universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Latent Relationships and Networks of Influence in Gothic Fiction

verfasst von | submitted by

Florian Klement BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | 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:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Tara Andrews

I would like to thank my supervisor Univ.-Prof. Mag. Dr. Tara Andrews for guiding me through this academic journey.

Sincere gratitude goes out to my parents for their unwavering support over the past years, Marina Seido for the moral support, Dusty Keim for advice and proofreading, and DeepL Write for straightening out winding Germanic sentences to be more palatable to an English audience.

Table of Contents

1 Introduction:

This master's thesis explores the use of state-of-the-art computational methods for the discovery of latent structural patterns in Gothic Fiction texts. To achieve this, a number of natural language processing (NLP) techniques commonly used by the Distant Reading community—literary scholars that employ quantitative means—have been employed. The study utilized topic modeling to investigate the salient attributes of the genre. Specific authors and their texts were grouped based on their contribution to a given topic. Network analysis and similarity metrics were then used to cluster the texts and make inferences about influence among them, clusters and groupings within the corpus of 182 texts.

The corpus comprises texts from three distinct sources. The first source is a list of books collected by Caroline Winter and Eleanor Stribling as part of their work exploring the color space used in Gothic fiction texts.[1] The second source is a collection of texts considered by Gothic scholars David Punter and Glennis Byron to be the most prominent contributions to the genre in their seminal works on the Gothic, curated by digital humanist Ted Underwood.[2] The third source comprises texts drawn from Project Gutenberg's 'Gothic Fiction' shelf.

Questions regarding the defining characteristics of literary genres are a crucial aspect of literary criticism. Numerous comparative literature scholars have dedicated themselves to comparing extensive amounts of texts to comprehend the elements that constitute a genre that transcend characteristics used in a particular period or language. In the 20th century, scholars of Formalism and Structuralism made significant progress in codifying the abstract makeup of texts and creating systematic models of their integral relations. Although most of these investigations remained qualitative and relied solely on a scholar's judgment of the texts they had read to make assessments of the makeup of texts and frequencies of patterns, one Formalist stood out for his quantitative methods. Boris Yarkho[3] and his team of researchers painstakingly conducted manual word counts and recorded the number of entrances and exits in Romantic tragedies to quantify their assessments.

[1] https://www.uvic.ca/humanities/english/home/news/archive/gothic-colours-pycon-2017.php [08.07.24]; https://github.com/eleanorstrib/gothic [08.07.24]

[2] See Underwood, "The Life Cycles of Genres;"; https://zenodo.org/records/51361 [08.07.2024].

[3] Eg Yarkho, "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)."

A new generation of researchers has taken up the aim of structured analysis of texts. Scholars of Distant Reading, such as Ted Underwood,[4] David Mimno,[5] Ryan Heuser,[6] and Andrew Piper,[7] have employed statistical methods to address these questions with computational means and large selections of texts. In the digital age, natural language processing (NLP) provides literary scholars with a toolkit to interpret large quantities of texts and infer common structural elements.

The formulaic nature of many works of Gothic fiction, with its consistent cast of character roles and recurring themes, invites analysis of its core building blocks. This has made it a frequent subject of exploration in traditional literary criticism, particularly within Structuralist and Formalist frameworks. A notable example of this is Tzvetan Todorov's seminal book, *The Fantastic*,[8] which is a prominent work in Structuralist theory on the function and compositional elements of genre. The investigation set forth in this master's thesis aimed to align these perspectives with the toolkit of modern digital scholarship and apply popular NLP methods in order to contrast the results of traditional criticism with the results of a more quantitative perspective on text.

While the comparison of NLP method results with the qualitative assessments of David Punter, Glennis Byron, and Fred Botting shows promising hints at the composition of early traces of influence within the genre, further study is needed to situate these findings within the larger canon. The collaboration between Distant Reading and regular literary criticism opens up new avenues for investigation that will prove mutually enriching. In an increasingly digitized age, alternative methods for evaluating literary history and recurring themes that tap into larger wells of preserved, digitized, and accessible human knowledge beyond human readability are a promising opportunity for future scholarship.

> Humanists have always been explorers […] They sail not the seas of water but on seas of color, sound, and, most especially, words.[9]

[4]Eg Underwood, *Distant Horizons - Digital Evidence and Literary Change*.
[5]Eg Mimno, "Computational Historiography."
[6]Eg Heuser, "Mapping the Emotions of London in Fiction, 1700–1900."
[7]Eg Piper, *Enumerations*.
[8]Todorov, *The Fantastic*.
[9]Smith 1984, 20. as cited in Slingerland et al., "The Distant Reading of Religious Texts," 1011.

2 Form and Genre within Literary Studies:

In order to study the development of a genre over time, it is necessary to understand the circumstances in which it arose and to shed light on its influences. The study of literary form has a long and complicated history, dating back to Aristotle's *Poetics*. He presented a fundamental prescriptive text of elements such as style, agents, theme, plot, and form. Many classical texts followed the format laid down by Aristotle, and convention from the Renaissance until the early eighteenth century would urge poets to adopt and revive these customs. While the classicists of the early eighteenth century valued the formal conventions of antiquity, the following generation of Romantic writers rejected the formal constraints and moral precepts of their predecessors. The requirements for poetic writing aimed for mimesis, the imitation of natural life, the retention of clear boundaries of form, and the depiction of the perseverance of morally good behavior. Many argue that Gothic fiction is a specific derivative of the Romanticism movement. The Romantics embraced new modes of expression; they created freer forms, that would come to define the styles of much of modern fiction. Romanticism saw the loosening, breaking up, reconfiguration and recreation of textual genres through the use of new techniques and materials. Further defining characteristics are the inclusion of wonder, awe and childlike creativity, a longing for something past and mythical, as well as a reappraisal of emotional drives as opposed to the position of reason within the modern industrial society.[10] Their preoccupation with an enchanted, glamorized nature, in addition to the tendency towards analogical association and the recreation of folkloric themes, is something that connects them again with prominent figures of the academic study of literary form. Many of them, like Johann Bodmer or the Schlegel brothers would go on to write their own theories of poetic form, combining aesthetic production and literary criticism. Vladimir Propp was a prominent figure within Russian Formalism, who dedicated himself to the analysis of myth, the morphology, and the elements within folktales. His approach is exemplary for a tradition of disassembling texts into small constituent elements, analyzing the individual pieces and drawing conclusions about the composite nature of the larger whole.[11] More general structural approaches will be discussed in greater depth at in Chapter 5.

Contemporary genre theorist David Fishelov recounts how the use of biological and evolutionary terminology has become more widespread since the Formalists brought them into the literary tradition and argues in favor of such borrowings, because of the rich conceptual language of

[10]Cf. Simpson, “Romanticism, Criticism and Theory,” 7ff.
[11]See Propp, “Morphology of the Folktale.”

distinction and development that they entailed. It promoted the extension of the existing frame of reference through functional analogies, while enabling more precise evaluation of the genealogy of genres if done mindfully.[12]

Many of the contemporary approaches to the study of tropes, forms, and narrative devices in current narratology owe the beginnings of their field to their roots in Russian theory from the late 19th century onward. Formalism as a tradition came into being as a critical reception of the mythopoeic and vast interdisciplinary works of Alexander Veselovsky. It is regrettable that his texts have not yet been translated directly. However, the current revival of investigations into his works in Slavic studies is promising in this regard. The first coinage of the term 'genre' can be traced to his historical studies, which engaged with the concept on a much broader and more general scale than any of the Formalists who followed in his wake, tracing the rise of individual literary forms and styles from a common historical oral tradition.[13]

Veselovsky considered literary form as a recursive, mediated response to historical processes. Literary traditions and their forms emerge out of shared socio-historical conditions, insinuating that similar literary forms would develop in parallel if given similar circumstances. He conceptualized the shaping force of literary forms throughout time, as a radically continuous shared experience creating similarities, that are perforated by non-synchronous, stratified, and heterogeneous experiences of individuals on the outskirts. Older, forgotten forms and styles that were once popular would unpredictably reemerge, seemingly as novelties, if the times reflected their demand.[14] Veselovsky's approach to form is comparative, focusing on the evolution of literature as a set of recurring motifs, themes, and styles shared within interconnected cultures dialogically, as opposed to innovation born out of the relations between individual authors and movements. Some of his important insights centered around genre theory, motif migration, and sociopsychological roots of specific styles of writing.[15]

> […] [T]exts appear as intrinsically hybrid entities, whose semantic layers speak in different voices, which may be audible or silent at various moments in the text's reception, entering into new or recurrent constellations that are largely outside authorial control.[16]

This conception of literary conventions views texts as vessels of unintended, codified facets of a social experience. These facets survive only to be reawakened in future works that grapple with

[12]Cf. Fishelov, "The Strange Life and Adventures of Biological Concepts in Genre Periodization."
[13]Cf. Nikitina and Tuliakova, "Genre Studies in Russian Literary Research."
[14]See Maslov and Kliger, Persistent Forms: Explorations in Historical Poetics, 4f.
[15]See Maslov and Kliger, 21.
[16]Maslov and Kliger, 7.

them anew. Framing structural components as informed by social circumstances bridges the gap between text-focused and context-focused methods of investigation. This unifies their influence on the atomic elements of a composition. A move that subsequent Structuralists in the late 20th century replicated as well, when they introduced psychoanalytic elements into their literary investigations. An alternative way to frame Veselovsky's distinction between a text and the compositional forces that it strains against is by using the term 'horizon of expectation,' as coined by the reception theorist Hans Robert Jauss.[17] This term signifies the frame of reference against which a given text is compared at the moment of its inception, providing a backdrop for attempts at understanding and contextualizing it. This concept can be situated at the intersection of what Veselovsky would consider the synchronous and the heterogeneous, the unexpected and unique crossing of the boundaries of convention.

Tzvetan Todorov played a major role in the rise and development of French Structuralism. His translations of many of the critical texts of Russian Formalism, as well as his own works, proved foundational to the movement. Of particular interest here is his work *The Fantastic – A Structural Approach to a Literary Genre.*

Todorov's conception of theoretical genre is quite comprehensive and broad. There exists the tendency within literature to delegitimize the epigonal and formulaic as the realm of popular fiction. In opposition to this, he poses broad questions on various levels of generality, bringing up the unifying need for classification and standardization in language to allow for descriptive communicative intent. The rejection of genre and the insistence on uniqueness as the sole indicator of quality denies the reality of pattern formation in literary discourse. This transgression is only possible through the warranted notion of a specific form. Therefore, it is necessary to codify a norm and set a relay point of reference for a chance at interpretation. In order for a literary text to establish referential relationships, it must adhere to the postulates of the system of signification, ensuring that its validity can be evaluated.[18]

Theories of genres are based on notions of abstract properties and laws governing the relations between them. Todorov distinguishes between three interrelated perspectives on works of literature, against which investigations of genre need to be compared: First, verbal aspects encompass the concrete sentences, which are bound to a given narrative perspective or are considered as utterances of individual speakers. Second, syntactic aspects relate the structure and

[17]Eg Jauss and Benzinger, “Literary History as a Challenge to Literary Theory,” 12.
[18]See Todorov, The Fantastic, 7ff.

components of the work to itself, they compose the structure through logical, temporal and spatial means. Third, semantic aspects pertain to the articulated themes that are transformed, combined and contrasted. An individual text does not need to coincide with a single category, as categories merely pose as constructed abstraction, while works are an expression within the tensions of a given system of narrative components and social relations. Texts participate in, but do not belong to genres. Their forms transform and develop from each other.[19] This investigation has excluded syntactic aspects during preprocessing and distilled verbal aspects to semantic aspects using latent dirichlet allocation. Further details will be provided in Chapter 6.

Much of the contemporary research on the genre of Gothic fiction comes from the field of popular culture studies. There, genre is defined as a collection of shared styles, forms, and content that relies heavily on the reader's contribution to keep the established conventions in mind and to help fill in intentional gaps or contextualize intentional deviations.[20] Here, genre is instead conceived in a more rigid format that is collectively maintained and becomes more apparent when broken or intermixed. This analysis becomes particularly relevant in fast-paced forms of media, where feedback from the audience plays a crucial role. The blending of tropes and hybridization is closely related to the appeal to established forms. It is worth noting that the more established a genre is, the more likely it is to become subject to parody. Genre studies heavily incorporate Structuralism and archetypal criticism, which were developed by and in response to Northrop Frye.[21] These approaches also draw heavily on mythology and the work of Carl Gustav Jung.[22] For the purpose of this investigation this would raise the assumption that the point in time where popular parodies arise would mark a turning point where a first clear image of generic requirements had been reached.

When typical elements of a genre from different categories are combined, they form motif complexes, for example a setting and an object. When multiple motif complexes are aligned into a recognizable narrative configuration, they form a sub-formula. For example, the development of penny dreadfuls as a cheap form of entertainment depended on the ability of writers and publishers to recombine recognizable and easily interchangeable elements to enable mass production. Their success depended on their reproducibility and inherent sensationalism, while avoiding direct references to contemporary social and political circumstances.[23] Hoppenstand

[19]Todorov, 20ff.
[20]Jenkins, “The History and Logic of Genre Study,” 85ff.
[21]Cv Todorov, The Fantastic, 8ff.
[22]Jenkins, “The History and Logic of Genre Study,” 89.
[23]Hoppenstand, “Genres and Formulas in Popular Literature,” 103f.

attributes much of the success of early Gothic fiction to its moral polarization and its appeal to repressed emotions. According to Hoppenstand, for instance, Walpole,[24] who blended folklore elements into his works and emphasized escapist notions and violence, while Radcliffe[25] catered to a female audience and incorporated elements of romance.

In recent years, the term 'genre' has become less popular in literary studies. Some of its applications have been incorporated into discourse studies that focus on continuity and socially constructed perspectives, bringing additional attention to the temporal plurality of modes of framing texts. Anis Bawarshi's efforts to unify Foucault's concept of the author's function with genre in interdisciplinary ways have yielded interesting results. The author function refers to a discursive space of regulated intertextual interpretive weight beyond and behind a given text. The author carries weight on their collective reception beyond the content of any individual piece of text they have written.[26] Bawarshi proposes that the concept of genre functions as a framework for reading texts, interpreting the intent and constructing a contextual space to assign agency and intent to actors within them. This framework proposes constructing temporary identities within the context of the actors involved at the time of creation.[27]

In contrast to this, the notions of genre brought forth by Franco Moretti seem polemic and polarizing and have been criticized by his contemporaries. He undoubtedly popularized the practice of Distant Reading and strove to bring new unifying impulses into comparative literature, but he has since been disavowed by larger sections of the community.

He suggests that genres rarely survive more than 25 years and work in overlapping generational periods, claiming that continuity among longer or reoccurring patches of text might be independent projects driven by demand and evolution.[28] Both Matthew Jockers[29] and Ted Underwood[30] found decisive evidence against this suggestion when analyzing literary genres in their studies. Underwood concludes that genre is a more mutable set of relations between works that resemble each other and link to one another in varying degrees. His analysis posits that genres are a mixed category with different life cycles and levels of textual coherence.

---

[24]Eg Walpole, *The Castle of Otranto*.
[25]Eg Radcliffe, *The Mysteries of Udolpho, a Romance*.
[26]Bawarshi, “The Genre Function,” 340ff.
[27]See Bawarshi, “The Genre Function,” 342.
[28]Moretti, Graphs, Maps, Trees, 3–30.
[29]See Jockers, Macroanalysis, 82ff.
[30]Underwood, *Distant Horizons - Digital Evidence and Literary Change*, 40; Underwood, “The Life Cycles of Genres,” 19.

In summary, various conceptions of genre characteristics have been discussed in this section. Through Veselovsky's lens, genre can be understood as a convention within a larger dialogue that shares historical ties with a plurality of voices and a common embedding of social circumstances.

Todorov's conception, on the other hand, focuses more on participation in a selection of various registers that retain a stable set of codes and a scope of references. The study of popular fiction suggests that a genre can only be subverted once its boundaries have been firmly established, and the reader's knowledge of the conventions is essential to understand the dialogue an author enters with their work on its generic requirements. Contemporary conceptions support the interactivity of the concept of genre, while the works of Moretti propose clear temporal delineations of 25 years per genre. The following section will examine Gothic fiction in greater detail.

3 Gothic Fiction:

> A more radical claim would be that there are very few actual literary texts which are 'Gothic'; that the Gothic is more to do with particular moments, tropes, repeated motifs that can be found scattered, or disseminated, through the modern western literary tradition.[31]

Todorov is also well known for his detailed analysis of fantastic literature of all types. While his broad strokes categorizations have not been taken up equally well by all critics,[32] the bigger picture of his analysis and the crystallized characteristic traits he assigned different types of fantastic texts continue to impact the field.[33] He defines the fantastic as a specific literary genre located between the uncanny and the marvelous. In the marvelous, the reader knows that irrational forces are operating in the world and can draw causal connections to events within the text to these forces; here Todorov lists fairy tales and science fiction. In the uncanny, however, weird events are given a rational explanation to compartmentalize them, be it madness, intoxication or a dream; the ambiguity in these texts creates structural tension. Ann Radcliffe is especially known for deflecting the tension created in her works through retrofitting mundane explanations onto supernatural occurrences.[34]

However, it is not just structural tension driving these texts. Most critics in the last decades have agreed that Gothic fiction is a fruitful target for the use of psychoanalytical methods of analysis, given the high degree of transgression, social taboo, desires and analogical reasoning present in the genre. As the preceding chapter recounted, the Romantic period introduced more elements of wonder, awe, explorations of unusual affects, and appreciation for emotion and suffering, a desire for authenticity and nature, and a retreat from the rationality of modern industrial society. Many of these qualities present in Romanticism are carried over into its darker, more popularized relative, Gothic fiction. A number of the authors active in the more highbrow movement, including E. T. A. Hoffman, Friedrich Schiller, Ludwig Tieck and Lord Byron, produced Gothic fiction as well. Goethe dismissively remarked on the aggravating, defeatist, and lachrymose attitudes of his contemporaries.

> Die Poeten schreiben alle, als wären sie krank und die ganze Welt ein Lazarett. Alle sprechen sie von dem Leiden und dem Jammer der Erde und von den Freuden des Jenseits und unzufrieden, wie schon alle sind, hetzt einer den andern in noch größere Unzufriedenheit hinein. Das ist ein wahrer Mißbrauch der Poesie,

[31] Punter and Byron, *The Gothic*, xviii.

[32] *See Brooke-Rose, "Historical Genres/Theoretical Genres: A Discussion of Todorov on the Fantastic."*

[33] *Cf. Parisot, "The Aesthetics of Terror and Horror."*

[34] See Punter, *The Literature of Terror - Volume 1*, 60.

> die uns doch eigentlich dazu gegeben ist, um die kleinen Zwiste des Lebens auszugleichen und den Menschen mit der Welt und seinem Zustand zufrieden zu machen.[35]
>
> [The poets all write, as if they were sick and the whole world a sick bay. They all talk of the pain and the wailing of the earth and the joys of the after world, and discontent, as they all are, one agitates the other to even greater discontent. This is a true misuse of poetry, which is actually given to us, in order to balance the small scuffles of life and to make humans at ease with the world and their state.]

In *The Aesthetics of Terror and Horror*, Parisot analyzes 18th century criticism, detailing depictions of the emotional states specific to Dark Romanticism and the Gothic. The genre's prevalent sensations, horror and terror, fulfill different functions, even if the usage of these terms had not been disambiguated for a long time. Horror is the experience of contraction and recoil, often accompanied by a paradoxical pleasure. Terror leads to an imaginative expansion of one's sense of self, an introspective sensation, in part evoking the sublime and transcending or expanding one's scope of reference. It is an excitement and fear that marks an uplifting thrill. It is a distinct contraction of the imminence and unavoidability of the threat. After horror glimpses invasion, terror expels, reconstituting the boundaries that horror has seen dissolve.[36] The pricking of moral sentiment was perceived as an exercise of one's benevolent sensibilities, leading to sympathy or, if overdone, revulsion. The nuance required to balance the pleasurable impulses for internal reflection, which some even ascribed pedagogical qualities to, and the moral violence of its readers was not always a given, and the lack of it could attract other, more unseemly attribution.

