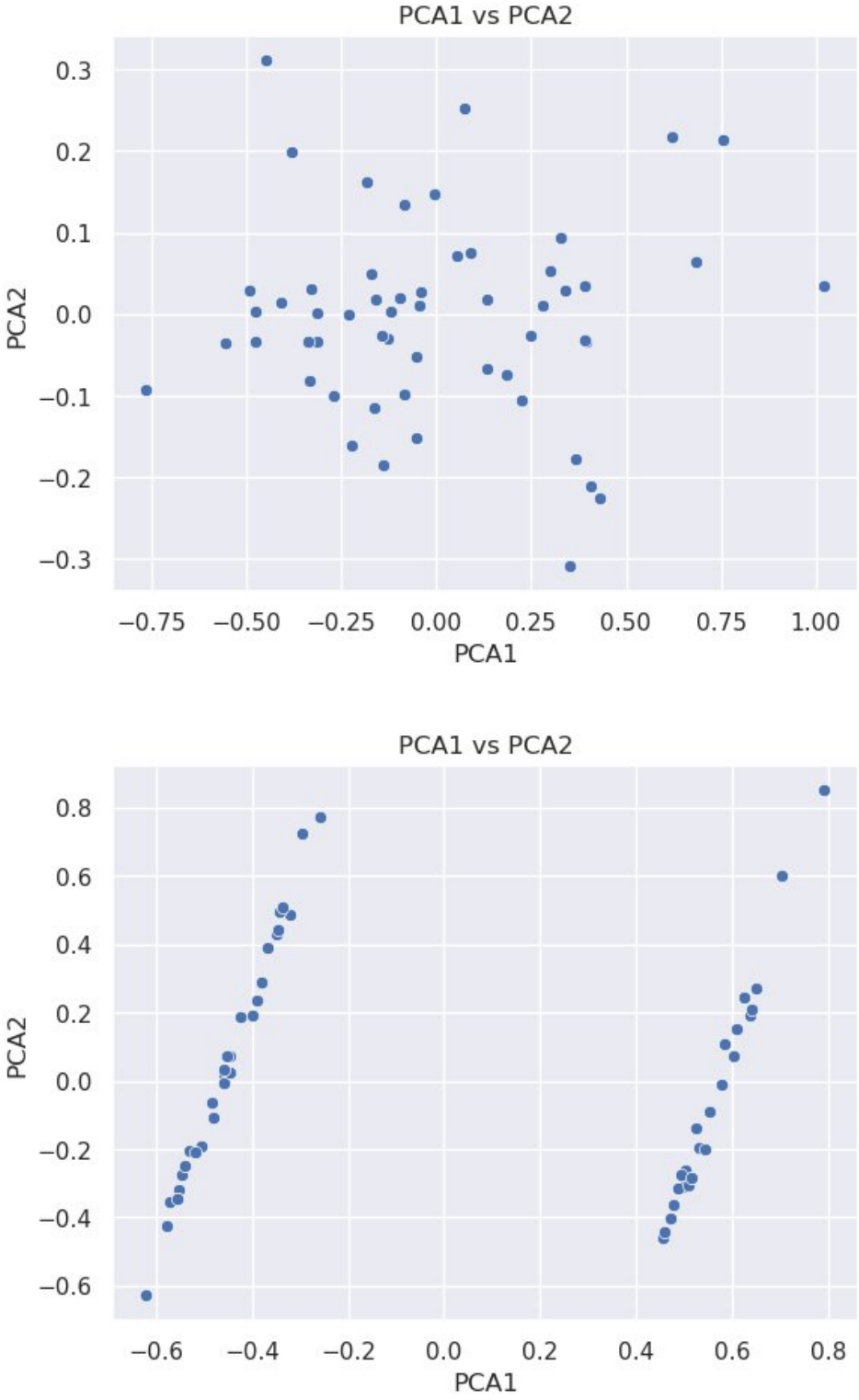

Figure 26: PCA with the period variable excluded and included

Investigating the two thereby produced scatterplots reveals that the second one, which includes the period-feature, displays a clear segmentation of the data into two distinct groups, while the first PCA attempt on the other hand spreads out the different instances much more evenly. This result furthermore corresponds to the initial clustering results for k-means, which also left out the period-column at first

and got much more overlap between the two clusters. On the other hand computing only one value for PCA and plotting it against the period variable also reveals a clear dispersion of the instances along the y-axis between the two groups.