Furthermore, Parisot hints at another feature of Romantic literature, noting that it is characterized by a strong incorporation of the sublime in their aesthetic visions.[37] The term 'sublime' was originally coined by Burke,[38] but came to define much of Romantic writing. The sublime can be described as an emotional experience of a person's diminutiveness in the face of larger forces at play. This can entail awe and fascination, but also reframe an individual's subjectivity within a larger whole. In many cases, the source that elicits these feelings is a force of nature, but it can also be unnatural or otherworldly.

Todorov's conception can be linked to Parisot's; in uncanny works, supernatural elements are used as an agent to provide contained emotional relief. This can be achieved by either putting one's own experience into perspective when exposed to sublime or overly extreme content, or

[35] *Johann Peter Eckermann, "Gespräche Mit Goethe in Den Letzten Jahren Seines Lebens. 24. 09. 1827."*
[36] See Parisot, "The Aesthetics of Terror and Horror," 290ff.
[37] See Isbell, An Outline of Romanticism in the West, 36.
[38] See Burke, *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*.

through exploring gruesome themes that the reader can indulge in from a safe distance. This safe distance is either created through appalling qualities that allow one to recoil, or through dispelling the veracity of the experience and providing a moral framing.

The boundlessness of an overly ornamental style and the transposition of the plot into a fictional medieval setting full of grandeur, looming sights of natural beauty, and archaic architecture evokes associations of feudal and dark times. In many of the texts, aristocratic characters wear the trappings of chivalry and romance, but the values portrayed are decidedly from the realm of family, domesticity, and bourgeois individualism. The imitations of outward form and stylistic aesthetics do not paint a coherent picture, yet it does create a setting sufficiently removed from the everyday life of its readership to openly test conventions and their subversion.[39]

Many of the main themes center around sexual transgression, as well as physical and psychological harm of protagonists under duress. The causal chain between infiltration of norms and their restitution through punishment within the narrative is sometimes only perfunctorily maintained. Some of the excess and magical thinking can be explained as insurgences against the conventional 18th century demands for simplicity, probability, and restraint. The naming convention itself speaks of the glamorized image of an emotional well of wilderness and the drawing of inspiration from a hidden source of power, given the limited knowledge and glamorously ambiguous portrayal of research into their predecessors at that time.[40]

> External forms were signs of psychological disturbance, of increasingly uncertain subjective states dominated by fantasy, hallucination and madness. The internalisation of Gothic forms reflected wider anxieties which, centering on the individual, concerned the nature of reality and society and its relation to individual freedom and imagination. […] The disturbance of psychic states, however, does not signal a purely subjective disintegration: the uncanny renders all boundaries uncertain and, in nineteenth-century Gothic writing, often leaves readers unsure whether narratives describe psychological disturbance or wider upheavals within formations of reality and normality.[41]

Not all the boundaries that were crossed and certainties that were dissolved concerned the sphere of the individual. Many of the texts were covert critiques of social structures, whether familial or societal. While many Gothic romances deal with forced marriage, family history, and the loss or gain of an inheritance, the texts of Charles Brockden Brown or William Godwin explore and question tumultuous and cruel societal structures.

With the advent of criminology and the entrenchment of Darwinism in popular discourse in the 19th century, the depictions of deviation from societal norms took on a more overtly

[39]See Botting, Gothic, 5ff.

[40]See Punter, The Literature of Terror – Volume 1, 5.

[41]Botting, *Gothic*, 7.

psychological nature, as texts like *Dr Jekyll and Mr Hyde*, *The Devil's Elixirs, Siebenkäs or the Sandman* dealt heavily in doubles and alter egos, thus combining topics of alienation with the uncanny, emphasizing the struggles of the animalistic, instinctual sides of humanity that are heavily laden with unconscious desires, as well as explore the release of repressed energies and antisocial fantasies.

The first heyday of the genre can be situated in the period between 1760 and 1830, featuring many tyrannical fathers, dead mothers, daughters with a remarkable ability for survival, and awe-inspiring, endlessly resourceful villains in pursuit of opaque ends. The ensuing demographic changes in future waves of production are also reflected in the portrayed scenery. The frequency of castles and abbeys and the preoccupation with purity fades into urban city sprawls, the archetypal laboratory, and depictions of fragmented individuals.

Punter speaks of 'paranoiac fiction,' fiction that animates its readership towards apprehension and doubt with regard to the certainty towards the congruity of the text; unreliable narration is a staple of the genre, thereby separating the texts Todorov would regard as marvelous from the uncanny.[42] The Gothic atmosphere anticipated many aspects of modernism, emphasizing the subjectivity of reality and the collapse of objectivity and the self. However, fears and disconnect from the products of human imagination still largely remained externalized.

Although the images and figures are reiterated, the projected cultural fears and fantasies carry variations and differences in significance depending on the cultural and historical point in time in which they stem from. This results in discontinuities and breaks in convention as the genre continues to develop and morph. What did stay consistent was its eerie veneer, emphasizing its distance from the everyday and the fragmenting effect of the encounters of its participants. What might loosely be referred as the 'Gothic tradition,' with a common set of motifs, settings, creatures and drives, is quite fragmented in much of its form and diverse in the sources it draws from, which includes Shakespearian Drama, Spenserian poetry, medieval romances, ballads, and folklore.[43] This looseness in its conception allowed for a productive overlap with science fiction, sensational fiction, crime fiction, historical fiction, romance, early adventure novels, political opinion pieces, and/or generational dramas. This led some of the writers of early twentieth-century supernatural fiction to prefer the self-description of 'weird fiction,' focusing their conception on heightened emotions and transgressed boundaries of reality.

[42]See Punter, The Literature of Terror – Volume 2, 183.
[43]See Botting, Gothic, 10.

Of further note is the genre's frequent use of antiheroes, sympathetic yet strong villains, gender ambiguous coded characters, and strong female characters, which presents a strong contrast to the otherwise heavily stereotyped characters.[44]

The use of dichotomies of emotional ambivalence, the sacred and the unclean, attraction and disgust, and dread and pleasure provide a nostalgic glimpse at a negative psychology of resistances within bourgeois society at times of advancing industrialization and changes in the speed and shape of communal life, through symbolism and self-mythologization.[45]

David Punter argues that Gothic fiction contributed to the development of modern novels with its convoluted and strong plot lines, which strengthened the sway of narrative complexity and opened a new discursive field. He summarizes the potential approaches to the genre as five different vectors of analysis: First, its reach for the divine and transcendence, like its architectural inspiration, second, its advances to plot structure, third, the narrative difficulties arising out of them—he argues that the unresolvable complexity is an evasive response to taboo—fourth, an agonistic attitude towards realism and lastly, the distinct and subterranean themes and the style it deals with.[46]

In summary, gothic fiction can be described as a common set of motifs, settings, creatures, and drives that serve a form of tension and relief set up on an individual, interpersonal, or societal level; in many ways, it can be described as a form of transposed wish-fulfillment or escapist fiction in unreal or unreliable circumstances. It contains strong contrasts and dichotomies, and has a high degree of overlap with other forms of fiction in existence at the time of its conception. Much of its interpersonal complexity and attempts at textual distortion of its psychological undercurrents would then go on to influence much of modernist fiction to come.

[44]See Day, In the Circles of Fear and Desire, 169.
[45]See Punter, The Literature of Terror - Volume 2, 188ff.
[46]See Punter, The Literature of Terror - Volume 1, 16ff.

4 Distant Reading:

4.1 Concerns and Perspectives

Computational efforts in Literary Criticism have been gaining momentum int the last 30 years and have since expanded their range of methods. However, a growing community has existed since the 1960s. One of the pioneers of string comparison metrics, Fred Damerau, developed authorship attribution tasks in the 1970ies, while many of his contemporaries searched for ways of fitting standard statistical hypothesis testing onto textual data. The techniques used were particularly resource efficient by today's standards, and the research questions were detail-oriented. A persistent challenge has been the lack of a unified methodology that can be considered a coherent framework for building upon previous work. However, some practices have remained just as widely used, including principal component analysis and linguistic techniques, such as lemmatization, part-of-speech tagging, and word co-occurrences, as well as many of the tools used in authorship attribution and stylometry. These tools have maintained close ties with computational linguistics and have proven to be fruitful.[47] *Computation into Criticism – A Study of Jane Austen's Novels and an Experiment in Method* by J.F. Burrows, which employed such methods, is a particularly comprehensive work of that time. It is frequently cited for its novel approach to the use of linguistic elements that are often regarded as 'noise,' which, in this context, refers to disruptions in the measured pattern. This noise is then removed, and thus overlooked in many more expansive research approaches outside the field. Burrows examined the use of punctuation, pronouns, and conjunctions in Austen's novels. He drew conclusions about the interrelationships of the characters that were consistent with the results of popular works of close reading, adding further weight to them.

In the 2010s, machine learning methods have become increasingly prevalent in computational literary scholarship. This has expanded the range of available tools and approaches for tackling research problems. Supervised learning methods, such as support vector machines,[48] random forests,[49] and logistic regression,[50] have been widely used in computational literary studies.

[47]Beausang, "A Brief History of the Theory and Practice of Computational Literary Criticism (1963-2020)," 183–186.

[48]See eg Hou and Jiang, "Analysis on Chinese Quantitative Stylistic Features Based on Text Mining."

[49]See eg Saccenti and Tenori, "Multivariate Modeling of the Collaboration between Luigi Illica and Giuseppe Giacosa for the Librettos of Three Operas by Giacomo Puccini."

[50]See eg Underwood, "The Life Cycles of Genres."

Additionally, vector space representations for textual comparison and clustering methods have gained popularity. The same can be said for the use of network analysis to map relationships.[51]

With the inclusion of bootstrapping, cross-validation and permutation tests, or even the modulation of research questions to be fit for hypothesis testing, many of these techniques can reach a robustness of validity and significance in their results. But Distant Reading, and Digital Humanities more broadly, face not only technical challenges, but also the challenge of realigning their tools with their research questions.

> In literary criticism, scale is often a function of similitude. A reader discovers a feature that is common between texts, or between a text and the conditions of its creation. If that feature (a plot device, a formal innovation, a character) can be meaningfully linked to an underlying commonality among a group of texts (a shared genre, period, or author), if it is similar in function or form in related texts, then the feature can be argued to be meaningful at much larger scales. The similitude that permits such scalar arguments depends upon abstraction: to be compared across scales, texts are reduced to the features that they have in common. The new economies of scale evident in the quantitative turn in literary criticism share this process of abstraction, but replace similitude with measurement, reversing the order of the analysis.[52]

Distant Reading aims to establish unifying and familiar approaches to literary texts to facilitate continuity in reference, study, and discourse. However, many studies have been criticized for their lack of objectivity and for failing to adhere to conventional structures and clear, objective language, which has resulted in inconsistencies in scope, research questions, and approaches to generalization, which has strained the continuity of efforts and results under different signs, times, and methodologies of literary history. As Beausang has regrettably pointed out, the same concerns about the validity of the results or the insignificance of the scale and the level of generalization of the work that the following papers lament,[53] have been raised by previous generations of researchers about methodologies that were less robust and less standardized. They are repeated almost verbatim, despite the advances in practices and the integration of more and more interdisciplinary efforts to enrich, consolidate, and unify efforts in digital philological work.[54]

According to Hammond, Distant Reading's contribution to literary criticism is either highly speculative and not in line with the established consensus on the texts in question, or it is stuck in a cycle of validating its methodology against the results of close reading. However, he does not acknowledge the epistemological significance of validating existing research with different tools

[51]Beausang, 191.

[52]Algee-Hewitt, "Distributed Character," 751.

[53]Eg. Drucker, "Why Distant Reading Isn't."; Accord Bode, "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History."; Accord Hammond, "The Double Bind of Validation."

[54]Beausang, "A Brief History of the Theory and Practice of Computational Literary Criticism (1963-2020)," 193.

that do not carry the same assumptions as orthodox scholars. Nevertheless, Hammond rightfully draws attention to one of the core self-prescribed tasks of Digital Humanities as a whole: To reconcile quantitative and qualitative paradigms and needs, and to interlace their methodologies and the interpretation of results.[55] Bode advocates for the application of an intertwined attempt at 'mid-range reading,' a technique that combines attributes of both close reading and Distant Reading and historicizes its object of study sufficiently to bring the caution necessary for proper interpretation of its data points. Andrew Piper supports this, stating that close reading should complement Distant Reading in order to 'validate whether a model captures the [...] frameworks, it is meant to approximate.'[56]

> These analyses demand new kinds of abstraction: ones that can take into account minute effects in single texts replicated across a corpus at large. Establishing metrics, finding patterns, and linking these metrics and patterns with meaningful concepts of literary criticism: these are tasks that digital humanities now faces. In doing so, it must ask new questions about the nature of what it studies and how computational results are compatible with the established practices of literary criticism.[57]

Fotis Jannidis notes that empirical quantitative practices are often smaller in scale, and more fractured and delicate in their questions, to the extent that they appear incompatible and unattractive to the body of research of their qualitative branches. This leads to little response and reception outside its own circles. Only when its results are taken out of their narrow confines, extrapolated, and integrated into the original frame of reference established by the discipline does it gain more attention, even if at the cost of methodological tensions. This leads to the dichotomy of the cases discussed above, which facilitates a more fragmented research community and field.[58]

Piper discusses the possibility of a circular flow in literary criticism to refresh, review, and potentially restructure current ideas on how to conceptualize specific categories in literary investigations. Close reading analyzes the individual and its characteristics, integrating these beliefs into a model of the larger whole. Distant Reading can measure these concepts by attempting to model their form and specifications. The results of large scale quantitative analysis can then, once processed and prepared accordingly, invite further close reading and interpretation of the found patterns.[59] This interpretation can serve as the basis for redefining the abstract concepts in use within the larger discipline, leading to the formation of new beliefs and avenues

[55]Eg. Hammond, “The Double Bind of Validation,” 8.
[56]Piper, *Enumerations*, 655.
[57]Algee-Hewitt, “Distributed Character,” 751.
[58]Jannidis, “Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft — ein Essay,” 661.
[59]Piper, *Enumerations*, 10.

for investigation. The integration of regular criticism and computational approaches can be achieved through a conscious discourse on the different presumptions of representation and what it means to produce worthwhile literary research results. This would allow for a fruitful collaboration between the two fields of study, mutually enriching them in the process.

Stephen Ramsay posits that, in order for computational methods in the humanities to fulfill their role effectively, they must be incorporated into a framework which comprises multiple layers of analogue criticism and close reading.

> [Works of computational text processing] are able to disrupt not because they lay claim to deep textual truths, but because they are capable of presenting the bare, trivial truths of textuality in a way that allows connection with other narratives—in particular, those narratives that seek to install the text into a network of critical activity.[60]

The last major attempt to provide a foundational set of concepts for Distant reading was ambitious, but overly bold in its claims and generalizations in the hope of creating a level playing field. Since the turn of the century, Franco Moretti has advanced the efforts of computational philology and comparative literature through a number of influential, yet polarizing papers, which were later collected in books. His premise is to expand the efforts of world literature to encompass larger, global interactive patterns outside the scope of existing criticism. He coined the term 'distant reading' to describe this approach. The main intent in his two books, *Distant Reading* and *Graphs, Maps and Trees,* is to overcome the tension between these different methodologies. Moretti poses a number of fragmented theoretical claims, supported by a set of hand drawn graphs. The hypotheses were collected through comparative reading of numerous works of literary history or manual sifting through thousands of texts, inferring larger patterns or underlying developments. He has faced criticism for issues related to reproducibility, transparency, euro-centric eclecticism, and methodological gaps in reasoning.[61] Many of his texts aim to reconstruct world literature and trace the evolution of literary movements and genres as driven by economic needs and desires, with adaptations influenced by consumptive patterns. Moretti presents a duality of evolutionary processes that involve increasing differentiation and adjustment to market demands. Additionally, he discusses a center and periphery dynamic of colonial dissemination, using the symbols of waves for bottom-up processes and trees for top-down processes.[62] However, he acknowledges that these concepts do not provide enough variation to be a stable exclusive frame of reference, but rather serve as an

[60]Ramsay, *Reading Machines*, 79.
[61]Eg. Goodlad, "A Study in Distant Reading," 497; See also Underwood, "A Genealogy of Distant Reading," 7f.
[62]See Moretti, *Distant Reading*, 43–62.

incentive to build new narratives and perspectives.[63] Furthermore, the author's idealized concept of 'the great unread'—a vast collection of texts waiting to be rediscovered in archives—as an alternative method for evaluating literary history and recurring themes, does not align with the current state of preserved, digitized, and accessible human knowledge.

In response to Moretti, a core group of researchers at the Stanford Literary Lab has steadily expanded its reach. However, its methodology has shifted toward purely computational methods, while the original theoretical texts and the alternative of comparative analysis of secondary literature on literary history have receded in its wake and in reference to it. Mark Algee-Hewitt, Matthew Jockers, and Ryan Heuser have made important contributions to literary analysis. In the last decade, several other scholars have also emerged, including David Mimno, Andrew Piper, Ted Underwood, Alexandra Schofield, Mark Olsen, and the ARTFL project at the University of Chicago, as well as John Unsworth.

Jockers provides a list of specific questions and approaches that demonstrate how Distant Reading can aid the field of Literary History going forward. These methods are particularly well-suited for quantitative analysis. Among the questions to consider are how individuals relate to a larger whole or to each other, quantitative timelines for regions or demographics, evolutionary processes within style, aggregations, and linkage of individuals within a cultural embedding, the waxing, and waning of themes, changes in tastes and preferences, correlations between discrete socio-demographic attributes or styles, genre or literary categories, and the development of schools of thought. It is also important to examine the effects of marginalization or canonization on literary features.[64] The upcoming section will further discuss how those questions align with the most prevalent research topics in the field; Chapter 7 will see these questions applied to this thesis' analysis.

In summary, Distant Reading has been a statistical literary practice since the 1960s, but only in the last 20 years has a consistent and shared body of research taken shape. Although many methodological questions remain and there is still a need to harmonize the discourse with orthodox philology to achieve mutual enrichment, several avenues of investigation and appropriate questions have already been identified. To discover latent structural patterns, make inferences about the spread of influence across a genre, and interpret socio-demographic

[63]Moretti, Franco. Graphs, Maps, Trees, 92.
[64]Jockers, *Macroanalysis*, 27.

attributes for their influence on the composition of Gothic Fiction texts, this thesis will now examine current examples of the technical implementation.

4.2 Digital Methods in Use

Various techniques have been developed and improved over the last few decades in order to achieve the goals of literary critics using computational means. The following section will further discuss the most promising methods that are applicable to the task at hand. The methods employed by proponents of Distant Reading depend heavily on the specific research question. Computational literary criticism shares its methodology with approaches in related fields. This is largely due to the closeness of the task to models of quantification that are inherent to existing domains of inquiry within these disciplines. For tasks of stylometric analysis, the proximity of research in computational linguistics brings with it the use of sentiment analysis, word-co-occurrence analysis, and frequency based metrics, such as term-frequency-inverted-document-frequency (TF-IDF) representations, or cosine similarity within a vector space representation. For tasks such as the mapping of actors, their attributes, and relationship to each other, tools for the purpose of relational extraction, named entity recognition, and network analysis come into play. A similar set of tools is used for tasks that focus on interaction patterns and the distribution of term usage under specific conditions. For the purpose of pattern recognition, many unsupervised methods come into play, such as various means of partition-based clustering and density-based clustering. To reduce complexity, many studies rely on principal component analysis (PCA), and for the purposes of text modeling, text classification, and collaborative filtering, Latent Dirichlet Allocation has become a staple tool. A more unconventional approach takes existing metadata as categories and investigates and interprets what textual features supervised machine learning algorithms will choose to discriminate between them in order to classify them as such.

As Kuhn mentions, in many of these textual applications, the dividing line between the analysis of thematic or stylistic elements of literary language depends on the textual elements taken into account: Nouns allow for the analysis of topics, verbs for actions and interactions, adjectives and adverbs for cases of polarity, emotions, and associations, while many syntactic elements are essential for stylometric analysis. The more parts of speech a model includes in addition to nouns, the closer it veers from an analysis of places and actors to their creators—the writers and their styles.[65]

[65]Kuhn, "Computerlinguistische Textanalyse in Der Literaturwissenschaft? Oder: »The Importance of Being Earnest« bei Quantitativen Untersuchungen," 17.

In order to facilitate these forms of analysis, techniques such as parts of speech tagging, and stemming or lemmatization are used to label the different elements of speech, determine which ones will be used for further applications, and strip them of flections in order to compare and aggregate base forms of terms with one another.

Topic Modeling

Topic modeling is a method used to identify latent thematic structures within extensive document sets. The most established variant, designated as Latent Dirichlet Allocation (LDA), is incorporated into numerous standard Natural Language Processing (NLP) libraries. It is an intuitively interpretable procedure that can offer insight into the contents of texts and the distribution of those contents throughout a collection of documents. There are several different implementations available, with the most widely used being a Java-based software package called Mallet, written by Andrew McCallum. However, the Gensim implementation,[66] as well as the R package topicmodels,[67] also have a large user base.

A multitude of applications in the humanities have sought to integrate this technique with diverse NLP methodologies, with the objective of contrasting the outcomes or embedding them within a contextual framework that encompasses successive readings of the results of data transformations. This approach is intended to facilitate a deeper understanding of the composition of extensive textual bodies. Slingerland et al. investigated mind-body concepts and the associations of specific terms in early Chinese literature using hierarchical clustering, word collocation, and topic modeling. The collocation was corrected to adapt the association between terms and their synonyms for their frequencies through an agglomerative use of adjusted t-scores and mutual information, in order to extract more significant semantic relationships. All three methods provide results in agreement with one another; the mind-body image is judged as monistic. Clustering allowed for a clearer, more divisible picture, and topic modeling emphasized a stronger context.[68]

Jockers and Mimno analyzed the distribution of topics within a corpus of 4456 works through topic modeling. They ensured the robustness of their selection by employing metadata, extensive

[66]Řehůřek and Sojka, "Software Framework for Topic Modelling with Large Corpora."
[67]Grün and Hornik, "Topicmodels."
[68]Slingerland et al., "The Distant Reading of Religious Texts."

bootstrapping, other permutation methods, and hypothesis testing. They made use of the Stanford Named Entity Recognition package (NER) to exclude all characters and person names. To make for a precise term pre-selection, they used part-of-speech-tagging, a nearest-shrunken-centroid classifier and to narrowed down the selection to nouns. They found a gender coded skew in the distribution of topics, then visualized and plotted the results to investigate them more closely.[69] Their approach is particularly well-suited to the subject matter, as it exemplifies methods of chaining methods like NER to enhance the quality of downstream results.

Schöch investigated a set of 750 French plays to identify shifts in the distribution of topics over time. The relationship between topics and dramatic subgenres was examined using heat-maps, word-clouds, and time-series data. They further applied Principal Component Analysis (PCA) to discriminate between genres by topics and visualized the commonalities.[70] PCA, a method of dimensionality reduction, has been a staple tool in machine learning as well as Distant Reading, due to its robustness and ease of use. Given a set of variables in a dateset, it attempts to reduce the number of axes of variation to a minimum, resulting in a condensed display of the analyzed pattern. This condensed display can be easily visualized on a graph, providing researchers with a clear and concise representation of the data. This method has been applied in further studies to analyze shifts in thematic popularity in eighteenth-century American newspapers,[71] or as a means of generating features to improve classification tasks.[72] These investigations underscore the potential for genre analysis through the integration of LDA with time series data and the application of PCA to differentiate between various textual groupings. Sections 7.2 and 7.3 will seek to follow this example.

Matt Erlin chose to investigate a collection of 154 German novels written between 1731 and 1864 to examine the prevailing thematic groupings and relations between the texts.[73] He used topic modeling and, following common practice, segmented the texts into sections of 1000 words in order to gain more precise information about prevalent elements. Additionally, this allowed him to being able to group and compare recurrent, analogous stretches of text more easily. He used Mallet for this. Erlin iteratively settled on 100 topics, an amount that aligns with the consensus of the research community— on the contrary Wilkerson and Casas determined to have

[69]Jockers and Mimno, "Significant Themes in 19th-Century Literature."
[70]Schöch, "Topic Modeling Genre."
[71]Newman and Block, "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper."
[72]Hettinger et al., "Genre Classification on German Novels."
[73]Erlin, "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864."

found an optimum at 50.[74] Additionally, Erlin created a network between text chunks by linking works that share at least one-thousand-word passage on the same topic with a participation strength of 20% or higher. The strongest indicator of similarity, shared authorship, is a common observation in the research community. Further analysis was conducted by additional networks via the removal of edge weight to increase clarity. This revealed complex gender groupings, genre associations, and reflected interpersonal relationships between authors. Erlin's illustration of the search for compositional elements of a genre and the utilization of these elements as input for a network of influence within the genre provides the framework for the investigations presented in section 7.4.

Although it is an intuitively interpretable procedure its optimization still poses difficulties even two decades after its introduction. Due to its widespread use, there have been numerous attempts to make the procedure more robust and approachable. Ramage et al. utilized LDA to detect patterns and trends in the social sciences. They introduced their own adaptable implementation, the Stanford Topic Modeling Toolbox, which enables practitioners to make quick and easy visualizations to inductively steer parameter choice. This improves accessibility and trust in the results.[75] Tangherlini and Leonard utilized a sample-based sub-topic approach to reinforce their hyperparameter selection when approaching topic modeling.[76] Wilkerson and Casas took a similar, but more computationally intensive approach. They created 17 models using the same dataset and consolidated the 850 topics into those with a cosine similarity surpassing a certain threshold.[77] Riddell analyzed the topic shifts in newspapers as trends changed throughout time, investigating the difficulties of automatic parameter optimization.[78] This is a reoccurring issue that many other authors have found eclectic means of addressing. Cheng et al.[79] compared the results of optimization-based performance changes for topic modeling with standardized human evaluation methods. The study participants found that task-based human interpretation drastically outperforms the likelihood-based measures. Mimno et al. replicated these findings, highlighting the potential for improving topic coherence and quality with the help of co-occurrence statistics.[80] El-Assady et al. developed a public modular visual online toolkit as part of their VisArgue project for the iterative optimization task of topic model training. The toolkit

[74]Wilkerson and Casas, "Large-Scale Computerized Text Analysis in Political Science," 538.
[75]Ramage et al., "Topic Modeling for the Social Sciences."
[76]Tangherlini and Leonard, "Trawling in the Sea of the Great Unread."
[77]Wilkerson and Casas, "Large-Scale Computerized Text Analysis in Political Science."
[78]Riddell, "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models."
[79]Chang et al., "Reading Tea Leaves."
[80]Mimno et al., "Optimizing Semantic Coherence in Topic Models."

includes visualizations for topic matching, topic summarization, parameter distribution analysis, and document relevance feedback to enable users to make controlled adaptations.[81] These efforts collectively underscore the necessity for measures that reinforce the quality of results topic models present and document the extent of this search. As Section 6.2 will demonstrate, the latest endeavors to fortify this validity have employed vector space representations and diverse metrics to bridge this gap. Nevertheless, the pursuit of optimized methodologies for this class of models persists.

Topic modeling has become a fundamental tool for investigating the composition of content and its evolution over time. It is widely used for the investigation of correlations between attributes of metadata and specific thematic textual features, as well as input data for further study of interaction patterns, thematic similarities, and influence between texts. Despite the challenges in optimizing it, there are several mitigation strategies and best practices available. Erlin's setup is particularly interesting, as it combines topic modeling with a similarity-based network analysis. This study's base premise will be the focus of investigation in this thesis, but it will be aided by some of the methods Schöch and Jockers applied.

Classification and Clustering

Many pattern detection projects have also made use of supervised learning, particularly logistic regression[82] or random forests,[83] due to their ease of interpretability. The features of classification can be used to provide answers about the discriminating qualities of the texts in question. This section seeks to elaborate on the aforementioned methods by elucidating how the exploratory data analysis and PCA set forth by Schöch and Jockers can be augmented by incorporating the attributes of a model as an additional source of input. Alternatively, these attributes can serve as the foundation for supplementary close reading endeavors.

Broadwell and Mimnos analyzed the categorization of tropes for subgenres in folklore classification by exclusively consulting places and actors for their correlations.[84] Ted Underwood's studies on changes within literary genres over time are another prominent example.

[81]El-Assady et al., "Progressive Learning of Topic Modeling Parameters."

[82]See Kessler, Nunberg, and Schutze, "Automatic Detection of Text Genre."

[83]See Saccenti and Tenori, "Multivariate Modeling of the Collaboration between Luigi Illica and Giuseppe Giacosa for the Librettos of Three Operas by Giacomo Puccini."

[84]*Broadwell, Mimno et al., "The Tell-Tale Hat."*

He examined the reliability of classifying texts correctly based on their expressions at different points in time, comparing texts of different genres or those labeled as belonging to the same category by different sources.[85] Underwood questioned many of Moretti's assumptions, as well as the tendency for gradual development of clearer conventions rather than diversification within detective, Gothic, and science fiction. For his analysis of Gothic Fiction, he considered a corpus from the Stanford Literary Lab and Library of Congress genre tabs, as well as a list of relevant authors provided in Punter and Byron's *The Gothic.*[86] He concluded that genres are a mutable set of relations between works that resemble each other and link to one another in varying degrees. Genres may be a mixed category with different life cycles and levels of textual coherence. Interestingly, the analysis of the blurry resemblances focused heavily on prepositions, punctuation, and conjunctions, features which are usually disregarded or filtered out in such investigations. His findings contradict not only premises set out by Moretti, but also popular genre studies set out in Chapter 2. Both studies draw heavily on the unearthed data points for their interpretation.

Underwood generally favored logistic regression in his book *Distant Horizons*,[87] which features many classification use cases. In contrast with the objective of forecasting an outcome, he regards the outcome as a predetermined certainty and examines the inferences his models have drawn in classifying these texts into distinct categories. He covered the attribution of prestige and general reception of a given text, as well as an intricate study on the use of gender-coded language, gender characterizations, and differences in the language use among genders.

In his book *Macroanalysis,*[88] Jockers covered a broad range of techniques. He explored Irish American fiction through data analysis and the visualizations of the chronological distribution of socio-demographic attribute counts from metadata. He also plotted time series of specific word occurrences within different subsections of the dataset and interpreted the developments. Another analysis addressed decision boundaries, hierarchical clustering, and text classification by genre. For this, the texts were split into ten segments each to ensure more precise differentiation between textual features. Again, like Underwood, many of his classification features focused on conjunctions and terms that would normally be removed as stop words, thus showing more stylistic differentiation in the following analysis. This is an approach that,

[85]Underwood, "The Life Cycles of Genres."; Underwood, *Distant Horizons - Digital Evidence and Literary Change*, 34–67.
[86]Punter and Byron, The Gothic, xi-xii.
[87]See Underwood, *Distant Horizons - Digital Evidence and Literary Change*.
[88]See Jockers, *Macroanalysis*.

according to Kuhn, would bring him closer in line with authorship investigation. Further analysis addressed decision boundaries, hierarchical clustering, and text classification by genre. The corpus was divided into ten segments to improve text differentiation. He again applied a version of nearest centroid to evaluate the sharpness of separation between metadata categories in order to investigate the noise and to isolate a specific signal of features through the occurrence of paradigmatic high-frequency terms attributed to the given genres, tracing them over time—for all ten segments simultaneously.[89] Additional explorations dealt with the weight of metadata attributes on categorization and clustering and PCA to disambiguate sets of authors with particularly polarizing features through a selection of style defining prepositions. Jockers thus added his claim to the often mentioned insurmountable weight of the individual signals of an author's expression to clump together.[90] His experience might have been mitigated with a stronger limitation in the choice of words to process. In his application of the Mallet LDA implementation, he again used named entity recognition and part-of-speech tagging to preselect only nouns, and then plotted various distributions for sub-selections. His last study dealt with the influence among authors via the use of Euclidean distance as a measure of similarity of style, where he created network graphs to illustrate the relationships between a set of texts.[91] These investigations provide further evidence of the combination of multiple techniques, the utilisation of socio-demographic data or metadata as input, and the investigation of inferred distinctive textual attributes between groups of texts.

In his book *Enumerations,*[92] Andrew Piper conducted studies on the use of punctuation in poetry and its syntactic importance. He also conducted numerous word occurrence studies, used network analysis to examine shifts in character interactions within a text, and provided an in-depth interpretation and investigation of the roots and occurrences of the topics of a topic model based on a corpus of German novels published throughout a time span of 150 years.

Ultimately, these investigations highlight the potential for unsupervised methods like PCA or various forms of clustering to disambiguate categories of investigation, as well as the potential for machine learning models to be used in unusual ways. In some cases, the features used to arrive at a classification were more beneficial to the investigation than the decisions themselves. This is relevant to this thesis because, it highlights the adaptability of the methods in use in

[89]Jockers, 88.
[90]Jockers, 101f.
[91]Jockers, 161f.
[92]Piper, *Enumerations*.

distant reading and the possibility of using features of machine learning models as input for additional cycles of investigation, either exploratory data analysis or network analysis.

Network Analysis

Network analysis is a method used to study relationships between entities. The elements it deals in are interconnected nodes, as well as edges, the links connecting them either unilaterally or bilaterally. These elements can carry various attributes, although edges mostly carry numerical features for large quantitative, non-interactive investigations. These networks can be used to analyze and visualize the strength of connections, pathways, and quality of routes. They can also be used to study transmission and influence processes, detect communities, dynamics, and information flow, and provide an overview of connectivity patterns and the overall structure of a network. A number of different implementations are freely available. The software Gephi[93] is widely used, but offers only basic features. For a more detailed analysis of the importance of nodes to their network, the Python and R implementations called igraph[94] and a more modern package in Python called NetworkX[95] are essential.

Important metrics in this regard are density, clusteredness, robustness, and, most importantly, centrality and betweenness. While degree centrality is determined by the number of connections a node has, eigenvector centrality reflects the interactions that occur in a node's immediate neighborhood. Betweenness reflects the importance of a node in terms of how many shortest paths between other nodes must pass through it to reach its destination, indicating the position of a mediator.

While the most common application of network analysis in Distant Reading remains the creation of networks between individual characters within novels or plays, as demonstrated by Algee-Hewitt for Shakespeare's plays[96] or Piper for German novellas,[97] it has many other use cases.

Matt Erlin used network analysis to reflect on the shared occurrence of topic distributions among different texts and to interpret group dynamics in this way.[98] Jockers used it to plot the similarity

[93]Bastian, Heymann, and Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks."
[94]Csárdi et al., "Igraph for R."
[95]Hagberg, Schult, and Swart, "Exploring Network Structure, Dynamics, and Function Using NetworkX."
[96]Algee-Hewitt, "Distributed Character."
[97]Piper, *Enumerations*, 51.
[98]Erlin, "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864."

between authors' works and their influence on one another.[99] Meanwhile, Eder used it to evaluate the sharpness of separation within different clustering algorithms and the distance metrics they use, in addition to the shifts in the relationships experienced by authors.[100]

This section outlined that topic modeling is a common method for exploring common attributes of large bodies of text, that the features created when modeling text corpora are a rich source for exploratory data analysis, distinguishing between individual authors or genres, and are often used as subsequent input for further investigation. The use of network analysis to visualize and explore connections in literary texts was also detailed.

[99]Jockers, *Macroanalysis*, 161f.
[100]Eder, "Visualization in Stylometry."

5 Formalism, Structuralism, and Genre Studies:

5.1 Perspectives

The relationship between Distant Reading and Formalism is evident in many of the individual analyses that aim to demonstrate technical applications and contextualize them within a given tradition. However, most individual papers do not explore the origins of their vocabulary and leave these associations implicit. This thesis aims to fill that gap.

For the purpose of interpreting and analyzing works of Gothic fiction from the perspective of Distant Reading, a good starting point is to trace how researchers situate themselves within the realm of literary theory and, through references to canonical voices, draw comparisons and construct a coherent narrative.

If one looks at the most programmatic examples, the works of Franco Moretti, what one sees is frequent references to the works of Viktor Shklovsky,[101] the polarizing vanguard of the Formalists and head of the Society for the Study of Poetic Language. Mentions of both the movement and Shklovsky fill his texts, and he often invokes the language of these authors, speaking of 'formal mutation' and 'cultural selection.' In *Distant Reading,* he states that 'the conjunction of Formalism and literary history has been a constant (perhaps the constant) of my work'.[102] Indeed, the first publication of the Stanford Literary Lab, the computational literary studies research group he founded, was even titled *Quantitative Formalism*.[103]

Matthew Jockers invokes the prominent Russian Formalist Tynjanov's notion that literary studies requires a focus on interrelationships in order to do literary history justice, but that the inherent complexity of these relations demands a methodology that reflects the current abundance of available data and the technological advances. He urges for a new means of evidence-gathering and meaning-making beyond the anecdotal selection of choice texts.[104] He goes on to quote Alexander Veselovsky, a prominent literary historian who sought to redefine literary history and poetics by studying the evolution of artistic form and theorizing about the patterns underlying its development. Veselovsky's ideas on genre have been explored in Chapter 2, but will be explored

[101]See, eg. Moretti, Franco. Graphs, Maps, Trees, 4, 16, 17, 20, 63, 73, 91; Moretti, Franco. Distant Reading, 31, 71, 144.
[102]Moretti, Distant Reading, 89.
[103]Allison, Sarah et al., *Quantitative Formalism: An Experiment*.
[104]Jockers, *Macroanalysis*, 9.

further in Section 5.2. In essence, he focused on the parallel development of poetic forms based on comparable social and historical circumstances. Many of his studies were concerned with taxonomies of genres, their transmission into different contexts, and their independent formation in different places at different times.[105] He remained a prominent influence on Russian Formalism, but has been largely forgotten, with most of his texts remaining untranslated to this day. Jockers took interest in the descriptions of migratory patterns that Veselovsky set out to trace. These migratory patterns exist independent of individual authors and texts, and the creativity expressed in them. Veselovsky focused on elaborate systems of borrowing textual attributes and the idea of implicit influence without direct connection to another source. These concepts informed Jockers' application and subsequent interpretation of topic modeling of Irish Fiction.[106] Moreover, Jockers' syuzhet[107] package for studying the changes in sentiment over time in narratives takes the Russian word for 'plot' as its name.

In his monograph *Enumerations,* Andrew Piper invokes Roland Barthes, a prominent Structuralist. He compares the dissolution of language and a distributed, more schematic approach to meaning making to the vectorization of sentences and word count matrices.[108] In addition, in setting up an analysis of the parts of speech involved in characterizations in different genres of fiction, he refers to the grandfather of folklore studies, the Formalist, Vladimir Propp, who is famous for his schematic categorization of the narrative structure of folktales into thirty-one discrete plot functions.[109] According to Propp, these functions make up the basic building blocks of every magical folktale or fairy tale. Meanwhile, his characters were reduced to a simple typology based on the nature of the actions assigned to them by the narrative.[110]

Underwood's *Genealogy of Distant Reading*[111] points to the work of Janice Radway as one of the first quantitative works of literary criticism, which sought to analyze the Romance novel genre diagrammatically in a structure of symmetrical, binary oppositions, with measurements of polarity. The method was strongly influenced by the work of the Structuralist anthropologist Claude Lévi-Strauss.