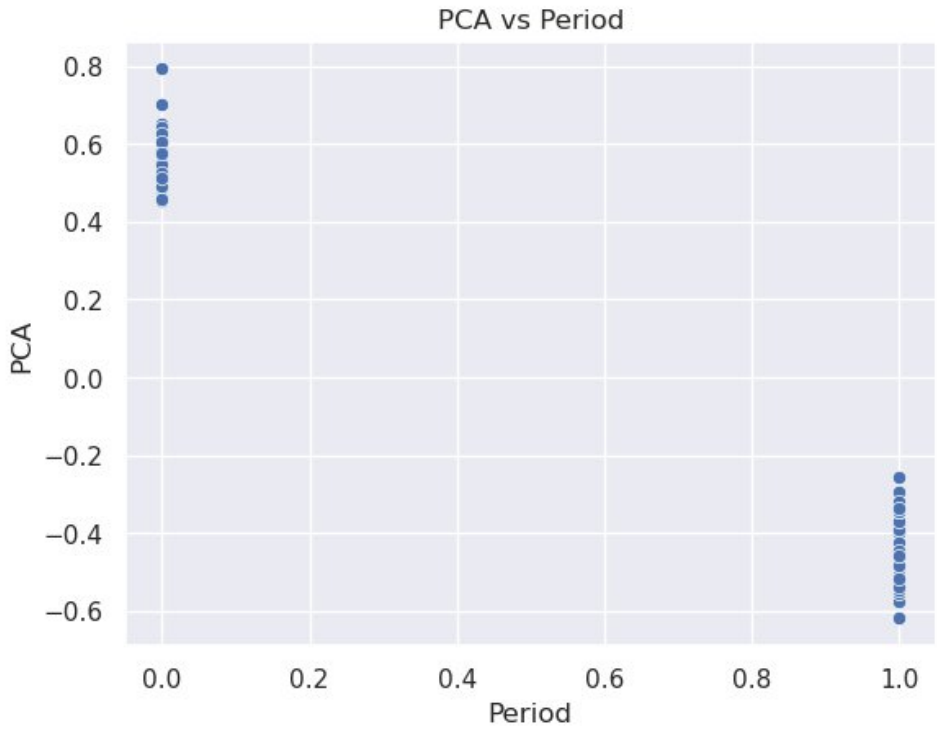

Figure 27: PCA vs period variable

After the separation of the instances into subgroups it furthermore makes sense to produce some basic classification models to explore the discerning qualities of the features even further. According to Ted Underwood the probabilistic nature of statistical models such as logistic regression allows for a quantitative proof of certain notions and beliefs by assigning a granular likelihood score to different instances and a corresponding label.[112] Thus, in the case of the task of finding distinguishing attributes between the different segments of the texts, the pipeline relies on training a multiclass logistic regression model, which tries to assign a given sequence of text to one of the five segments with a certain likelihood based on the features computed beforehand. The following code then includes all the features which had displayed a significant p-value in the earlier ANOVA-tests (except for the count of locations since that feature proved to reduce the model's performance when included initially) and uses the assignment column of the different segment numbers for its classification labels. It then splits the data into training and testing increments and finally prints out the common evaluation metrics of the resulting model.

[112] Cf. also Gareth, James et al. (2021): pp. 133–140.

	precision	recall	f1-score	support
1	0.32	0.50	0.39	12
2	0.00	0.00	0.00	18
3	0.33	0.11	0.17	9
4	0.11	0.29	0.16	7
5	0.44	0.70	0.54	10
accuracy			0.29	56
macro avg	0.24	0.32	0.25	56
weighted avg	0.21	0.29	0.23	56

Figure 28: Results for logistic regression

While the model in general does not display a very high accuracy, which can first and foremost be explained by the rather small number of instances, one can nevertheless clearly see that it performs comparatively well when classifying the first and the last segment of a given text compared to the other three segments. This finding suggests that there exists the most distinction between these two classes which for example then also corresponds to the notion of the basic narrative structure discussed before with the exposition at the end and the conclusion in its final stage. In the specific case of the detective story one might then also argue that the classic structure of the initial introduction of the crime and its subsequent solving in the end is reflected in these probabilistic scores. The fact that the model is on the other hand unable to properly classify instances of the second segment shows us that this part of a given narrative might contain the least importance for the overall progression.

In addition two support vector machines (SVM) are deployed – one for the same classification task described above for the logistic regression, and one for the binary discrimination of the two periods. In general SVM is more suited towards handling data with high dimensionality and finding underlying correlations accordingly, which provides an additional double-check against the more straightforward logistic regression approach. Secondly, it tries to find the best decision boundary between its instances, which appears particularly useful when it comes to the binary classification task of separating the two periods.[113]

[113] Cf. also Gareth, James et al. (2021): pp. 367–387.

	precision	recall	f1-score	support
0	0.62	0.75	0.68	20
1	0.84	0.75	0.79	36
accuracy			0.75	56
macro avg	0.73	0.75	0.74	56
weighted avg	0.77	0.75	0.75	56

Figure 29: Results for support vector machine

Again the SVM for discerning the different segments produces similar results as the logistic regression model and also performs (relatively speaking) the best when labelling the first and the last segment. Also note that the second segment again remains unclassified. Overall the accuracy of the SVM remains quite poor but in comparison to the logistic regression it performs a bit better. Turning to the binary classification task of assigning a given segment to one of the two possible periods, the SVM model achieves a quite high accuracy of 75%, which again adds even more weight to the hypothesis of the discerning qualities of this particular feature.