---

[105]See eg. Zherebin, "Aleksandr Veselovskijs Konzept der Historischen Poetik und sein Echo im kulturwissenschaftlichen Diskurs Russlands und Deutschlands," 7.
[106]Jockers, *Macroanalysis*, 119.
[107]Jockers, Matthew, "Syuzhet.".https://github.com/mjockers/syuzhet.
[108]Piper, *Enumerations*, 13.
[109]Piper, 119.
[110]Cf. Propp, "Morphology of the Folktale."
[111]Underwood, "A Genealogy of Distant Reading," 18.

In summary, the burgeoning tradition of Distant Reading draws on the methodology of earlier schools of thought that focused on the atomization of literary works into interdependent building blocks, building blocks that perform reproducible, generalizable functions for entire types of texts at large. In order to provide the present thesis with a set of building blocks and tools for framing the object of study, applicable features of Formalism and Structuralism will be examined.

The Society for the Study of Poetic Language, the core group of Formalists, was established to redefine literary criticism by creating a set of tools based solely on intrinsic points of reference, namely the formal and inherent elements of the texts in question. This was in response to the predominance of historical approaches, such as the strong influence of Veselovsky, and approaches stemming from the social sciences and other fields of the humanities in literary criticism at the turn of the 19th century. Proving the autonomy of the aesthetic function within various forms of literary writing was a core project. This tradition was strongly influenced by developments in linguistics and frequently borrowed its language and ways of framing human expression.[112] This section will provide context to their work and single out individual perspectives important for the work at hand.

Saussure established an essential differentiation within a linguistic sign, which is composed of a signifier (verbal and written signs) and a signified (the represented concept). The represented concept is distinct from the material conditions that constitute the inherent idea. Reality exists independently of language, but our understanding of it is shaped by mental representations and concepts of sound that interact with each other through discourse. According to Saussure, there is a distinction between langue, which is the ideal system of language, and parole, which is the actual use of language in individual utterances.[113]

The tradition of Russian Formalism had several centers, including Moscow, Saint Petersburg, and later, Prague. The presence of cosmopolitan polyglot members like Roman Jacobson in Prague, as well as the first French translations and interpretations of seminal texts by Tzvetan Todorov and Julia Kristeva sparked the formation of French Structuralism. Structuralism integrated the culturally predominant psychoanalytic notions of its time and generalized the formal framework to interactive systems of arbitrary origin. It opened literary structures to

[112]Erlich, "XI. Literature and 'Life' - Formalist and Structuralist Views," 206.; cv. Roque, "Towards a Computational Approach to Literary Text Analysis," 101f.
[113]Alt, "Theorien literarischer Evolution bei Šklovskij, Tynjanov und Mukařovsky," 20.

references to cultural structures and focused on uncovering the ways in which meaning emerges from systems of signs, the interrelationships between linguistic elements, and their implications for the creation and decoding of cultural artifacts.[114]

A notion that most members of both Formalism and Structuralism, including Veselovsky, emphasize is the primacy of the system and the developments it undergoes, while the individual authors and their texts within it are merely expressions of trends and developments taken to form. An important contribution to the discipline as a whole was also the overcoming of the Aristotelian division between form and content as two separate spheres.[115] Images and themes, as well as stylistic and structural features, and any decomposed cultural artifacts, form the composite constituent systems composed of elements with specific functions that make up the text. Proponents of Formalism, such as Eichenbaum and Shklovsky, searched inductively for the generalizable rules of composition, attempting to identify the narrative techniques and devices at play.[116] Barthes would later employ the terms: dissection, fragmentation of elements, and the articulation of the rules of association.[117] As will become evident, analogies to biology are a very common occurrence, as is an aversion to subjectivity in interpretation and the intention to ground their study in the empirical sciences of their time. Eichenbaum summarizes it as follows:

> We were interested in the very process of evolution, in the very *dynamics* of literary form, insofar as it was possible to observe them in the facts of the past. For us, the central problem of the history of literature is the problem of evolution without personality—the study of literature as a *selfformed [sic] social phenomenon*. As a result, we found extremely significant both the question of the formation and changes of genres and the question of how "second-rate" and "popular" literature contributed to the formation of genres.[118]

One of the most widespread and universal contributions of Formalism is the relationship between 'fabula' and 'syuzhet.' Syuzhet represents the chronological structure, which includes all elements of arrangement and presentation, while fabula encompasses the world at large on which the plot is based, including all of its thematic elements and specific motifs. While the specific terms have found numerous reformulations, notably in French theory as 'discourse' and 'histoire,' the duality and basic definition have remained a constant.

Structuralism, on the other hand, had a tendency toward interdisciplinary work and worked on the premise that structures in different spheres would mutually illuminate each other. If unconscious patterns govern one area of social life that is structurally reducible, another area that

[114]Roque, "Towards a Computational Approach to Literary Text Analysis," 101.
[115]Gius and Jacke, "Are Computational Literary Studies Structuralist?," 3ff.
[116]Cf. Eichenbaum, "The Theory of the 'Formal Method'".
[117]Smithson, "Structuralism as a Method of Literary Criticism," 147f.
[118]Lemon, "Boris Eichenbaum," 83.

is also structurally reducible can benefit from its patterns and thus be enriched and explained this way. Many proponents have imagined literature as a dialect of language that can be analyzed in terms of Saussurean structural linguistics, which assumes a self-regulating system of language, or as a version of myth mediated by social and psychological approaches, with the intention of opening up the infrastructure of meaning inherent in the texts.[119]

> The goal of all structuralist activity, whether reflexive or poetic, is to reconstruct an 'object' in such a way as to manifest thereby the rules of functioning (the 'functions') of this object. Structure is therefore actually a simulacrum of the object but a directed, interested simulacrum, since the imitated object makes something appear which remained invisible or, if one prefers, unintelligible in the natural object.[120] […] Structuralism does not withdraw history from the world : it seeks to link to history not only certain contents […] but also certain forms, not only the material but also the intelligible, not only the ideological but also the esthetic [sic].[121]

An example of early analogue pseudo-quantitative applications is the work of Claude Lévi-Strauss in *The Structural Study of Myth.* Lévi-Strauss analyzed different versions of the Oedipus tradition by assembling segments of the narrative into a matrix along the axes of chronology and thematic relatedness. The author's method of pattern recognition involves arranging elements through analogy and association, drawing parallels and comparisons across cultures, times, and representations. This approach is crucial to the project of establishing elements of human culture that transcend individual regional representations of society and reveal some of its inner workings. His idea for the future was the atomization of texts into individual sentences and basic elements, in order to shift and align them between texts with even greater granularity in order to find points of analogy. The method he envisioned taking place on small strips of paper bears resemblance to bag-of-words models or term-document matrices. He argued for the use of statistical methods, similar to later proponents of Distant Reading. Moreover, he noted that the goals of Structuralism, if scaled to the necessary level of dimensionality and complexity, could only be achieved through sophisticated feats of mathematics.

> When we use several variants of the same myth for the same tribe or village, the frame of reference becomes three-dimensional and as soon as we try to enlarge the comparison, the number of dimensions required increases to such an extent that it appears quite impossible to handle them intuitively. […] [M]ulti-dimensional frames of reference cannot be ignored, or naively replaced by two- or three-dimensional ones. Indeed, progress in comparative mythology depends largely on the cooperation of mathematicians who would undertake to express in symbols multi-dimensional relations which cannot be handled otherwise.[122]

Tomashevsky's concepts of theme and motif as basic principles were particularly influential for many of the Structuralists, but Lévi-Strauss extended his conception to include the instance of

[119]Cf. Free, "Structuralism, Literature, and Tacit Knowledge," 65.

[120]Barthes, "The Imagination o f the Sign," 214f.

[121]Barthes, 219.

[122]Lévi-Strauss, "The Structural Study of Myth," 436.

the author as a bricoleur, an active mediator, designer, and assembler: 'The literary artist begins with three givens: motifs (ideas, events, or images), which belong to his experience; themes, which belong to the imagination's means of relating motifs into structures; and the motivating energy to create, which belongs to his psychology.'[123]

It has been established that Formalists aimed to atomize the components of fiction in order to create a repertoire of building blocks. They do not differentiate between form and content as two separate spheres, but between narration/story and content/plot as expressed by syuzhet and fabula. Furthermore, Structuralism extends the conception of these forms to include patterns that govern other areas of life that are structurally reducible.

Tomashevsky's repertoire has been particularly influential and will prove fruitful when applied to the task at hand. His primary opposition was between bound and free motifs. Bound motifs are fixed and constitute the essential plot. They have remained constant over time and across different movements and epochs; free motifs, on the other hand, are what constitute the unique contribution, style, theme, and leitmotifs of a given genre or period of literary production and are considered digressions. Dynamic motifs are transitional elements that change the state of a narrative and carry the development from one to the other. Static motifs describe a given situation, landscape, character, and provide attributes or embellishments that are not essential, but add verisimilitude and immersion.[124] Motifs that change the situation are dynamic motifs; those that do not are static. Free motifs are usually static, but not all static motifs are free.

Tomashevsky was of the opinion that a fabula contains a horizontal section of progression that decouples the essential developmental elements of exposition, initial situation, exciting force, peripeties, climax, and ending from their expressed representation within the story.[125] For him, a theme is organized into a moving set of motifs that aggregate and shift in and out of focus. While the fabula can be reconstructed independent of the reader, the syuzhet requires a specific point of view inherent in the given narrative. This progression through dynamic motifs is inherently driven by conflicts between characters and its resolution, or more generally by the collision of states with incongruous attributes. These state changes are accompanied by the suppression or emergence of traits as motifs, or their transference from one character to another.[126] In this

[123]Free, “Structuralism, Literature, and Tacit Knowledge,” 69.
[124]MacKenzie, “Narratology and Thematics,” 537.
[125]Garcia Landa, “The Structure of the Fabula (II),” 15.
[126]Garcia Landa, “The Structure of the Fabula (II),” 10ff.

respect, Tomashevsky mirrors Propp, reducing all characters to a living embodiment of a conglomerate of motifs along a thread of changing states.[127]

> The protagonist is by no means an essential part of the story. The story, as a system of motifs, may dispense entirely with him and his characteristics. The protagonist, rather, is the result of the formation of the story material into a plot. On the one hand, he is a means of stringing motifs together; and on the other, he embodies the motivation which connects the motifs.[128]

In other words, the temporal progression of narration, the foundational elements for its logical conclusion, and the base archetypical components are independent of any particular genre, style, or period of writing. These core universal functions are what he calls bound motifs. Those would be expected to be seen as an underlying baseline in the topics that do not define authorial style, or defining characteristics. Situations, landscapes, and characters do not progress the story but rather define the genre, while certain genre-defining themes or leitmotifs may still interact with the narration. This assessment adheres to common Distant Reading practices for studying themes. When preprocessing a corpus, it is standard practice to exclude proper names, temporal expressions, prepositions, and verbs that signify basic actions. It is widely regarded as essential to retain all nouns as they convey topical content.[129] However, there is a divide regarding the use of verbs, adjectives, and adverbs for additional generic content, atmosphere, scene, subjects, and expressivity.[130] In this thesis nouns, verbs, adjectives, and adverbs will be used, additional deliberations are part of Section 6.1.

The Structuralist Barthes also contributed one of the most complete inventories of tools for literary analysis. He aimed to break down texts into their underlying codes, analyze their connotations, and uncover the terms that hold significance in the text even when absent. These terms could be substituted to impact the interpretation of the text.[131] For him, functions are the smallest units of content, a 'Signified'—the content represented by the Signifier—that is very small but necessary for the operation of the entire 'system' of the text. They can be divided into three levels. Cardinal functions are points of risk in narratives, points of choice. Sequences from this point of view can be seen as 'threatened logical units', subject to risk or to choice.[132] Indicator functions or indices establish an atmosphere, a personality trait, a philosophy, a feeling, and often work implicitly or even non-verbally, hinting at concepts without dealing with the

[127]Garcia Landa, 19ff.

[128]Lemon, “Boris Tomashevsky,” 61.

[129]See eg. Jockers and Mimno, “Significant Themes in 19th-Century Literature,” 754.

[130]Cf. Schöch, “Topic Modeling Genre,” 5. ; Cf. Erlin, “The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864,” 61.

[131]Greenwood, *Structuralism and the Biblical Text*, 4.

[132]Francis, “Appropriating the Theory,” 54.

underlying operations of the narrative. For Barthes, popular fiction carries fewer indices and is more explicit in its themes. The core essential elements that define the progression and branching out of a plot are what he calls 'nuclei,' which unfold themselves one after another and interact. They fill the space between them, complement them, make events happen, and add to the discourse by enriching it with attributes and temporal or consequential framing. While all elements that do not influence the fabula, but only serve as embellishments in the chronological unfolding of the narrative, are named 'catalysts.' The third essential set of functions he calls informants. These provide specificity to the narrative, root it in a particular space and time, and lend it credibility and verisimilitude. Not all functions are purely of one type alone; they can overlap and blend together.[133] A consecutive connected chain of nuclei can be considered a sequence and can be bound together by a predetermined syntax; the study of these chains he likens to the work of Propp.[134]

Barthes argues that the notion that characters are optional for a fabula—a sentiment that has persisted since Aristotle in the work of most Formalists—has applied to less and less fiction since the advent of Romanticism, and is especially true for the works of Modernism. This is because in many of these texts characters are more pronounced and their personalities constitute much of their role within the narrative itself. He distinguishes between agents, for whom this old assessment still holds true, who act as a vehicle for functions and the unfolding of the plot, and persons, who are characters for the sake of personalization alone.[135] His methods of framing the role of characters within the action boil down to paradigmatic structures of opposition in how the agents participate in the action, action that can be filled by many actors mobilized within a given scene. The basic sets of oppositions are subject and object, giver and recipient, adjutant and opposer. These are located within articulated practices that can be subsumed within plots that deal with desire, communication, or struggle.[136]

In summary, Barthes' perspective provides noteworthy elements for the study at hand. His conception of the elements that compose the fabula and drive its progression as a narrative mirrors Tomashevsky's. Both conceive of a core set of functional elements that enable both plot and syuzhet to unfold their specific constellation in ways that are independent of archetypical, generic, or stylistic dealings. Style, genre, setting, and references to bodies of thought are

[133]Barthes, “An Introduction to the Structural Analysis of Narrative,” 247ff.
[134]Barthes, 252.
[135]Barthes, 256.
[136]Barthes, 258.

separated out into their own categories. Where they start to differ is in the specificity of their tools. Tomashevsky distinguishes between the core states and stages of a plot (bound motifs) and the actions and circumstances that drive their progression (dynamic motifs). Barthes makes a distinction between plot related (nuclei) and chronological actions (catalysts), but differentiates between concepts and moods (indices) and specificity that provides a contextual embedding (informants). His separation of agents and persons, based on their role as a narrative vessel or a portrayal of a personality, is also an important distinction. Considering the unique role Gothic fiction plays at the intersection of highbrow and lowbrow literature during the outset of modernity, it can be suggested that more popular texts within the corpus may contain fewer topics that serve as 'indices', while still maintaining the same core makeup. This may result in a reduction of atmospheric layers, multifaceted characters, philosophical references, and emotional depth. Furthermore, this suggest that there are fewer topics dealing with introspection and emotional expression in lowbrow texts outside Gothic romances.

Janina Jacke argues, from a contemporary linguistic perspective, that the analysis of a work's fabula can be attempted independently of its context, a point of debate for many generations of narratologists who drew heavily on the tradition of Structuralism. In making this assessment, she draws on the models of Propp and Greimas. The latter was heavily invested in categorizing words into hierarchical thematic clusters, which he organized in binary opposition to one another.[137] Jacke remains adamant that the syuzhet of a work is overly dependent on causality and chronology and presupposes the integration of too many elements that computational methods would seek to atomize.[138] Many critics of Distant Reading share this assessment. However, practitioners seek to address this concern by focusing their research on specific aspects of the syuzhet, such as the emotional trajectory or the distribution of terms signifying particular mental states. This line of research is greatly aided by computational linguistics and can also yield promising results with the necessary adjustments to research questions.[139]

Jacke argues that by allowing for interpretations based on common world knowledge, it is possible to infer the kind of pattern recognition on which many more motif-bound studies focus.[140] Given the findings of many of the detailed studies of Chapter 4, this thesis can only support this notion.

[137]Cf. Greenwood, Structuralism and the Biblical Text, 64.
[138]Cf. Jacke, "Is There a Context-Free Way of Understanding Texts?"
[139]An example for such approaches: Schmidt, "Plot Arceology."
[140]Jacke, 136ff.

In conclusion, this section has established a connection between Distant Reading and the traditions of Structuralism. It has explored several conceptions of the primary compositional elements of text, made assumptions about which elements can be expected to be found in upcoming topic modeling tasks, and provided tools for grappling with and differentiating between some of these elements. The section also provides backing for a fabula-based computational investigation.

5.2 Ties to Distant Reading

> Moretti's big questions are reminiscent of those that inspired the first generation of literary theory in Russia at the turn of the 20th century. Distant Reading is concerned with world literature, with the 'laws of literary evolution', with literary form, with the device as a minimal unit, with narrative universals. It is interesting to compare Moretti's approach to these questions with the foundational contributions [of] Alexander Veselovsky's Historical Poetics [...][141]

As discussed in Chapter 2, much of Moretti's writing on literary evolution is informed by Darwinian evolutionary biology and Immanuel Wallerstein's economic world-system theory, as well as Marxist historical materialism. His primary aim is to establish links between comparative literature, literary movements, and the development of trends and literary constants. He does this by examining various distribution systems that track transmission and ascribe a demand-driven logic to adaptations and changes. His theories revolve around a selective appropriation from the periphery, with a very hegemonic distribution of new forms from the centers of cultural production. He agrees with many of the Formalist assessments of the dispersal of plot to the periphery, while style remains more localized and is absorbed only in exemplary cases, creating hybrid texts. This perspective is at odds with that of Veselovsky, who saw literary forms as more cyclical in nature. Yet, he is closer to Moretti's views in attributing the origin of change in narratives to social and historical transformations.[142] Veselovsky, like the Formalists following in his wake, regularly employs fluid biologisms, framing the periphery as a space of counter-currents and unique associations that drive developments in times of ebb.[143]

Veselovsky's theory of the history of world literature is based on ritual repetition and re-contextualization. He emphasized the implicit normative elements of literary production over time, observing how these norms change, arise independently of one another, are transmitted, and what elements persist across time.[144] He also explored how different texts influence each other and shape styles and epochs. His work has sought to order and categorize the progression of different forms and stages of canonical texts, establishing relationships and a path of development among them. He employs arboreal metaphors to describe the common heritage, essential unity, and particularization, likening them to the 'leaves of a tree.' He was also the first comparative literary scholar to use the term 'motif' as a minimal formal unit of plot. The

[141]Merrill, "Distant Reading in Russia: Franco Moretti and the Formalist Tradition," 2.

[142]Merrill, 4.

[143]Cf. Etherington, "World Literature as a Speculative Literary Totality."

[144]See eg. Zherebin, "Aleksandr Veselovskijs Konzept der Historischen Poetik und sein Echo im kulturwissenschaftlichen Diskurs Russlands und Deutschlands," 7.

language he used was organic, implying convergence, correspondence, amalgamation, and fluid developments of associations rather than direct causality. Much of his focus was on the global evolution of genres on a scale of long-term development.[145] His studies dwelt on the development and common elements in song and myth prior to their clear divide. Through his decentralized perspective on popular and folk literature and the perceived continuous process of hybridization and blending through the negotiation of difference, he stands to offer a more intuition- and association-driven alternative to the other thinkers which have been discussed. Poetic elements of earlier epochs thus represent a sediment that allows for spontaneous reactivation, where social upheaval and times of strife may offer new opportunities for a revival.[146]

> These moments of rupture between epochs are when the synchronicity of multiple narrative systems allows for the greatest hybridization and the most active processes of reactivation of old strata as recycled tropes, motifs, and plots are renewed and acquire new meanings within this polysemous universe … the recycling and transformation of motifs and plots as a response to the changing religious and historical circumstances.[147]

While Shklovsky and Eichenbaum set out to define a set of tools and concepts for the systematization of knowledge, they also carried with them an inherent apprehension about the potential of freely applying these terms to periods and categories outside their respective studies. They understood literature as too volatile and unpredictable to account for all the changes in this way. An example of this volatility is embodied in Shklovsky's notion of defamiliarization, which captures the tendency of literature to enrich one's understanding of familiar events through the use of unusual perspectives and modes of framing to alienate the reader.[148]

For Shklovsky, stylistic forms go through a cycle, in which they first strive for dominance and canonization, only to eventually, through the passage of time and the effect of continuous modulation, become automated and degraded to auxiliary functions as elements no longer useful for defamiliarization. Changes in the course of literary history are, for him, a permanent process of erosion and regeneration of functionality. It requires a loss of commitment to a given stylistic form of literary genres in order to create new conventions. Parody is exemplary for the playful dynamics at work here.[149] The pace of these developments depends on the qualities of the

[145]Cf. Zherebin, "Aleksandr Veselovskijs Konzept der Historischen Poetik und sein Echo im kulturwissenschaftlichen Diskurs Russlands und Deutschlands."

[146]Holland, "Narrative Tradition on the Border," 440ff.

[147]Holland, "Narrative Tradition on the Border, 444f"

[148]Lvoff, "Distant Reading in Russian Formalism and Russian Formalism in Distant Reading," 32.

[149]Alt, "Theorien literarischer Evolution bei Šklovskij, Tynjanov und Mukařovsky," 9f.

elements themselves; their functions for the construction of a given work change over time, while the development of the literary system in which they are contained is ongoing.[150]

The most direct analogies to the methods of Distant Reading can be found in the work of a largely forgotten member of the movement. This scholar was more on the fringes of its circles, and his work has not received widespread recognition. Unfortunately, only fragments of his work have been translated and have reached a wider audience, and all of the practitioners of Distant Reading mentioned in this thesis have yet to acknowledge the parallels with his approach. Boris Yarkho employed the vocabulary of evolutionary biology—likening motifs and their occurrence in a given text to genotypes and phenotypes—and the statistical methodology of his day for the large-scale quantitative analysis of large corpora of text. He was a strong proponent of empirical validation of research findings in literature.

> [T]he researcher who has not conducted a quantitative account can never prove that an exception to the intuitively derived »basic forms« or essence, given by an opponent, is actually an exception or deviation. Secondly, the researcher's deductions do not prove that a certain feature constitutes the »essence« and predominates over all other features, for there is no measure of »qualitative« superiority and the quantitative weight of the remaining features are not known to him. … related artistic entities (authors, genres, schools, eras) differ from one another not so much by the presence of one or another feature, but more by the proportion of said feature; and these proportions cannot, of course, be derived in a deductive-intuitive way.[151]

While many of the Russian Formalists and some of the Structuralists used mathematical notation, and geometric interpretations of the works for the sake of visualizing events in the story-line was not uncommon, none of them except Yarkho ventured into the application of quantitative methods.[152] In one of his studies, *Speech Distribution in Five-Act Tragedy: Towards the Problem of Classicism and Romanticism*, Yarkho and his research group set out to identify the similarities and differences between 153 Romantic and Classical tragedies, and to compare them with an out-group of 50 canonical tragedies from outside this period. His main metrics revolve around measuring the ratio of the number of speakers to the number of characters per scene, along with the rates of entrances and exits. Yarkho spends long sections discussing the relevance of features for the subsequent comparison and his methods of calculation. He maps the attributes of the text of an individual authors onto the mean distribution of literary features within the epoch, assigning it its place on the continuum. From this he draws conclusions about the associations of different periods and how his results reflected common knowledge about

[150]Alt, 13.
[151]Yarkho, "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)," 14f.
[152]Lvoff, "Distant Reading in Russian Formalism and Russian Formalism in Distant Reading," 38.

them. He traces breaks in the transitions between these periods.[153] Yarkho's hypothesis about the wave pattern in the form of literary periods resembles some of Moretti's assumptions about the periodization of genre in *Distant Reading,* as well as assumptions about top-down stylistic developments that resemble the shape of waves.

For Yarkho, the development of literature was a dynamic and organic process, akin to cross-breeding of archetypal texts with others, resulting in the transmission of active or dormant traits in a generational process that defied clear chronological order. Traditions would dissolve and sources would agglutinate into new shapes. He likened the process of appropriation in the mimicry of certain traits independent of descent to hyperbolism. By comparison, in Moretti's conception, the actualization of traits involves the adaptation of potentialities inherent in a lineage of transmission and the possibility of representing attributes relevant to a given time and to the reality it strives to reflect and the influencing circumstances it entails. Regarding the question of whether economic demand plays a role, the two perspectives differ.[154]

The methods of word count and text segment shape comparisons Yarkho employed had, in his time, already been well established in classical philology for the study of poetry. He sought to formalize the emotional and ideological concept within texts by isolating statements from one another; the individual units were to be weighed with a so-called 'volume denominator' to evaluate their importance and contribution.[155] The categories of structuring and studying fiction involved a division into stylistics, phonetics, iconics, and the final composition of these elements. For the purposes of this study, of particular interest are his ideas on iconology, which deals with how images are linked together through thematic devices based on contrasts, similarities, or analogies, causal connections, or quantifying or specifying relations.[156] In his works on Germanic folklore, the author employed a method similar to the modern concept of topic modeling. He accomplished this by manually comparing texts, classifying terms within topics, and evaluating the importance of topics across different texts. Additionally, he created tables to compare the distribution of topic occurrences. In the untranslated *The Comedies and Tragedies of Corneille,* he compared the different modes and styles of speech, as well as the different actions between character roles, such as the occurrence of fraud, deception, bravery, and love across different genres. Through an immense amount of manual labor, he established

[153]cf. Yarkho, "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)." 57.
[154]Lvoff, "Distant Reading in Russian Formalism and Russian Formalism in Distant Reading," 45ff.
[155]Gasparov, "Boris Yarkho's Works on Literary Theory," 131ff.
[156]Yarkho, Boris, "The Elementary Foundations of Formal Analysis," 159.

stable points of reference between which individual works could be situated.[157] Like many of the contributors to Distant Reading who sought to crystallize the inherent concepts and patterns of genres or movements, for him,'[…] the ambition of any synthesis is to reduce the largest possible number of distinguishing characteristics to the smallest possible number of distinguishing concepts.'[158]

> The 'reality' of a sequence lies not in the 'natural' succession of the actions composing it but in the logic there exposed, risked and satisfied. […] [T]he origin of a sequence is not the observation of reality, but the need to vary and transcend the first form given man, namely repetition.[159]

It is abstractions like these that allow the search for recurring textual patterns to be framed as a search for the meaning inherent in text. With the argument that repetition serves as the core that reveals the structure of a work, and with the sentiment of the likes of Lévi-Strauss, that the humanities need to synchronize their methodology with the natural sciences, it can be argued that mathematical methods of pattern recognition based on correlations, co-occurrences, and repetitions have the potential to trace stable recurring sets of literary forms based on a given social framework in a given period of time.

Topic modeling can reflect such conceptions and facilitate comparative analyses between texts of a given genre, highlighting shifts or continuities in themes and perspectives. This is particularly valuable for the validation of canonical perspectives, as it provides evidence of persistence or evolution over time. On a more granular level, it can enable the identification of structural variation or deviation, and highlight texts or themes that challenge or subvert established norms and expectations, allowing for a more nuanced understanding. Network analysis, on the other hand, is uniquely suited to trace the relationships between the structures attributed to individual texts and authors and the context of contemporaries in which they are situated. It allows for hypotheses about the transmission and evolution of forms through influence and intertextual entanglements.

To summarize the specific points of tension identified here: For Moretti, the main source of innovation comes from the center of cultural production and is disseminated outward, while Veselovsky argues that the periphery is the main source of innovation, leading to integration inward. For Veselovsky, hybridization is a result of innovation at the periphery, and while he agrees with Hoppenstand that periods of social unrest lead to more hybrid forms, they differ in

[157]Gasparov, "Boris Yarkho's Works on Literary Theory," 141.
[158]Yarkho, "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)," 17.
[159]Barthes, "An Introduction to the Structural Analysis of Narrative," 271.

their assessment of the quality of these forms. Hoppenstand would attribute the rise of parody to a form that was firmly established and thoroughly rooted in a state of mass production. Shklovsky's assessment would support this, adding that, at this point, these forms had reached the end of their creative innovation and would continue to be eroded further until they fade out of use. At which state, according to Veselovsky, they would again be privy to unconsciously inform new conceptions of style and begin to be reappropriated in small fragments.

Tomashevsky, Barthes, and Hoppenstand have a narrower conception of genre fiction, claiming that it contains only a subset of the available functions. While for Hoppenstand the missing elements are of a social and political nature, for Tomashevsky and Barthes they are missing atmosphere, personality traits, references to philosophy, or elements related to the defining characteristics of a particular literary tradition of a particular time.

The evaluation of these and similar claims will be based on the results of computational analysis using the concepts of motif and function presented in this study. As demonstrated, analyzing the distribution of textual features of individual authors across a wider period or genre, as well as investigating the distribution of thematic devices across authors and times, are not novel techniques. However, what Yarkho accomplished through years of manual tracking is now achievable for current generations of digital humanists through computational implementations.

6 Dataset and Modeling:

For this thesis a selection of 182 Gothic fiction texts were collected, processed, and enriched with socio-demographic information. These texts were then used as input for a topic model, which was analyzed in order to investigate trends in the underlying texts. Furthermore, the distribution of these features was used to investigate relationships between texts and specific clusters of similarity between the authors. The following section will discuss the technical steps that preceded the analysis.

6.1 Dataset

The corpus was retrieved and normalized using the Python programming language and its standard data and text processing packages, along with gutenbergpy, a library for retrieving books from Project Gutenberg[160] using SQL. The texts were obtained through four different means. One set of texts was selected from an existing project conducted by Caroline Winter,[161] a Romanticist, and Eleanor Stribling, a programmer. Their project focused on the color space used in Gothic fiction texts. Another consulted body of research was a study by Ted Underwood, where he aimed to distinguish between Gothic fiction, science fiction, and crime fiction.[162] For his research, he consulted an index of important texts of Gothic fiction within *The Gothic* by Punter and Byron.[163] All other texts were obtained from Project Gutenberg directly in three rounds. First, a selection was made of all British or American texts Project Gutenberg labeled as horror stories, Gothic fiction, ghost stories, or supernatural fiction within the period of 1750 and 1910. Second, another selection was made of texts mentioned by Punter and Byron not yet retrieved by Ted Underwood. Third, several texts were manually downloaded from the Project Gutenberg website due to the low retrieval rate of gutenbergpy, which has not been in active development for several years.

After careful consideration, several texts present in the initial model run were removed before the final run. The results were unsatisfactory due to a skewed dataset. Additionally, many entries

[160] https://www.gutenberg.org/ [08.07.2024]
[161] https://github.com/CarolineWinter/gothic [08.07.2024]
[162] See Underwood, "The Life Cycles of Genres."
[163] Underwood, "The Life Cycles of Genres," 17; See Punter and Byron, *The Gothic*.

from the Underwood selection and some color corpus texts were manually re-evaluated and excluded from the corpus. This was because they were found to be too far removed from the core contents of the genre to be suitable for an assessment of the core features of Gothic Fiction. While those texts were influential to the genre, most of them were closer to historical fiction.

The final corpus has been cleaned, preprocessed, and chunked, resulting in 182 unique texts, 222 unique segments of 5000 word chunks of text with a reduced vocabulary, and 89 unique authors. Out of the segments, 110 entries (60.44%) were obtained from the color corpus, 47 entries (25.82%) were manually retrieved according to Punter and Byron, 18 entries (9.89%) were taken from relevant Project Gutenberg shelves, and 7 entries (3.85%) were sourced from Underwood's list.

Some manual cleanup was required to remove lengthy biographies, bibliographies, advertisements, scanners' notes on the license used, publisher subsections, or indexes. The color corpus contains highly relevant metadata, and through extensive manual research, the rest of the corpus has been enriched to provide the same features. These include 'date', 'gender', 'birth-date', 'nationality', and 'source'. Additionally, 'period', 'mode', 'genre', and 'role' are included for those taken from the color corpus.

The depreciated gutenbergpy proved incapable of parsing any kind of dash, quote, or empty line, and instead created mojibake in their place. This was solved by converting each character to its actual UTF-8 character. Then a regular expression was used to find all remaining byte-like sequences (e.g. \xe2) and allow detection and removal of these characters.

The corpus was preprocessed by removing any elements that are not relevant to the analysis or may impede its interpretability. The spaCy NLP package offers a language model that assigns syntactic types to individual words based on their context. It includes an English tokenizer, a part-of-speech tagger, a parser, stop word lists, and named entity recognition features. It is standard practice to exclude proper names, temporal expressions, and prepositions. Basic action verbs should also be excluded. However, it is widely regarded as essential to retain all nouns, as they convey topical content.[164] There is a divide regarding the use of verbs, adjectives, and adverbs for additional generic content, atmosphere, scene, subjects, and expressivity.[165] In this case, it was decided to retain all nouns, verbs, adjectives, and adverbs to maintain polarity and

[164]See eg. Jockers and Mimno, “Significant Themes in 19th-Century Literature,” 754.

[165]Cf. Schöch, “Topic Modeling Genre,” 5. ; Cf. Erlin, “The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864,” 61. ; Cf. Uglanova and Gius, “The Order of Things. A Study on Topic Modelling of Literary Texts.,” 59.

expressions of intent. As Kuhn states, the more parts of speech a model includes in addition to nouns, the closer it veers from an analysis of places and actors to their creators—the writers and their styles.[166] This issue which will be mitigated in a later step. The subsequent step involves the use of regular expressions to eliminate unwanted terms, punctuation, symbols, and spaces. Finally, the texts are tokenized into lists of individual words. Although it is common practice in unsupervised NLP studies to lemmatize or stem words, numerous studies have definitively proven that for topic modeling, the effects are negligible at best and reduce coherence at worst. In most cases, topic models are capable of conflating the vocabulary effectively as it is.[167] Additionally, it is good practice to curate the vocabulary used in a topic model. Removing highly infrequent terms can decrease unnecessary complexity of the model, and removing the most common ones increases interpretability. Opinions on the specifics vary. Mimno suggests removing terms that appear less than 5–10 times or in more than 5-10% of the documents.[168] Uglanova and Gius recommend a minimum frequency of 3 texts, with a maximum of 50% of the texts.[169] For this thesis, a minimum occurrence of 5 times was chosen and the top 5% of the most common words were removed, which significantly improved the coherence of the topics. In addition, it is common practice to segment texts into shorter segments, which are then processed individually.[170] This creates less ambiguity in the topics and reduces the influence of the author's style, thereby minimizing the impact of additional parts of speech in the process. The generally accepted range for segment length is between 500 and 5000 words.[171] After a trial and error process, a section length of 5000 words was chosen to provide a larger context window. To group these segments together later, a unique identifier was created using an incrementing number, sections of the author's name, and the text name. This identifier can be used to group or differentiate the sections, depending on the type of analysis.

Finally, the Gensim NLP package was used to create Word2vec embeddings, which are vector representations of the language space of a corpus of texts.[172] These representations are essential

[166]Kuhn, "Computerlinguistische Textanalyse in Der Literaturwissenschaft? Oder: »The Importance of Being Earnest« bei Quantitativen Untersuchungen," 17.

[167]Schofield, "Comparing Apples to Apple," 298. ; Mimno, "Computational Historiography," 5. ; Eads, "Separating the Wheat from the Chaff," 3.

[168]Mimno, "Computational Historiography," 3.

[169]Uglanova and Gius, "The Order of Things. A Study on Topic Modelling of Literary Texts," 59.

[170]Jockers and Mimno, "Significant Themes in 19th-Century Literature," 754. ; Uglanova and Gius, "The Order of Things. A Study on Topic Modelling of Literary Texts," 58.

[171]Schöch, "Topic Modeling Genre," 5. ; Erlin, "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864," 61. ; For a more general discussion see: Schofield, "Text Processing for the Effective Application of Latent Dirichlet Allocation," 26.

[172]See Mikolov et al., "Efficient Estimation of Word Representations in Vector Space."

for many language-based machine learning models as they enable the evaluation of the context of a word. These embeddings are generated to be used as input for the evaluation metrics in the modeling section, as well as input for one of the models itself, 'ETM: Topic Modeling in Embedding Spaces'.[173] While Word2Vec representations emphasize situating a given word within the context of a few sentences, another technique called GloVe focuses on situating a given word within a text segment as a whole. Unfortunately, many model architectures do not support it, so the created GloVe embeddings had to be discarded.

Dataset Composition

Roughly half of the texts are written by British authors, with an additional 20% consisting of Scottish, Irish, or Welsh texts. American texts make up just under 30% of the distribution, while other English-speaking sources are rare. Two-thirds of the documents have male authors. The distribution of publishing dates generally reflects waves of literary production, which somewhat aligns with Moretti's claim of two peaks in the genre's production at 1800 and 1830.[174] However, there are two additional peaks around 1770 and 1900. The corpus ends around the turn of the 20th century to avoid confusion with the emergence of weird fiction at the beginning of the century.

The color corpus is the only source of information on the period, text type, and role within the larger canon within the corpus. Two-thirds of the labeled texts fall within the Romantic label, and roughly one-third within the Victorian period. Their distribution reflects the shifts in publishing dates, with the former covering the two peaks of the late 18th to the early 19th century, and the latter accounting for the peak at around 1900. Around 50% of the labeled texts are novels, with an additional 25% being short stories and novellas. Poetry, drama, and other forms are not as well-represented. It is expected that the number of short stories is under-reported due to the inclusion of a large number of short story collections and the tendency of major contributors to the corpus, such as Poe, Machen, and Blackwood, to write exclusively in short fiction formats.

[173]See Dieng, Ruiz, and Blei, "Topic Modeling in Embedding Spaces."
[174]Cf. *Moretti, Franco. Graphs, Maps, Trees*, 15.