Last but not least the pipeline turns to the visualisation of the general narrative arc which appears to be latent in the data and is based on the sum of all the individual features. For this once again PCA is utilised and then the average narrative arc of both clusters and per period is retrieved.

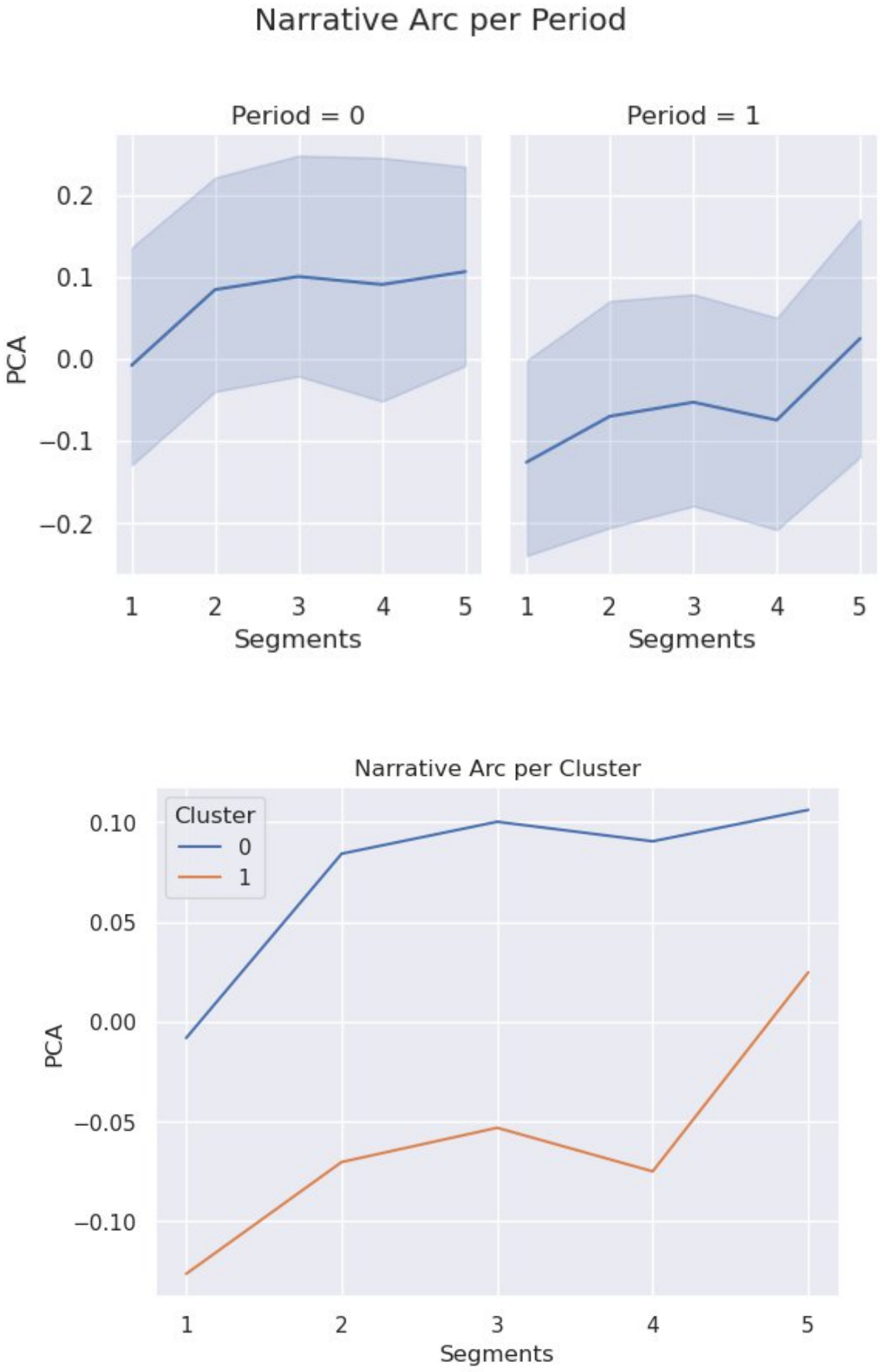

Figure 30: The narrative arc of *Sherlock Holmes*

The hereby produced visualisations finally show that the results of the k-means clustering and the grouping of the data per period both produce almost identical results regarding the overall narrative arc, thereby providing an even stronger proof for the existence of a changing of the structures of the stories as the publication cycle of *Sherlock Holmes* approached its end. While the relative, overall trajectory of the different texts does still appear quite similar, one can nevertheless clearly see that the first period of Doyle's writing on average displays a higher amount of narrative density (i. e. a higher occurrence of the relevant features) across the different segments than the second period following Holmes' 'death'. One common and obvious explanation for this pattern might then be found in the creator's growing contempt and general dissatisfaction with its most famous protagonist when compared to his other literary endeavours such as the writing of historical novels, which where always overshadowed by the adventures of the beloved detective but to which the author also devoted a considerable amount of work and passion during his lifetime. This observation finally also echoes the initial theories of the Russian formalists and here especially Tynianov's concepts on the general evolutionary trends (and consequent 'decay') of literary genres, thereby coming in the end full circle and closing a decisive gap between the traditional methodological roots of distant reading and its recent practical advancements.

To summarise, the here developed pipeline can be seen as an effort to provide meaningful NLP techniques for the analysis and evaluation of recurring plot structures within literary texts. Not only do the different tasks of feature engineering try to capture the basic formal qualities of genre texts (the syuzhet), they also go on to look into the topical contents of its specimens by examining recurring events and topics (the fabula) across narrative time. While the code (and especially its respective parameters) have been tuned to the specific domain of detective fiction and the adventures of Sherlock Holmes, it seems plausible to argue that its usage may also be applicable to other corpora of texts and/or genres – requiring mostly (if any) minor tweaks provided by the literary scholar and his specific domain knowledge. Of course further research with larger quantities of (different) text data is still necessary to verify or falsify this claim in full. So far, it can be concluded that the present examination of the stories of Sherlock Holmes through the lens of distant reading has produced some notable and statistically significant results. Especially the engineering and testing of features such as the temporal usage and the emotional progression displayed a considerable contribution to the general narrative arc of detective fiction. In addition, the hypothesis that there exists a certain diversion – which could mainly be proven by grouping the data into two distinct periods – in the earlier and later writing of Doyle's detective stories appears not only as a novel insight but can also be backed up when looking at the general publication history of the adventures of Sherlock Holmes with its rising popularity on the one hand and the biographically proven fading interest of its creator on the other.

8. Conclusion

The present master's thesis not only aimed to recount the theoretical foundations of distant reading and computational literary studies by tracing back the beginnings of the statistical and formalistic analysis of literary texts to the mathematically inclined schools of Russian formalism and structuralism, but also set out to reconcile these more abstract concepts with programmatical state of the art NLP approaches as well as the surrounding research laid out by the likes of Moretti, Underwood or Piper. In this light this work owes much to a (surprisingly) rich legacy dating back to the early 20th century, which provided a considerable legitimisation for the (still for the most part neglected) pairing of quantitative evaluations and the study of artistic language usage. Thus, one of the first goals of this thesis lay precisely in arguing against often encountered reservations against a more generalist and corpus-driven study of literature, which falls especially under scrutiny when it comes to its (alleged) opposition to the traditional hermeneutic or close readings of a few preselected individual texts. That such an opposition does not necessarily have to seen as antagonistic as it has often been understood by the critics of distant reading and like-minded, more orthodox scholars, could be shown via for example the practice of also taking a closer look at the outliers, which had been identified as such beforehand through statistical means, in a separate part of the carried out analysis.

The main principles of the methodological framework employed here, which were described in detail in the first few chapters, revolved therefore in essence around the promoting of the structural and formal study of literary language. This should first and foremost be achieved through the identification and subsequent extraction of relevant features and similar segments that can then be declared as conducive to the overall meaning as well as form of a given text. This specific analytical approach – the *operationalisation* of written language – relies on the basic formalistic axiom that aesthetic expression is best be captured when considered as a sum of smaller atomic parts that can be examined separately and/or in relation to one another. Combining said axiom furthermore with the more technical implementations such as the computational application of statistical algorithms and other data processing tools, pioneered for example by the Stanford literary lab of Moretti and his peers, enables the modern practitioner of distant reading in turn to explore much larger corpora of text as well as sophisticated mathematical calculations to retrieve latent trends and interactions within the given research data. Thus, while the early beginnings of distant reading and similar formalistic study attempts mainly relied on the manual extraction and calculation of features, today's technical advancements on the other hand have been able to radically enhance the accessibility and magnitude of possibilities when it comes to carrying out the ideas developed by the founding figures of the formal study of literary language such as Yarkho, Propp or Tynianov.