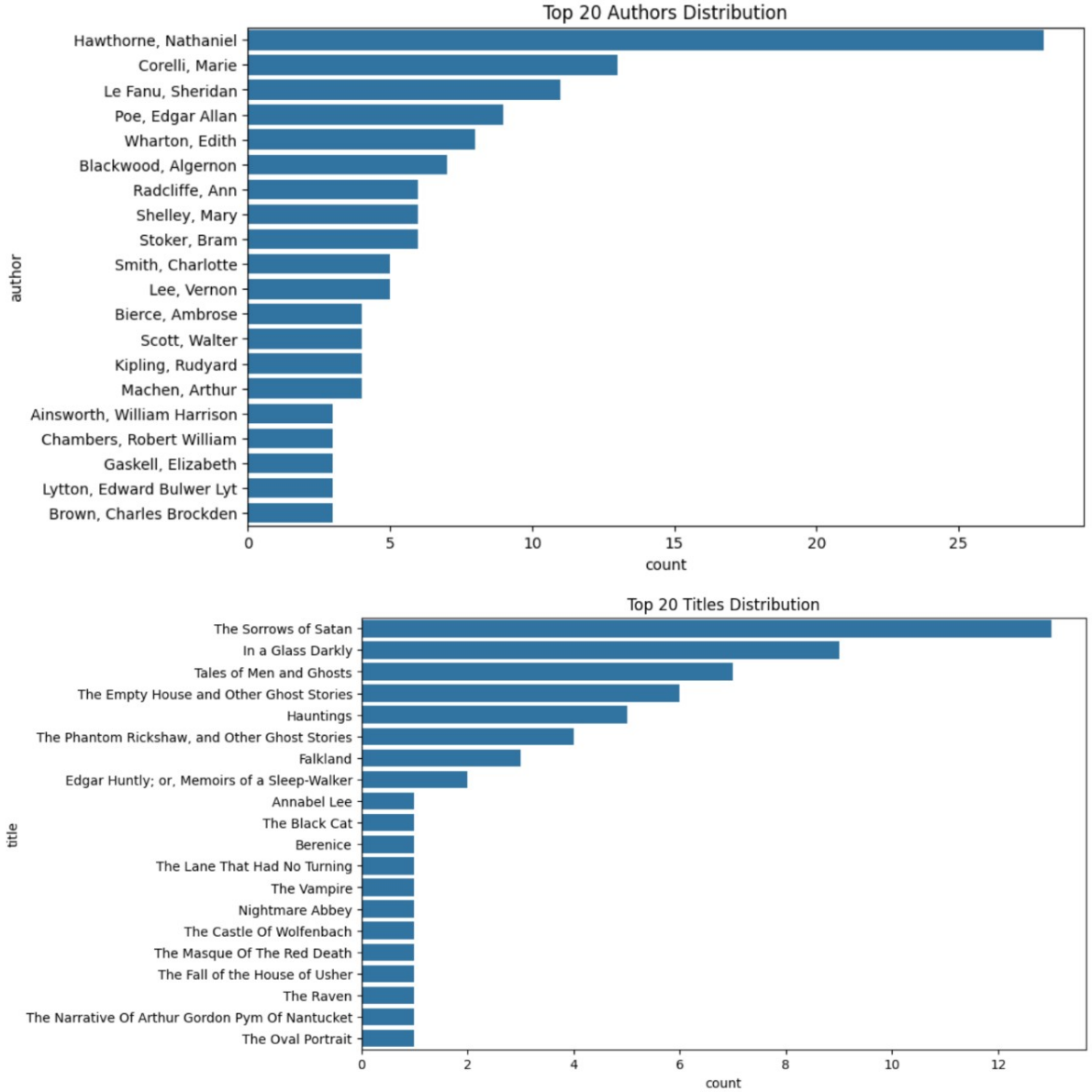


Figure 1: Texts and authors with the strongest contribution to the corpus

The corpus's main contributing authors are Nathaniel Hawthorne and Marie Corelli. Corelli, Blackwood, Machen, and Chambers are situated comparatively late within the development of the genre. The first contributors to the early Gothic, Mary Shelley, Ann Radcliffe, Ambrose Bierce, Charlotte Smith, and Charles Brockden Brown, appear only in seventh place. Between 1830 and 1870, the central publications were written by Nathaniel Hawthorne, Edgar Allan Poe, and Sheridan Le Fanu. Although female authors account for one third of the 20 most prevalent authors, only two of the 20 most prevalent texts are gothic romances.

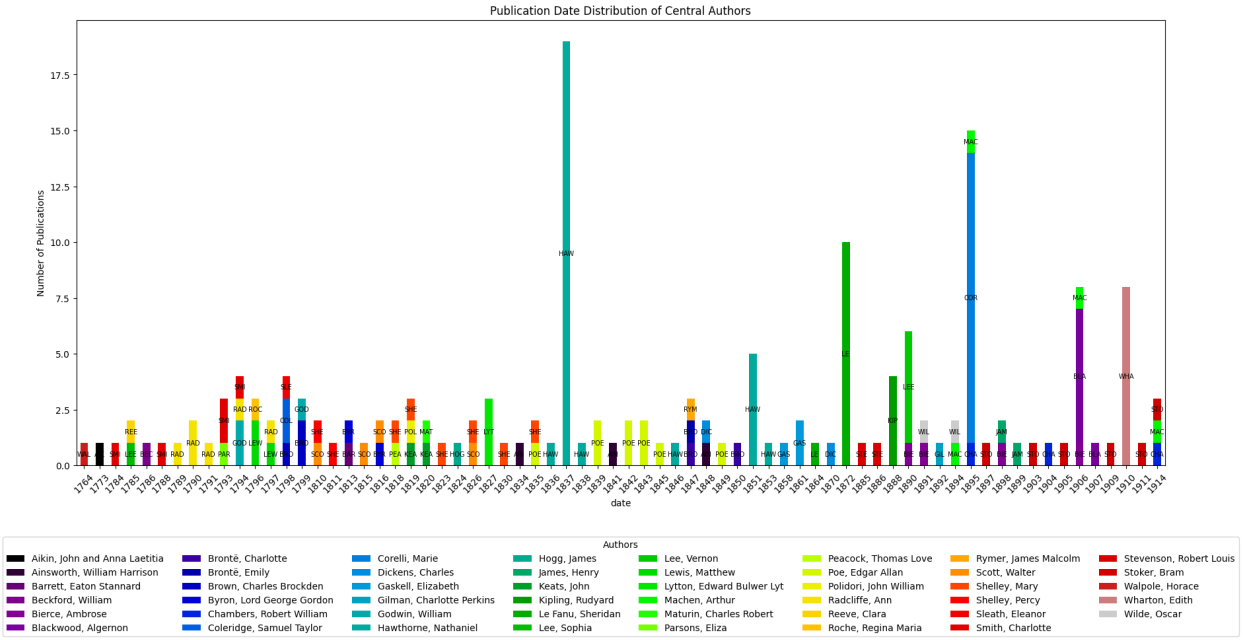


Figure 2: Distribution of publications by the most prevalent and central authors

Some of the largest contributing texts strongly influence this picture. In particular, these are the short story collections by Hawthorne in 1837, Le Fanu in 1872, Blackwood in 1906, and Wharton in 1910, as well as *The Sorrows of Satan* by Corelli in 1895. However, looking past these works reveals peaks in distribution during 1790, 1820-1830, and 1870-1900. This becomes more clear when the whole corpus is considered.

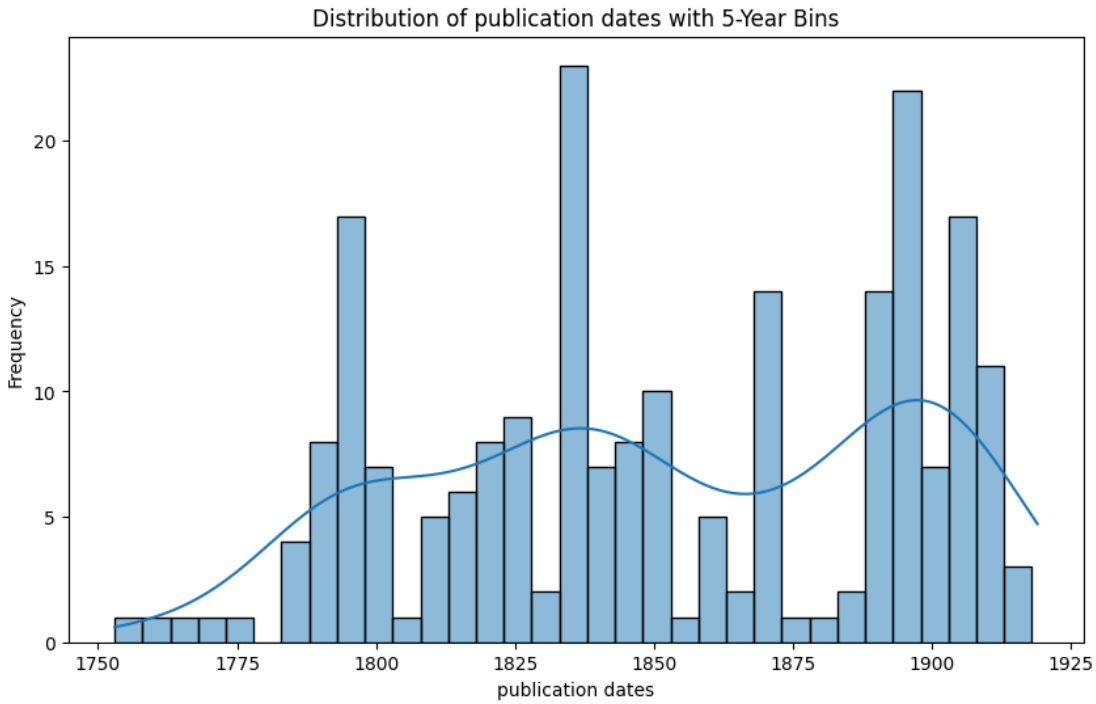


Figure 3: Distribution of all publication dates

6.2 Modeling

This section deals with the creation and optimization of topic models, as well as the enrichment of the topic distribution by document with the features taken from the preprocessing section. For the creation of these models the Python library *OCTIS: Optimizing and Comparing Topic Models is Simple!*[175] is used; the model is optimized with *Scikit-Optimize*[176] and enriched with *VADER (Valence Aware Dictionary and sEntiment Reasoner)*[177] sentiment scores on the texts.

Three different OCTIS models were optimized and trained:

- The standard latent dirichlet allocation (LDA)[178] in its Gensim[179] implementation.
- A contextualized topic model (CTM)[180] relying heavily on the use of the contextualized transformer model Bert.
- Topic Modeling in Embedding Spaces, (ETM),[181] a topic model that draws its topics and words from the embedding space of the original corpus using word2vec.[182]

Although each model has different parameters to optimize, they are all trained on the same corpus and optimized against the same metrics. The interaction with their implementations follows a general structure that is identical. In the subsequent section, only the results of LDA are discussed in depth. The CTM and ETM results are mainly used for comparison purposes to test the consistency of the topic distribution over time. More information on this can be found at the end of section 7.3.

The dataset variant used for model training is different from the one used for final analysis and interpretation only in that the former includes a split into training, test, and validation sets. As explained in the previous section, the dataset was preprocessed and divided into 5000-word sections. Terms that occurred less than 5 times overall were removed, as well as the top 5% most common words. As Schofield notes, 'Overly long documents and overly frequent words may

[175]Terragni et al., "OCTIS."
[176]https://scikit-optimize.github.io/ [08.07.2024]
[177]See Hutto and Gilbert, "VADER."
[178]For the original paper, see Blei, Ng, and Jordan, "Latent Dirichlet Allocation."
[179]https://radimrehurek.com/gensim/ [08.07.2024]
[180]Bianchi et al., "Cross-Lingual Contextualized Topic Models with Zero-Shot Learning."
[181]Dieng, Ruiz, and Blei, "Topic Modeling in Embedding Spaces."
[182]For a brief explanation of embeddings and word2vec see section 6.1.

slow convergence and produce poor topics; overly short documents and rare words may too quickly converge and produce similarly poor topics.'[183]

Since this analysis focuses mainly on the results of LDA, it is crucial to establish a basic understanding of how it works and to explain the modeling process choices in detail.

Topic modeling is a technique that helps to identify underlying themes or 'topics' in a collection of documents. The technique assumes that each document contains a mixture of different 'topics' and that a 'topic' can be understood as a collection of words that have higher probabilities of appearance in passages discussing the topic. The assumptions behind topic modeling are that documents are produced by discourses rather than authors, and that there is no way to infer the topics exactly because there are too many unknowns. To infer the topics, probabilistic topic modeling techniques like Latent Dirichlet Allocation (LDA) are used. LDA is a way of extrapolating backward from a collection of documents to infer the discourses ('topics') that could have generated them. The technique assigns words to topics randomly and then keeps improving the model to make the guess more internally consistent until the model reaches an equilibrium that is as consistent as the collection allows.[184]

Each topic is formally framed as a distribution over terms, while the eponymous Dirichlet allocation stems from the distribution that is used to draw the per-document topic distributions, which is called a Dirichlet distribution. In the generative process of word assignment and improvement for LDA, the result of the Dirichlet is used to allocate the words of the document to different topics.[185]

However, topic modeling has limitations because it requires subjective judgment calls that heavily influence the results. The resulting model is tailored to the researcher's preferences in ways that are difficult to replicate. As the resulting topics are a term distribution, researchers must use their discretion to assign labels to collections of terms. This aspect of the process cannot be quantified objectively. In addition, probabilistic methods require significant processing time and memory.

OCTIS is a relatively new library that aims to enhance existing state-of-the-art topic modeling solutions by integrating them into a unified pipeline for training, evaluation, and optimization.

[183]Schofield, “Text Processing for the Effective Application of Latent Dirichlet Allocation,” 26.

[184]See Blei, “Introduction to Probabilistic Topic Models.”; For a non-technical introduction into its application for literary use cases: Underwood, “Topic Modeling Made Just Simple Enough.”

[185]Blei, “Probabilistic Topic Models,” 78.

Although well-known implementations such as MALLET and Gensim provide access to standard coherence metrics, they do not, by default, cover diversity and significance metrics, embedding-based evaluation metrics, or neural topic models. In addition, their techniques for hyperparameter tuning using maximum likelihood estimation are particularly resource-intensive and time-consuming. OCTIS seeks to enhance the experience for the researcher, by offering more functionality in an easily accessible fashion.[186]

During the modeling process, the integrated Scikit-Optimize implementation of Bayesian probabilistic inference, which is a resource-intensive stochastic approach suitable for optimizing or evaluating difficult-to-optimize models treated as black-box inputs, was compared with the hyperopt[187] hyper-parameter optimization package. While Hyperopt's variant of a Tree-structured Parzen Estimator Approach (TPE)[188] proved noticeably less time-consuming and computationally expensive than the Bayesian approach at hand, the latter's seamless integration into the OCTIS framework's methods provided a user-friendly experience that outweighed Hyperopt's merits.

The concept of Bayesian optimization involves computing the objective function on the median of past runs with the same parameters, using previously evaluated model configurations to estimate the performance metric value and selecting a new promising configuration to evaluate.[189] In the case of LDA, there are three parameters: α and β for the two Dirichlet distributions, and the number of topics. The α parameter determines the distribution of topics over documents, while β determines the distribution of words over topics. A higher value results in a denser distribution, while a lower value results in a sparser one. In simpler terms, lower values result in a stronger fit into clear opposing categories, while higher values lead to less extreme and more centralized attributions. Terragni and Fersini suggest that effective ranges for both α and β fall between 10^{-3} and 10.[190] Regarding the number of topics, the majority of studies agree that a range of 50 to 100 topics is appropriate. Tangherlini et al. argue that maintaining a certain minimal threshold of topics is important to prevent nascent conceptual groups from being overlooked. They suggest that the risk of using too few topics is greater than the risk of using too

[186]Terragni et al., "OCTIS."

[187]See Bergstra, Yamins, and Cox, "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures."

[188]See Bergstra et al., "Algorithms for Hyper-Parameter Optimization."

[189]Terragni et al., "OCTIS," 265.

[190]Terragni and Fersini, "OCTIS 2.0." 4.

many, as with an excess, existing topics are merely cannibalized by new ones instead of creating uniform splits.[191]

As for the additional trained models, The Contextualized Topic Model (CTM)[192] and Topic Modeling in Embedding Spaces (ETM)[193] are two methods used for topic modeling. CTM uses a contextualized Bert transformer model, while ETM draws its topics and words from the embedding space of the original corpus. Both methods employ parameters commonly associated with neural networks, such as learning rate, activation function, and number of neurons. For the purpose of this thesis, the models were considered auxiliary, and their parameter ranges for optimization were chosen based solely on the recommendations provided by the package creators.[194]

Regarding evaluation metrics, coherence measures have become the standard for topic modeling studies. These measures assess the semantic coherence of words within a topic as a function of the relationship between the most prevalent words in a topic, which has been shown to be a good proxy for human interpretability.[195] Several metrics have been developed to measure coherence in natural language the most prominent of them being UMass, UCI, and NPMI.[196] These metrics have two general limitations. They are either based on the relatedness of word tokens, such as Cv, which uses normalized pointwise mutual information and cosine similarity, or on the probability distribution of the words denoting the topics. While the former cannot adapt for lexicographical differences, which might denote a similar meaning, the latter suffers in efficiency due to the high dimensionality of vocabulary.[197] An additional caveat for the investigation at hand is that the aforementioned metrics base their evaluation on external corpora that focus on modern everyday language and are not specific to the corpus they are applied to. To evaluate the fit of a model on literary texts, the distinctness from everyday language use proves a problem, specifically for a corpus focused on a genre that frequently resorts to archaic terminology. This is why only measures that optimize based on embeddings trained specifically on the language in use were considered, to make use of the fact that vector representations of the words appearing in similar contexts tend to be close to each other.

[191]Tangherlini and Leonard, "Trawling in the Sea of the Great Unread," 731f.
[192]Bianchi et al., "Cross-Lingual Contextualized Topic Models with Zero-Shot Learning."
[193]Dieng, Ruiz, and Blei, "Topic Modeling in Embedding Spaces."
[194]Terragni and Fersini, "OCTIS 2.0."
[195]See Lau, Newman, and Baldwin, "Machine Reading Tea Leaves."; Used eg. in: Uglanova and Gius, "The Order of Things. A Study on Topic Modelling of Literary Texts," 60.
[196]Cf. Rosner et al., "Evaluating Topic Coherence Measures."
[197]Terragni, Fersini, and Messina, "Word Embedding-Based Topic Similarity Measures," 33.

The coherence metric used in this study is Word Embedding-Based Centroid Similarity (WECS). This metric averages word embeddings for the top k words belonging to a topic, computes their centroid, and estimates their similarity with other topics to evaluate the contextual closeness of topic content. When used as an optimization technique, the score is averaged across all topics, and its improvement is tracked.[198]

Although coherence metrics are valuable for assessing the interpretability and quality of individual topics, they do not consider the breadth of information covered by topics as a whole. Diversity metrics, a less commonly used category, aim to measure the distinctiveness of different topics from one another and help to reduce unwanted overlap and similarity among them. In addition to WECS, Word Embeddings Inverted Rank-Biased Overlap Centroid has been tracked. Although the optimization did not specifically factor in this metric, it was the defining criterion for the final selection of parameters. Among the 10 most coherent parameter sets, the one with the highest diversity within its topics was chosen. The Word Embeddings Inverted Rank-Biased Overlap Centroid measures the significance of topics by comparing the centroid of word embeddings for a given topic to the centroids of other topics, computing inverted similarity. This method gives more weight to the most significant words within the topic.[199]

[198]See Terragni, Fersini, and Messina, "Word Embedding-Based Topic Similarity Measures." 36.
[199]Cf. Terragni, Fersini, and Messina, 37f.

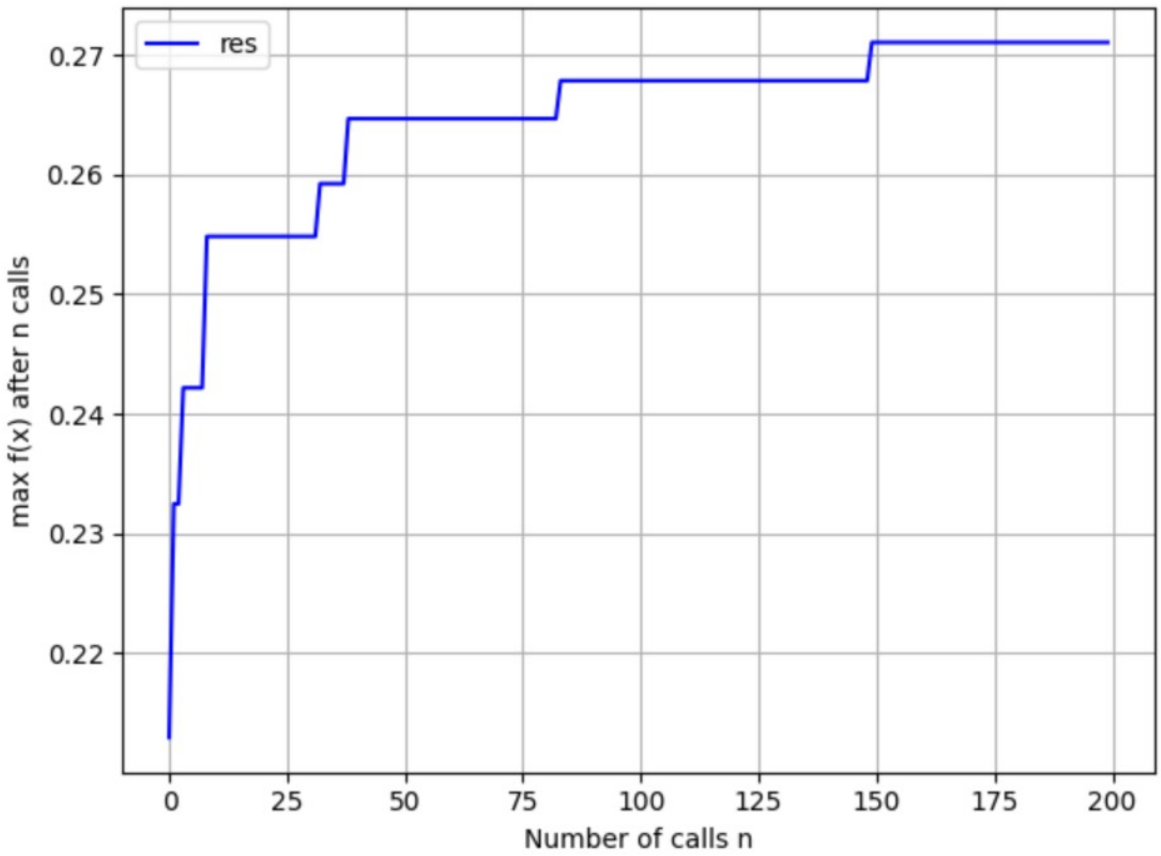


Figure 4: LDA optimization convergence

The following graph shows the optimization process for the LDA, where the initial parameters can be considered a particularly good fit, given the substantial initial performance. The starting point was 60 topics, with an α of 0.09 and a β of 5. As can be seen here, most of the coherence improvements could be achieved within the first 40 iterations, but the following 140 runs still brought a noticeable improvement. The optimization can be considered successful. It is interesting to note that the performance of the initial model run, which included texts that were subsequently excluded due to their incompatibility with the corpus, exhibited considerably higher WECS scores. Presumably, the embedding space benefited from the additional terminology in its attempts to compare and disambiguate categories.

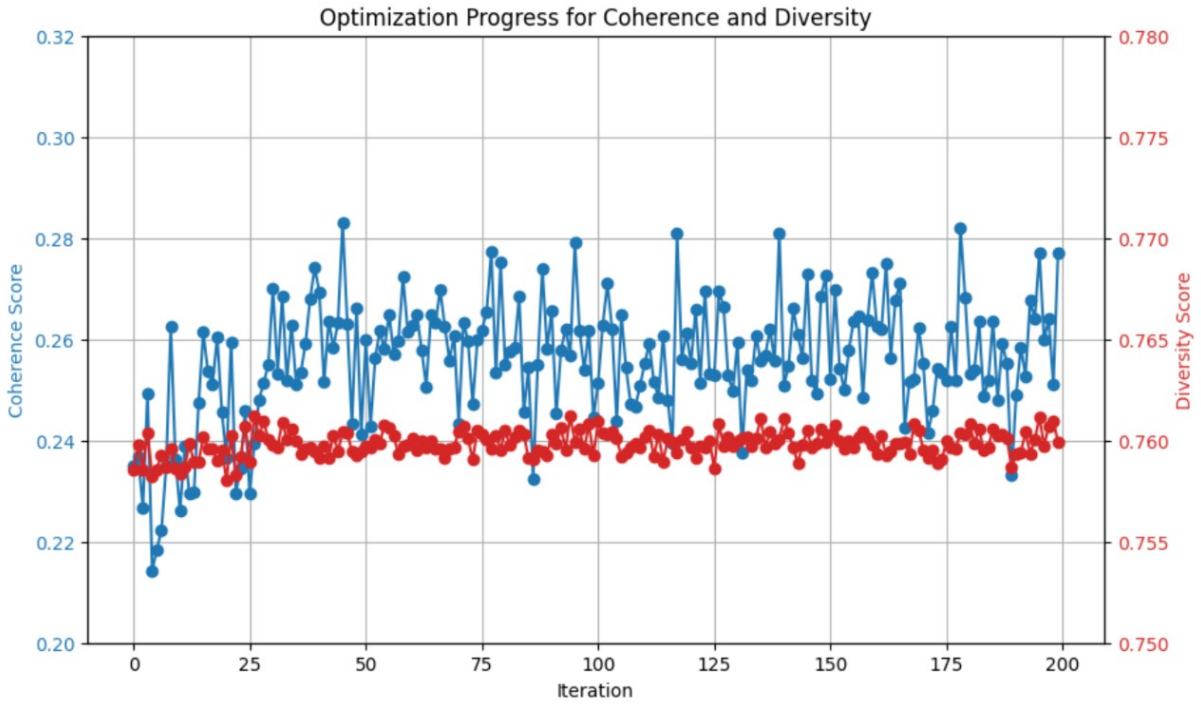


Figure 5: Coherence and Diversity metrics during the optimization

The final set of parameters, chosen by the aforementioned criteria, was 73 topics, with an α of 0.065, and a β of 2.942, which yielded a coherence score of 0.277 and a diversity score of 0.761.

In terms of lucidity and reusability, LDA was the clear winner. While ETM appeared to optimize better for the metrics in use, its practical application was found to be lacking.

The rows of the original dataset are sampled at random when passed into OCTIS in case the split parameter is set to 'True,' which is its default. This is a necessity to allow for hold-out data in the parameter optimization. In order to retrain the model with optimal parameters on the whole corpus, the parameter is set to 'False,' which leaves the order of the documents maintained. The separation into training, test and hold out or validation sets is necessary for hyperparameter tuning. By evaluating the performance of the model on the validation set, overfitting can be avoided. Overfitting would mean that the model captures noise or peculiarities in the training data instead of identifying the more general patterns. An overfitted model would be good at explaining the data it was trained on, but poor at generalizing to new, unseen data.

The distribution of the adherence to each individual topic for each text segment within the dataset set, which will henceforth be referred to as the topic distribution of the texts, has been joined with the original data frame to allow for comparisons across the individual categories to

which the texts adhere. In order to further enrich the data, sentiment scores for each text segment were generated and added as features to the data frame, using the de facto standard for sentiment analysis, VADER.[200]

This concludes all the enrichment of the processed texts that were used to interpret and explore their textual patterns. Any unspecified subsequent changes in Chapter 7 are the result of aggregations and comparative analysis. Other machine learning or NLP methods in use in this thesis include multidimensional scaling in section 7.1, PCA and K-Means in section 7.3, and network analysis in section 7.4. Their implementation details are discussed in their respective sections.

[200]See Hutto and Gilbert, "VADER."

7 Interpretation and Results:

This section explores the structural patterns found in 182 Gothic Fiction texts modeled using latent dirichlet allocation (LDA). The previous section outlined the process leading to these results. The full source code and documentation on the individual steps, as well as the instructions to replicate these results, can be found on GitHub.[201]

The underlying assumptions of topic modeling posit that documents can be understood as products of discourses rather than authors, and that there is no precise way to infer the topics, given the multitude of unknown variables. Consequently, a more flexible definition of the term 'topic' is employed. Topics are defined as any collection of shared contexts, themes, or styles. Among the questions particularly well suited to works of Distant Reading that Jockers proposes,[202] this thesis considers how individuals relate to a larger whole or to each other, aggregations and linkage of individuals within a cultural embedding, the waxing and waning of themes, correlations between discrete socio-demographic attributes or styles and genre or literary categories. For the pioneer of the approach, the Formalist Yarkho, this would fall within the realm of iconology, which deals with the manner in which images are linked together through the use of thematic devices based on contrasts, similarities, or analogies; causal connections; or quantifying or specifying relations.[203]

This chapter aims to apply Piper's conception of a circular flow in literary criticism to refresh, review, and potentially restructure the results of Close Reading. Close Reading analyzes the individual and its characteristics, integrating these beliefs into a model of the larger whole. The results of Distant Reading measure these concepts by attempting to model their form and specifications. The results of large scale quantitative analysis can then invite further close reading and interpretation of the patterns found.[204]

[201] https://github.com/f-klement/gothic-fiction-pattern-detection [08.07.2024]
[202] Jockers, *Macroanalysis*, 27.
[203] Yarkho, Boris, "The Elementary Foundations of Formal Analysis," 159.
[204] Piper, *Enumerations*, 10.

7.1 Topics

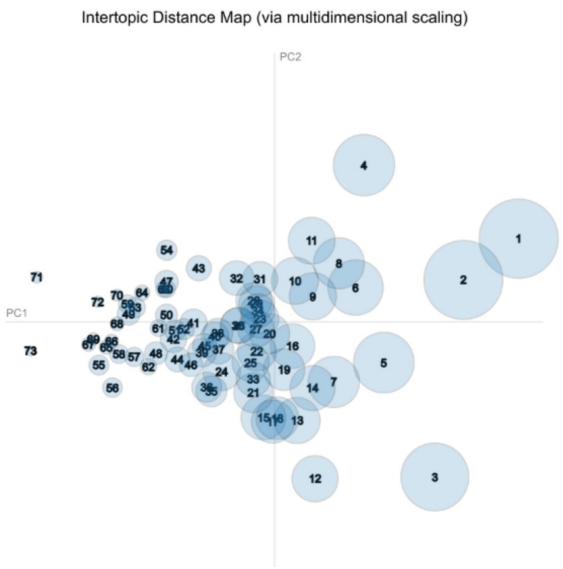


Figure 6: Intertopic Distance Map

The Python package pyLDAvis[205] provides a user-friendly approach to examining the keywords for each topic, their respective weights, and the interrelationships among and distances between the topics. It is a Python adaptation of the LDAvis package in R.[206] The tool extracts information from a fitted LDA topic model and enables interactive interpretation of the underlying topic space. Multidimensional scaling reduces the distribution of topic terms to a two-dimensional space, while still retaining both the importance of each topic within the corpus and their distance from one another. This is achieved using Jensen-Shannon Divergence[207] as the metric. Figure 6 shows that the first 19 topics cover the most ground individually, while subsequent topics are more closely packed together. After topic 50, topics once again share more individual ground, which is consistent with the overall importance of the topics. Although the topics at the core of the plot are widely used within the corpus, few of them reflect the unique contribution of an individual author or a specific peak in the genre's topical distribution. Instead, they represent a baseline of consistent associations. According to Tomashevsky, these would be considered bound motifs that convey the essentials of the plot without defining its characteristics. The topics that fall into this category include social altercations, travel, emotional dialogue, diplomacy, subterfuge, and moral conflicts. This supports Punter's assessment[208] that Gothic fiction

[205] https://github.com/bmabey/pyLDAvis [08.07.2024]
[206] See Sievert and Shirley, "LDAvis."
[207] See Sason, *Divergence Measures*.
[208] Cf. Punter, *The Literature of Terror - Volume 2*, 182.

anticipated the modern plot and introspective human interaction-driven novel. The content of these topics deals heavily with human interactions and emotional exchanges to drive the plot forward.

pyLDAvis includes two metrics for ranking the salience[209] and relevance[210] of terms within each topic. The most prominent members of each group were analyzed to synthesize underlying themes into coherent labels. The complete list of all 73 labels is provided in the appendix, and the lists of composing terms can be viewed via the HTML file of the third notebook. The following section will only cover individual topics or groups of topics that share common elements. In general, these topics can be classified into several main categories. This classification was derived inductively through a process of retroactively attempting to subsume all labels under overarching clusters.

- Emotional turmoil and psychological distress
- Physical violence and combat
- Social settings, diplomacy, and court
- Self-expression and frustration with society
- Myth, lore, and tales
- Forbidden truths and knowledge
- Adventure and exploration
- Ambition, greed, and regality
- Deceit and apprehension
- Science and reasoning
- Nature—woods, mountains, and harbors
- Religion and sacred rituals
- Monsters, demons, and undead
- Medieval settings, cities, and castles
- Dreams and illusions

Referring back to the themes of Gothic Fiction discussed in Chapter 3, the genre's diversity becomes immediately apparent. The model captures the social unrest, heightened emotions, and moral disputes that are inherent to the genre. The model also effectively represents many tangible elements, such as occult topics, religion, archaic medieval themes, monsters, violence, and scientific language. The thematic composition also supports the overlap with historical fiction, adventure novels, romance, and sensationalist fiction. The prevalence of concepts of logical reasoning suggests a connection to Mary Shelley's early science fiction and Edgar Allan Poe's early crime fiction, both of which are well-represented in the corpus. Other notable facets are the abundance of topics that deal in magical thinking, dream states, illusions, and other terms of dissociation and impaired judgment. The prevalence of these topics is in some ways attributable to the veil of disbelief that is often imposed on the final outcome of a narrative, as

[209]See Chuang, Manning, and Heer, “Termite.”
[210]As defined by Sievert and Shirley, “LDAvis.”

with Radcliffe, or inversely, a veil that has to be peeked behind and traversed in order to glimpse the sacred or forbidden, as with Machen. It is worth noting that many topics in the overall makeup deal with insanity, heightened emotions, excitement, distress, and vices. However, the theme of sexuality is surprisingly subdued and only hinted at. This may be due to implicit verbalization that is difficult to capture. Expressions of a philosophical nature are not uncommon in higher-numbered topics. Several topics explore individualism, conflicts with societal norms, and contemplation in natural settings, drawing strong parallels to classic Romantic themes.

As numerous topics share a significant thematic overlap, the ones most pertinent for the upcoming analysis have been grouped according to their thematic similarity. Below is a list of the topics in question, along with the numbers of topics they share overlapping motifs with and the associations they have in common.

Table 1: Core topics and those they share overlap with

Topic Nr.	Related Topics	Common Traits
1	11, 17, 70	Atmospheric: vast, archaic, refined
2	5, 10, 45	Emotions, Arousal, Fear, Secrecy
3	6, 5, 70	Individualism, Status, Excess
4	17, 70, 34	Myth and Crime
5	10, 38	Aggression and Emotion
6	8, 20	Nature & Reasoning
7	2, 19	Socializing, Courtship
8	9, 13	Faith, Knighthood, and Knowledge
9	8, 16, 65	Conviction and Adventure
10	5, 45	Intimacy and Conflict, Tragedy
11	17, 1, 34	Doom & Gloom
12	4, 65, 34	Home Invasion
13	4, 16, 19	Rituals, Dance, Magic
14	5, 65, 17	Conflict, Death
15	7, 5	Trickery and Science
16	9, 13	Desecrated Chapel
17	4, 11, 14	Undead, Judgment, and Grief
18	4, 17	Mystery and Adversity
19	10, 51, 13	Forlorn Carnival – Festivities and Decay
20	6, 8	Science and Nature
34	38, 12, 11, 4	Secrets, Mystery, Suspense
38	10, 17, 15	Psychology, Trauma, Secrets
45	10, 3, 2	Intimacy, Emotions, Identity
51	19	Disillusionment with Society
65	1, 5, 70	Battle, Atmosphere, Royals
70	4, 7, 65, 1	Myth, Wealth, Castles

7.2 Topic Distribution

The upcoming section will examine the distribution of topics over time and focus on a group of significant authors in more detail. The importance of these authors is determined by their prevalence within the corpus, which includes the 20 most frequently-occurring authors, as well as those considered core contributors to the genre by the romanticist Caroline Winter.

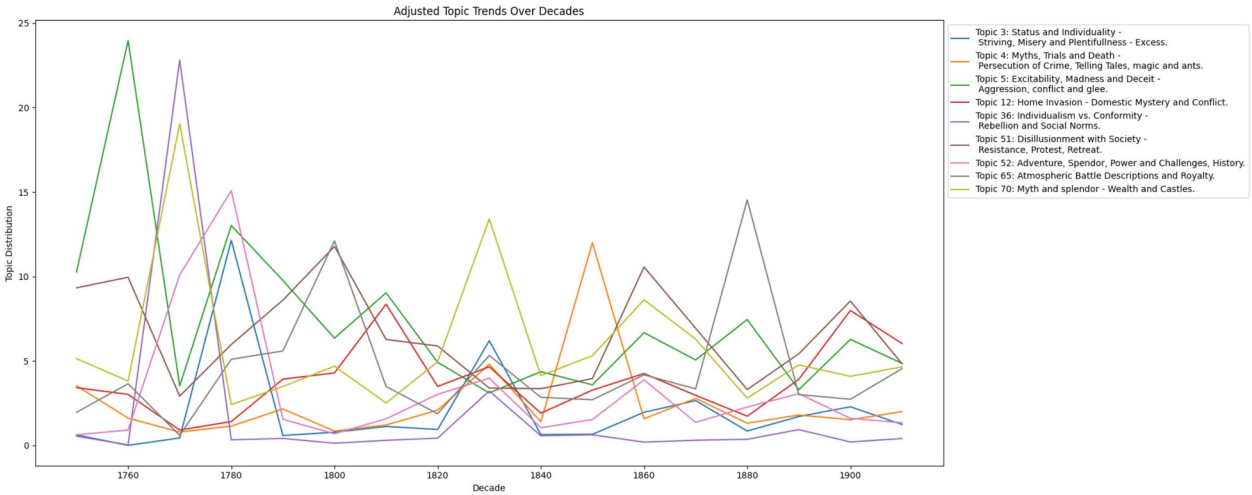


Figure 7: Adjusted Topic Trends over Decades

Figure 7 shows the progression of cumulative topic relevance per decade. The selected topics rise above an 8% share of any given decade's total at least once. Moretti argues that by 1820 the English book market would start to specialize and diversify, which could account for the shift from a large agglomeration of overlapping topics to more distinct peaks.[211] It is prudent to exercise caution in applying this observation here, given that the fundamental principles of the Gothic Fiction genre were established prior to 1820, it is equally plausible that the agglutination is the result of the concurrent exploration of clear themes and stylistic voices.

This selection provides a clear overview of the genre's main themes. The genre's most significant periods of textual representation occurred in the decades 1770, 1800, 1830, and 1900, and were accompanied by a rise in specific topics. Topics 3, 36, and 52 peaked before 1800 and then decreased in importance. Topic 3 and 36 are more strongly associated with societal discontent or unrest within power structures, whether through family intrigue, religious power struggles, or of individual quests against corrupt institutions. These notions are defining features for many of the early influential texts including Shelley's *Zastrozzi,* Godwin's *Caleb Williams,* or Brown's *Arthur Mervyn* and *Edgar Huntly,* as well as the works of Hogg. Topic 52 represents the early adventure

[211]Cf. *Moretti, Franco. Graphs, Maps, Trees*, 8.

and exploration-focused attitudes of wandering among decrepit ruins and the regal seats of old ruling houses. These attributes are commonly found in Walpole's works, but also in Radcliffe's Gothic romances, as well as the works of Sleath, Reeve, Roche, Parsons, and the gloomy Longsword by Leland.

Topic 5 exhibited a peak in the early stages, followed by a decline until 1830, after which it remained a constant undercurrent. In contrast, Topic 70 achieved prominence early on, declined in use, then later became a predominant influence in 1830, while remaining a stable baseline throughout. Meanwhile, Topics 51 and 65 reached decisive peaks in 1800 and secondary peaks in 1860 and 1880, respectively. In the case of topic 65, this peak can be attributed to Kipling. In the case of topic 51, it can be attributed to Corelli. Topic 4 also displays two peaks, one smaller in 1830 and a large spike around 1850. This is caused by Poe, Le Fanu, and O'Brien.