Speaking of accessibility, the popular and well-maintained programming language *Python* with its numerous libraries suited for the task of NLP analysis proved to be a reliable and at the same time also potent tool for realising some of the still untapped avenues that the enticing realm of distant reading entails. Thus, after having explored and stated the goals and trajectories of its research endeavours, the present master's thesis went on to implement a pipeline for the analysis of the object of its study, which was found in the collected adventures of Sherlock Holmes. The main reason behind the decision to work here with the well-known detective stories by Arthur Conan Doyle can be found first and foremost in the often made observation by numerous literary scholars that genre literature such as detective stories adheres most of the time to pretty strict and consistent patterns lending themselves as a result all the more to a formalistic and structuralist research approach. In this notion there had also already been identified – prior to the writing of this thesis – a quite sizeable amount of traditional literary research concerning detective fiction and/or the specific stories of the infamous private detective that focused on extracting and analysing certain recurring overarching forms, features and themes in a manually laborious but nevertheless systematic way. Consequently, this was precisely where the present work recognised considerable potential for further and extensive research on its own terms – not merely in a quantitative sense but also in a qualitative – by limiting itself not solely to a few anecdotal instances of text but rather by being able to widen the scope drastically and taking the in total sixty different stories, revolving around the beloved pair of Holmes and Watson, as a whole into consideration.

The deployment of said pipeline proved to be a quite extensive undertaking on its own, thereby finally resulting in a supplementary *Jupyter Notebook*, containing all the results, visualisations as well as the accompanying code. The first and crucial step of the computational part entailed the preprocessing of the textual data in order to create a normalised and consistent format that could then be fed into the different numerical operations of further downstream tasks. Here the most notable decision – besides the common removal of stopwords and similar words/sequences that do not carry any sizeable amount of meaning themselves – was to segment each text into five parts and handle each created segment as a separate instance within the dataset. Secondly, the pipeline turned to the engineering of the various features, which had been identified as conducive to the constitution of the plot of the stories of Sherlock Holmes beforehand. For this task the underlying mathematical operations stuck mostly to the calculation of the frequencies of certain keywords across the different segments of the given texts and then ran some statistical tests on this newly generated values in order to examine their relative changes as the narrative progresses. In addition this part of the analysis was also supplemented with a range of different exploratory plots as well as clustering approaches, which aided in checking for formal consistencies and narrative stabilities throughout all the sixty different stories. Last but not least, after having computed all the necessary features, the pipeline was rounded up by the employment of more sophisticated

statistical methods such as the classification models of logistic regression and SVM. Not only was the computational analysis able to prove then a considerable stability in the narrative qualities and structures of the literary corpus, it also verified a number of previously postulated claims and notions that traditional scholarship in the realm of detective fiction, and particularly the research surrounding *Sherlock Holmes*, had already been discussing. One noteworthy example in this regard can be found in Nelles' and William's hypothesis that detective stories of the likes of Sherlock Holmes' adventures follow an achronological ordering of its narrative segments. While Nelles and Williams were already able in their paper to analyse some exemplary texts in the context of this question quite satisfactorily, their sole reliance on manual methods hindered themselves in arriving at an exhaustive conclusion nonetheless. In this light the present thesis on the other hand proved to be able to provide said exhaustive backing by undertaking a more automated approach that consequently span over *all* the possible textual instances and their respective features.

All in all it can therefore be concluded that the main goal of bringing together the theory of formalism, structuralism and distant reading with the practical approaches of computational literary analysis proved to be a quite fruitful undertaking. Rooting the here carried out approach with its defined research focus firmly in the traditional problems tackled by scholars within the area of *Sherlock Holmes* and detective fiction was especially successful in bridging the often postulated gap between quantitative and qualitative literary studies, thereby challenging some previous reservations. Furthermore the thesis managed to verify its central claim that a consistent evolutionary pattern concerning the narrative arc of the adventures of Sherlock Holmes could be extracted and observed via the methods of computational processing and statistical modelling. Moving forward, the here laid out perspectives regarding the constitution of a 'futuristic' discipline of literary studies that may finally encompass a more interdisciplinary focus, with the primary focus here being the introduction of technical and mathematical methods, already indicate a step in the right direction. This especially holds true in the light of the very positive findings on the example of the chosen text corpus – while more extensive research with larger as well as more varied datasets has still to be carried out in order to infer from it in full a general agenda of the progression of a distant reading yet to come.

With that being said – if one may allow for ending this present thesis on a more light-hearted note by referring back to Underwood's previously discussed *Jurassic Park* metaphor – the present state of the field might then best be summarised with a question: As the dinosaur has now been let out of his cage, roaming free and mighty on our premises – what do we do with it, how can we best incorporate it into our existence, without it ending in either a cataclysmic moment of auto-destruction or the (renewed) tranquillisation of the so-called beast, thereby putting it back into its cage, where it has already been lying dormant for all those years (or even centuries to say the least)?

Sources

Datasets, Code & Additional Materials

The complete adventures of Sherlock Holmes were downloaded from https://sherlock-holm.es/ [21.4.23].

The complimentary code can be accessed at https://github.com/DavidSiegl/Sherlock-Holmes-Narrative-Arc [17.10.23].

Dictionary of keywords used for measuring the basic narrative arc: https://osf.io/wpcx8 [17.4.23].

matplotlib library: https://matplotlib.org/ [15.10.23].

NLTK library: https://www.nltk.org/ [18.4.23].

NumPy library: https://numpy.org/ [15.10.23].

pandas library: https://pandas.pydata.org/ [15.10.23].

scikit-learn library: https://scikit-learn.org/stable/ [15.10.23].

SciPy library: https://scipy.org/ [15.10.23].

seaborn library: https://seaborn.pydata.org/ [15.10.23].

spaCy library: https://spacy.io/ [15.10.23].

VADER library: https://github.com/cjhutto/vaderSentiment [17.4.23].

Literature

Abbott, Randy L. (2008): Roots of Mystery and Detective Fiction. In: Critical Survey of Mystery and Detective Fiction, ed. by Carl Rollyson. New York: Salem, pp. 1891–1900.

Aizawa, Akiko (2003): An information-theoretic perspective of tf-idf measures. In: Information Processing and Management 39, pp. 45–65.

Archer, Dawn (2016): Data Mining and Word Frequency Analysis. In: Research Methods for Reading Digital Data in the Digital Humanities, ed. by Gabriele Griffin and Matt Hayler. Edinburgh: Edinburgh University, pp. 72–92.

Archer, Jodie and Jockers, Matthew L. (2016): The Bestseller Code. Anatomy of the Blockbuster Novel. New York: St. Martin's.

Ascari, Maurizio (2020): Counterhistories and Prehistories. In: The Routledge Companion to Crime Fiction, ed. by Janice Allan et al. London & New York: Taylor & Francis, pp. 22–30.

Bode, Katherine (2017): The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object of Data-Rich Literary History. In: Modern Language Quarterly 78, 1, pp. 77–106.

Boyd, Ryan L. et al. (2020): The narrative arc: Revealing core narrative structures through text analysis. In: ScienceAdvances 32, 6, pp. 1–9.

Burrow, Merrick (2003): Holmes and the History of Detective Fiction. In: The Cambridge Companion to Crime Fiction, ed. by Martin Priestman. Cambridge: Cambridge University, pp. 15–28.

Campbell, Mark (2007): Sherlock Holmes. Somerset: Pocket Essentials.

Cawelti, John G. (1997): Canonization, Modern Literature and the Detective Story. In: Theory and Practice of Classic Detective Fiction, ed. by Jerome H. Delamater and Ruth Prigozy. London: Greenwood, pp. 5–16.

Ciotti, Fabio (2016): Toward a Formal Ontology for Narrative. In: Matlit 4.1, pp. 29–44.

Debnár, Marek (2018): Formalism and Digital Research of Literature. In: Digital Age in Semiotics & Communication 1, 1, pp. 113–120.

Deerwester, Scott et al. (1990): Indexing by Latent Semantic Analysis. In: http://wordvec.colorado.edu/papers/Deerwester_1990.pdf [15.10.23].

D'haen, Theo (2017): Sherlock's Queen Bee. In: Crime Fiction as World Literature, ed. by Louise Nilsson et al. New York: Bloomsbury, pp. 233–244.

Eichenbaum, Boris (2012): The Theory of the "Formal Method". In: Russian Formalist Criticism. Four Essays, 2nd Edition, ed. by Lee T. Lemon and Marion J. Reis. Lincoln and London: University of Nebraska, pp. 102–136.

Freeman, Austin R. (1924): The Art of the Detective Story. In: http://gadetection.pbworks.com/w/page/7931646/The%20Art%20of%20the%20Detective%20Story [3.4.23].

Gareth, James et al. (2021): An Introduction to Statistical Learning. With Applications in R, 2nd Edition. Berlin: Springer.

Gasparov, Mikhail (2016): Boris Yarkho's works on literary theory. In: Studia Metrica et Poetica 3, 2, pp. 130–150.

Genette, Gérard (2010): Die Erzählung. 3rd Edition. Paderborn: Wilhelm Fink.

Genette, Gérard (2014): The Architext. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, pp. 210–218.

Glazzard, Andrew (2018): The Case of Sherlock Holmes. Secrets and Lies in Conan Doyle's Detective Fiction. Edinburgh: Edinburgh University.

Goodlad, Lauren M. E. (2020): A Study in distant reading: Genre and the *Longue Durée* in the Age of AI. In: Modern Language Quarterly 81, 4, pp. 491–525.

Habjan, Jernej (2012): The Bestseller as the Black Box of distant reading: The Case of Sherlock Holmes. In: Primerjalna književnost, 35, 1, pp. 91–105.

Hutto, C. J. and Gilbert, Eric (2014): VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Proceedings of the International AAAI Conference on Web and Social Media 8, 1, pp. 216–225.

Hoyt, Long and Richard, Jean So (2016): Literary Pattern Recognition. Modernism between Close Reading and Machine Learning. In: Critical Inquiry, 42, 2, pp. 235–267.

Jockers, Matthew L. (2013): Macroanalysis. Digital Methods and Literary History. Urbana, Chicago and Springfield: University of Illinois.

Jockers, Matthew L. (2020): Introduction to the Syuzhet Package. In: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html [17.4.23].

Kayman, Martin A. (2003): The short story from Poe to Chesterton. In: The Cambridge Companion to Crime Fiction, ed. by Martin Priestman. Cambridge: Cambridge University, pp. 41–58.

Kim, Tae Kyun (2017): Understanding one-way ANOVA using conceptual figures. In: Korean Journal of Anesthesiology 70, 1, pp. 22–26.

Kim, Hoyeol (2022): Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons. In: http://www.digitalhumanities.org/dhq/vol/16/2/000612/000612.html#p10 [17.4.23].

Kliger, Ilya (2021): Dynamic Archeology or Distant Reading. Literary Study between two formalisms. In: Russian Literature 122–123, pp. 7–28.

Knox, Ronald (1929): Ten Commandments for Detective Fiction. In: http://gadetection.pbworks.com/w/page/7931441/Ronald%20Knox%27s%20Ten%20Commandments%20for%20Detective%20Fiction [3.4.23].

Lehnert, Wendy G. (1981): Plot Units and Narrative Summarization. In: Cognitive Science, 4, pp. 293–331.

Louis, A.L. and Engelbrecht, A.P. (2011): Unsupervised discovery of relations for analysis of textual data. In: Digital Investigation 7, pp. 154–171.

Lvoff, Basil (2021): Distant Reading in Russian Formalism and Russian Formalism in Distant Reading. In: Russian Literature 122–123, pp. 29–65.

McClure, David (2017): Distributions of words across narrative time in 27,266 novels. In: https://litlab.stanford.edu/distributions-of-words-27k-novels/ [17.4.23].

Moretti, Franco (2005): Graphs, Maps, Trees. Abstract Models for Literary History. London & New York: Verso.

Moretti, Franco (2013): Distant reading. London & New York: Verso.

Moretti, Franco (2013): Pamphlet 6. “Operationalizing”: or, the function of measurement in modern literary theory. In: https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf [9.4.23].

Motwani, Aditya et al. (2019): Extracting Evidence Summaries from Detective Novels. In: Proceedings of the Text2StoryIR’19 Workshop, ed. By A. Jorge et al.. Germany: Cologne, pp. 1–7.

Nelles, William and Williams, Linda (2021): Doing Hard Time: Narrative Order in Detective Fiction. In: Style, 55, 2, pp. 190–218.

Nicol, Bran (2019): Holmes and Literary Theory. In: The Cambridge Companion to Sherlock Holmes, ed. by Janice M. Allan and Christopher Pittard. Cambridge: Cambridge University, pp. 185–198.

Piper, Andrew (2015): Novel Devotions. Conversional Reading, Computational modelling, and the Modern Novel. In: New Literary History, 46, 1, pp. 63–98.

Piper, Andrew (2018): Enumerations. Data and Literary Study. Chicago & London: The University of Chicago.

Propp, Vladimir (2009): Morphology of the Folktale. Texas: University of Texas.

Propp, Vladimir (2014): Fairy Tale Transformations. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, pp. 50–67.

Redmond, Christopher (1993): A Sherlock Holmes Handbook. Quebec: Simon & Pierre.

Ross, Lindsay May (2019): Exploration of Story Arcs in Palliative Care Conversations Using Natural Language Processing. In: UVM Honors College Senior Theses 293, pp. 1–28.

Rzepka, Charles J. (2010): Introduction: What Is Crime Fiction? In: A Companion to Crime Fiction, ed. by Charles Rzepka and Lee Horsley. New Jersey: Blackwell, pp. 1–10.

Scaggs, John (2005): Crime Fiction. The New Critical Idiom. London: Taylor & Francis Routledge.

Schmidt, Benjamin M. (2015): Plot Arceology: a vector-space model of narrative structure. In: 2015 IEEE International Conference on Big Data, ed. by Howard Ho et al., pp. 1667–1672.

Shklovsky, Viktor (1990): Theory of Prose. Elmwood Park: Dalkey Archive.

Shklovsky, Viktor (2012): Art as Technique. In: Russian Formalist Criticism. Four Essays, 2nd Edition, ed. by Lee T. Lemon and Marion J. Reis. Lincoln and London: University of Nebraska, pp. 26–44.

Striedter, Jurij (1989): Literary Structure, Evolution, and Value. Russian Formalism and Czech Structuralism Reconsidered. Cambridge and London: Harvard University.

Swafford, Annie (2015): Problems with the Syuzhet Package. In: https://annieswafford.wordpress.com/2015/03/02/syuzhet/ [17.4.23].

Tobin, Vera (2006): Ways of reading Sherlock Holmes: the entrenchment of discourse blends. In: Language and Literature 15, 1, pp. 73–90.

Todorov, Tzvetan (1973): The Fantastic. A structural approach to a literary genre. Cleveland and London: Case Western Reserve University.

Todorov, Tzvetan (1977): The poetics of prose. New Jersey: Blackwell.

Todorov, Tzvetan (2014): The Origin of Genres. In: Modern Genre Theory, ed. by David Duff. New York: Routledge, pp. 193–209.

Tynianov, Yuri (2019): Permanent Evolution. Selected Essays on Literature, Theory and Film. Boston: transcript.

Underwood, Ted (2016): The Life Cycles of Genres. In: Cultural Analytics, May 23, pp. 1–25.

Underwood, Ted (2017): A Genealogy of distant reading. In: Digital Humanities Quarterly, 11, 2, pp. 1–17.

Underwood, Ted (2019): Distant Horizons. Digital Evidence and Literary Change. London: The University of Chicago.

Vonnegut, Kurt: Kurt Vonnegut on the Shapes of Stories. In: https://www.youtube.com/watch?v=oP3c1h8v2ZQ [10.4.23].

Worthington, Heather (2010): From The Newgate Calendar to Sherlock Holmes. In: A Companion to Crime Fiction, ed. by Charles Rzepka and Lee Horsley. New Jersey: Blackwell, pp. 13–27.

Yarkho, Boris (2016): The elementary foundations of formal analysis. In: Studia Metrica et Poetica 3, 2, pp. 151–174.

Yarkho, Boris (2019): Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism). In: JLT 13, 1, pp. 13–76.

Appendix

Abstract

The present master's thesis aims to show that the employment of computational text analysis and quantitative methods, such as statistical modelling, can produce new and interesting insights for a corpus-driven approach of (digital) literary studies. For this approach the collected stories of Sherlock Holmes written by Arthur Conan Doyle were found to be a suitable subject, given firstly that these kinds of formulaic narratives pertaining to the genre of detective fiction lend themselves particularly well to a more thorough exploration of narratological features, and secondly that this connection has also already been acknowledged by past research, which often focused on Doyle's plot structures rather than actual content. Drawing, historically speaking, from more traditional schools of literary scholarship, namely Russian formalism and structuralism, while at the same time also making use of more recent technological advancements in the realm of NLP, this thesis last but not least argues for a more interdisciplinary approach of literary studies, trying to bridge the still present gap between individual, qualitative judgements on the one side and more expansive, quantitative notions of aesthetic language and form on the other.

Keywords: natural language processing, narrative arc, Sherlock Holmes, Arthur Conan Doyle, distant reading, computational literary studies, sentiment analysis, named entity recognition, plot structure, formalism, structuralism, genre, detective fiction.

Die vorliegende Masterarbeit zielt darauf ab zu zeigen, dass der Einsatz von computergestützter Textanalyse und quantitativen Methoden, wie statistischer Modellierung, neue und interessante Einblicke für einen korpusbasierten Ansatz der (digitalen) Literaturwissenschaft liefern kann. Für diesen Ansatz wurden die gesammelten Geschichten von Sherlock Holmes, verfasst von Arthur Conan Doyle, als Untersuchungsgegenstand ausgewählt. Erstens deshalb, weil sich derartige formelhafte Erzählungen im Genre der Detektivgeschichten besonders gut für eine genauere Exploration narratologischer Merkmale eignen, und zweitens, weil dieser Konnex bereits von einer Vielzahl an vorangegangenen Forschungsarbeiten anerkannt wurde, die sich oft auf Doyles Handlungsstrukturen konzentrierten, anstatt auf inhaltlicher Ebene anzusetzen. Historisch betrachtet beruft sich die Arbeit auf traditionellere Schulen der literarischen Forschung, nämlich dem russischen Formalismus und Strukturalismus, und macht gleichzeitig aber auch Gebrauch von neueren technologischen Methoden im Bereich der natürlichen Sprachverarbeitung (NLP). Letztendlich plädiert die Arbeit für einen interdisziplinären Ansatz der Literaturwissenschaft, der versucht, die immer noch bestehende Kluft zwischen individuellen, qualitativen Beurteilungen einerseits und umfassenderen, quantitativen Konzepten von ästhetischer Sprache und Form andererseits zu überbrücken.

Schlagwörter: Natural Language Processing, Erzählverlauf, Sherlock Holmes, Arthur Conan Doyle, Distant Reading, computergestützte Literaturwissenschaft, Sentiment Analysis, Named Entity Recognition, Plotstruktur, Formalismus, Strukturalismus, Genre, Detektivgeschichten.