As will become evident in the subsequent network analysis, a combination of topics 5: 'Excitability, Madness and Deceit–Aggression, Conflict, and Glee', 51: 'Disillusionment with Society–Resistance, Protest, Retreat,' and 70: 'Myth and Splendor–Wealth and Castles' constitutes the primary defining characteristic of all the center texts within the early Gothic, although the specific composition varies. It can be argued that the peak of topic 70 in the 1830s can be attributed solely to the influence of Hawthorne on the corpus.

In summary, the texts of early Gothic fiction focus heavily on a certain aesthetic and locality, as well as the exploration of social issues. However, the 1830s to 1850s incorporate more mystery and crime elements, as well as elements of transgressed states of reality. The texts of the latter half of the century saw a rise in debate on social issues and a resurgence of historical portrayals, imbued with an underlying sense of supernatural unease. By 1820 many of the texts seem to diversify into clearer and more specialized perspectives.

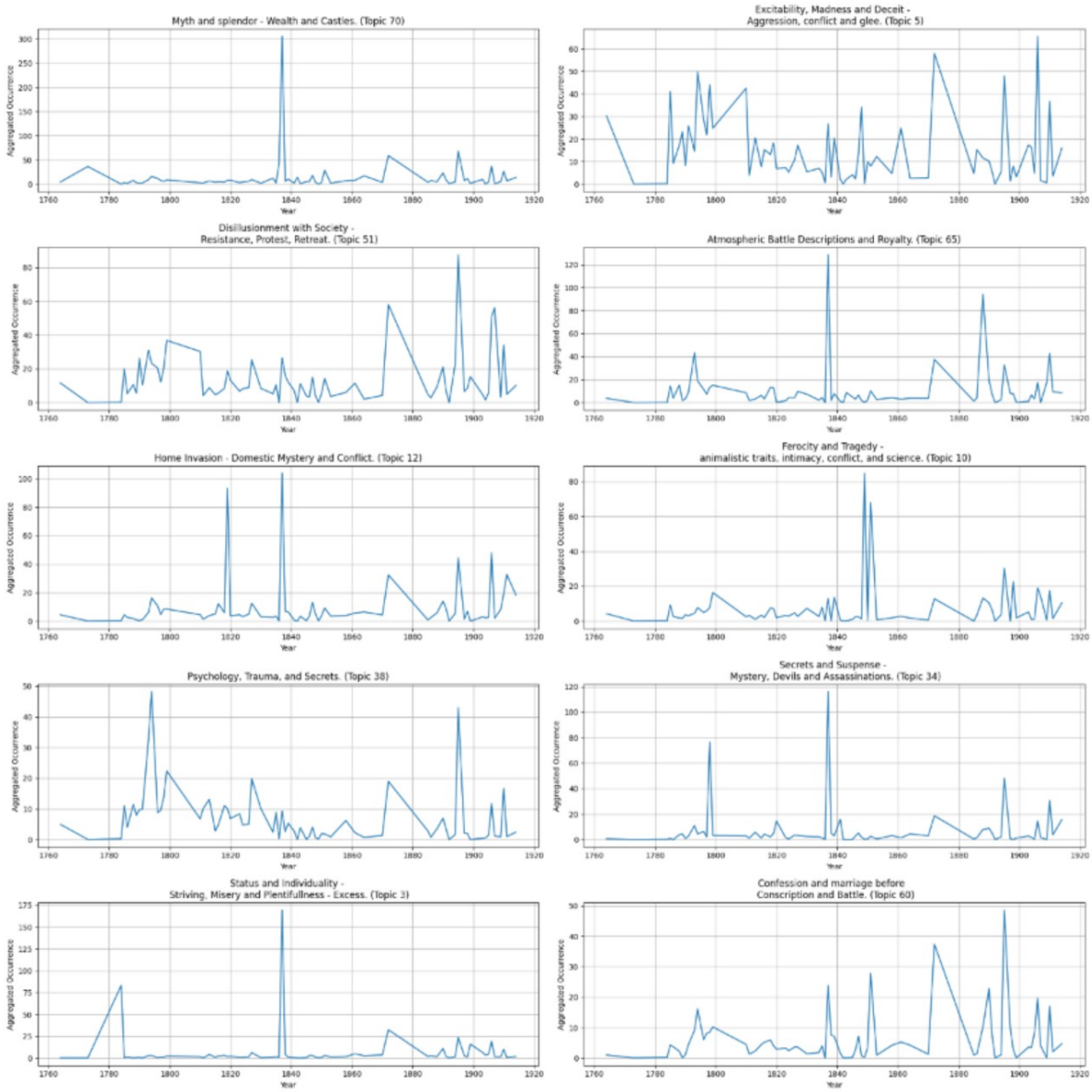


Figure 8: Topic fluctuation over time

A preliminary investigation of the most significant influences on trends over time and the compositional topics that have shaped them provides a framework for understanding the distribution and the voices that have contributed to its form. Figure 8 displays the yearly cumulative occurrence for the topics identified above. It shows an unusual spike in topics 3, 45, 34, 12, 65, and 70 around 1837. This spike is due to the significant influence that Hawthorne exerts on the corpus during this timeframe, and are composed of some of the topics most strongly associated with him. Although most of his texts do not heavily feature topic 34, 98% of 'Vision of the Fountain' is composed of this topic. This is fitting for a text focused on unraveling

the message conveyed in a dream state. The significant shift in influence during the 1870s was caused by Le Fanu, whose main contributing topics (60, 12, 51, 70, and 65) were heavily affected. This demonstrates the immense influence of his voice on the most prevalent topics of the corpus. Another spike in the strength of many of the observed topics is observed around 1898 due to Corelli and Machen. While Machen, like Le Fanu, has a very classical profile fitting the trend, Corelli's distribution of topics is highly unique, and deal with fighting, strife, and exploration.

In summary, the overall distribution demonstrates a pronounced focus on heightened emotions and the exploration of psychological themes and societal issues during the 1800s, followed by a resurgence of societal issues and familial topics towards the end of the 19th century. Additionally, there is a discernible peak in the representation of martial topics and threats to the sanctity of the private sphere between 1870 and 1890. The shape of the genre as a whole aligns with the image of early rapid growth and later specialization posited by Moretti for any fiction genres between 1770 and 1820.[212] The heavy focus on social issues and psychological distress could be attributed to the influence of the French Revolution and the Napoleonic Wars.[213] The vacuum of politics in the early novels was filled by highly political Romantic authors and Gothic fiction authors, who disguised the tremors of coming to grips with the Napoleonic Wars and industrialization with historical narratives. The heightened emotions and familial distress, common features of both Sensationalist fiction and Gothic fiction at this time, can be read as a deflection of the social into the private.

[212]Cf. Moretti, 8.

[213]See Punter, *The Literature of Terror - Volume 1*, 146.

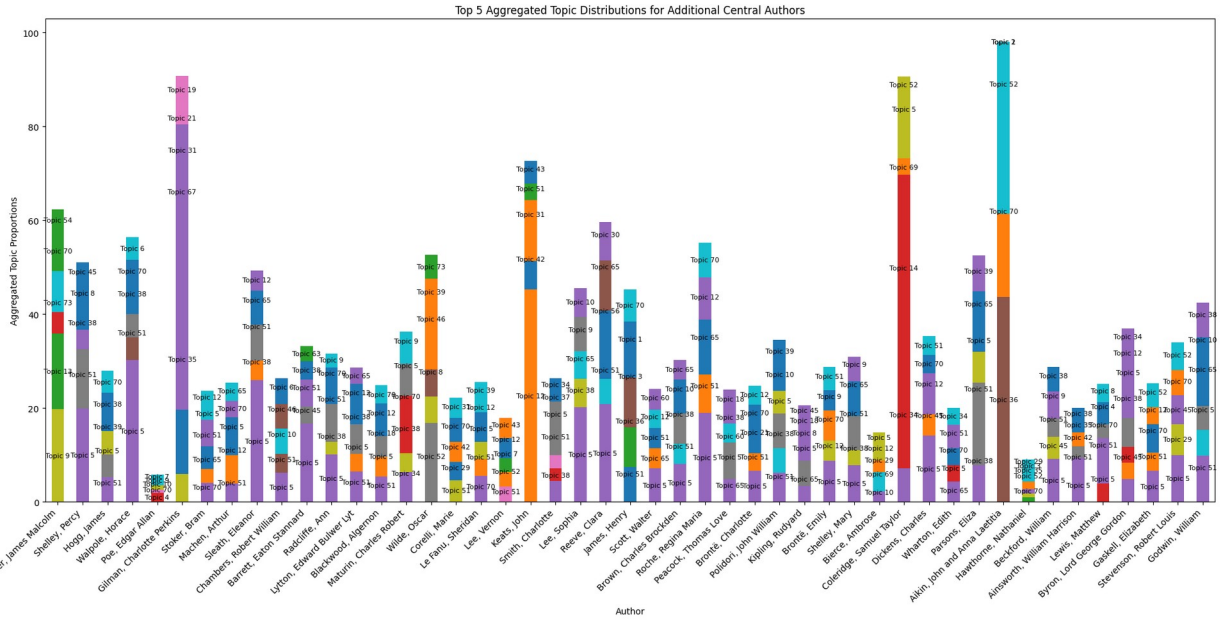


Figure 9: Top 5 topics for prominent authors

Figure 9 displays the cumulative occurrence across all texts of the top 5 topics for a selection of authors. The importance of these authors is determined by their prevalence within the corpus, which includes the 20 most frequently occurring authors, as well as those considered core contributors to the genre by the romanticist Caroline Winter. This comparison has been duplicated for both the 5 most important topics summarily and per median. A table comparing both is attached in the appendix.

A prominent focus here is that of Mathew Lewis on Topics 51: 'Disillusionment with Society,' 52: 'Adventure, Splendor, Power and Challenges, History,' 39: 'Quest for Meaning—Self-Discovery, Transformation,' and 12: 'Home Invasion—Domestic Mystery and Conflict'; for John Keats, 14: 'Conflict, Animosity, and Change—Emotional Changes, Death, and Construction' and 69: 'Seduction, Deception, Violence, Bureaucracy'; for Coleridge, 36: 'Individualism vs. Conformity—Rebellion and Social Norms'; for Aikin 52: 'Adventure, Splendor, Power and Challenges, History'; for Gilman 21: 'War, Punishment, and Exploration.' The aforementioned themes are characterized by a pervasive reflection of change, tumultuous interpersonal and social relationships, violence, processing a form of upheaval and striving.

Generally speaking, Topic 5: 'Excitability, Madness and Deceit' is a prevalent theme across the works of many authors, reinforcing the idea that Gothic literature frequently explores psychological instability and the darker aspects of human behavior. Walpole carries the highest values for topic 5, followed by Shelley ex aequo with Sophia Lee and Clara Reeve. Topic 51:

'Disillusionment with Society' appears significant for several authors as well, implying the importance of themes of resistance against societal norms and the exploration of characters who are at odds with their social context.

Topic 70: 'Myth and Splendor—Wealth and Castles' is prominent amongst authors like Charles Maturin, Arthur Machen, and Walpole, indicating a focus on grandeur, historical settings, and a reflection on the role of the past in shaping individual identities and social structures. Oscar Wilde's most prevalent topics, 52: 'Adventure, Splendor, Power and Challenges, History,' 39: 'Quest for Meaning—Self-Discovery, Transformation' and 34: 'Secrets and Suspense—Mystery, Devils and Assassinations,' highlight a drive for self-actualization and punishment delivered by society at large. Punter attributes an inner moral didactic exploration to Wilde as well.[214]

Eleanor Sleath, Eliza Parsons, Sophia Lee, and Clara Reeve have a significant presence in Topic 65: 'Atmospheric Battle Descriptions and Royalty', which reflects works that delve into grand conflicts and courtship, an association frequently found in both the early Gothic Romances and their later revival near the end of the 19th century. John Keats, Algernon Blackwood, and Bram Stoker have a considerable portion of their topic distribution dedicated to a combination of topics 5 and 12, suggesting a focus on personal turmoil and the encroachment of danger into personal spaces. This is consistent with both the depictions of hostile environments that invade the personal feeling of safety in Blackwood's The Willows, and Count Dracula as a looming threat that preys on people in their homes. Unsurprisingly, Stoker is the prime contributor to Topic 61: 'Vampires, Regality, Experiments, Festivities, and Sacrifice', befitting for his explorations of the boundaries of shifting social norms and the aristocratic Other for whom the experience of carnival as the state of exemption from societal restrictions takes on a different hue.

Edgar Allan Poe is a distinctive voice in Topic 28: 'Communion in Nature—Transformation, Relationships and Identity,' which resonates with Poe's themes of personal transformation, identity, and a profound examination of the development of inner turmoil experienced by characters. Poe's narratives frequently culminate in moments of epiphany or horror as his characters confront their own identities. In Poe's works, the environment in which a narrative is set is not merely a backdrop; it is an active participant in the narrative, influencing and reflecting the characters' mental and emotional journeys. Punter and Byron support this assessment.[215] Further topical associations focus on mystery, gloom, animal depictions and the undead.

[214]Cf. Punter and Byron, The Gothic, 173; Cf Punter, The Literature of Terror - Volume 2, 7.
[215]Cf. Punter and Byron, 156.

Hawthorne's works often grapple with the moral legacy of Puritanism, the past's weight on the present, and the ensuing questions of redemption from history, individual morality, and identity. His focus on isolation, obsessions, mental illness, guilt, and spacial storytelling related to his hometown of Salem, Massachusetts aligns well with topics 35 and 36, and to a lesser degree with topics 70 and 52.[216]

Machen's focus on the invasion of the domestic sphere, as hinted at by the heavy reliance on topic 12, stems from his interest in the vulnerability of personal space and the erosion of the boundaries between the safe and the profane. This theme often leads to a deep-seated unease, as the sanctity of home is breached by otherworldly forces, making the familiar uncanny. Machen's work could be seen as prefiguring the modern psychological horror genre, which includes works such as those of H. P. Lovecraft, Clive Barker or Thomas Ligotti, that frequently uses similar themes.[217]

Le Fanu's Gothic tales often revolve around psychological ambiguity, unreliable narrations, madness, deceit, characters at the grip of their own repressed self and frequent critique of social norms.[218] His stories, such as 'Carmilla', 'Uncle Silas', and those collected In a Glass Darkly, struggle with externalized aspects of their of psyche and the societal backlash for their enacting of predilections outside the norm. His topical makeup—5, 6, 11, 19, 70, 45—carries madness, longing, and a form of repulsive intimacy.

These associations all reflect a prevailing sense of unease and nervousness about identity formation or social dissolution, a defining aspect of modernist fiction that shaped the 20th centuries' adaptation of Romantic themes. However, they also exhibit a stronger focus on the processing of a shared history or the identification of fractures where it is lost.

Topical Groupings

The following is a list of dominant topical groupings and the authors that show a particularly strong association with one or more of the listed topics. These topical groupings are presented in order to offer a reading of Gothic fiction counter to the understandings of the genre put forth by Tomashevsky, Barthes, and Hoppenstand.

[216]Cf. Punter and Byron, The Gothic, 123f.
[217]Cf. Punter, *The Literature of Terror - Volume 2*, 22.
[218]Punter and Byron, The Gothic, 30 and 138.

Metaphysical and Philosophical Inquiry: Authors in this group explore a search for meaning in topics like 39: 'Quest for Meaning—Self-Discovery, Transformation' and 66: 'Hidden Knowledge, Learning and Secrets'. The relevant authors are Le Fanu, Shelley, Wilde, Coleridge, and Hogg.

Gothic Romanticism: This category includes authors whose works have strong associations with topics of Romanticism, often exploring the tension between desire and morality, freedom of expression, community, and sublime nature. Topics like Topic 28: 'Communion in Nature—Transformation, Relationships, and Identity', Topic 44: 'Companionship in Times of Trial and Distress,' and Topic 6: 'Nature and Reasoning—Creativity, Understanding, mixed with Fauna' are indicative of this category. The relevant authors are Poe, Kipling, Le Fanu, Hawthorne, Shelley, and Chambers.

Social and Political Commentary: These authors use Gothic elements to critique social and political structures. Topics that stand out include Topic 36: 'Individualism vs. Conformity—Rebellion and Social Norms' and Topic 51: 'Disillusionment with Society—Resistance, Protest, Retreat'. The relevant authors are Hawthorne, Brown, Lytton, Gaskell, Chambers, Ainsworth, Scott, Vernon Lee, Charlotte Smith, Stoker, Mary Shelley, Radcliffe, Wharton, Le Fanu, and Corelli.

Historical and Mythic Reconstruction: Works by these authors are characterized by a strong sense of history, archaic spacial settings, and the interweaving of myth within their narratives. Prominent topics are Topic 54: 'Medieval Cities, Castles, and Courtship' and Topic 70: 'Myth and Splendor—Wealth and Castles'. The relevant authors are Radcliffe, Hawthorne, Corelli, Wharton, Stoker, Vernon Lee, Scott, Machen, Ainsworth, and Gaskell.

Pioneers of the Psychological Thriller: This grouping is for authors who laid the groundwork for what would become the psychological thriller, focusing on the human mind's complexities and its vulnerabilities. Topics such as Topic 5: 'Excitability, Madness, and Deceit', Topic 38: 'Psychology, Trauma and Secrets' and 'Topic 44': 'Companionship in Times of Trial and Distress' are central to this grouping. The relevant authors are Le Fanu, Wharton, Blackwood, Radcliffe, Shelley, Stoker, Charlotte Smith, Bierce, Machen, and Chambers.

Conflict and Societal Restructure: These authors focus on the chaos and order of society, the collapse of old structures, and the struggle for new identities. Topics such as Topic 14: 'Conflict,

Animosity, and Change', Topic 37: 'Order and Chaos—Constrained Focus and Unchecked Emotions', and Topic 29: 'Bickering, Fighting, and Mountains' are highlighted by this grouping. The relevant authors are Bierce, Hawthorne, Marie Corelli, Radcliffe, and Charlotte Smith.

These groupings illustrate the extensive range of the genre in its intricacy. Gothic fiction serves as an anti-realist tradition and a conduit for Romantic writers to investigate questions of identity, a yearning for a past and mythical realm, as well as a reassessment of emotional drives. It is crucial to acknowledge the genre's pivotal role as a channel for the exploration of familial turmoil, the potential for societal change, and the philosophical musings on the trappings of Victorian morals. However, it is equally important to recognize the genre's capacity to facilitate a dialogue with many repressed aspects of the individual or collective psychological makeup. The concept of alienation, the animalistic, the instinctual sides of humanity, unconscious desires, and the release of repressed energies and antisocial fantasies were embedded within a context that allowed for the contrast with forces that reach for transcendence, and hope for human potential to persevere.

To reiterate, Tomashevsky, Barthes, and Hoppenstand have a narrower conception of genre fiction, claiming that it contains only a subset of the available functions. While for Hoppenstand genre literature is missing elements of a social and political nature, for Tomashevsky and Barthes genre literature lacks atmosphere, personality traits, references to philosophy, or elements related to the defining characteristics of a particular literary tradition of a particular time. This stands in contrast to the groupings put forward here, as a number of the authors participating in the societal, philosophical, or Romantic discourse are not largely acknowledged as highbrow poets. Neither Le Fanu, Chambers, Corelli, nor Machen are regarded as having contributed to the development of the discourse of their time in this regard. As Todorov notes, numerous writers of supernatural texts have projected personal and traumatic experiences into their work. The fantastic as a whole is defined by its lingering on challenging the border between the real and imaginary, thereby creating a space for association. Whether this lingering is about transgressing a social law or re-imagining the boundaries of what it means to be human or a functioning society in a changing environment, these boundaries of expression are made permeable, for attempts at transcendence, preservation, or dismantling. The Gothic served as a jester's license to explore themes that otherwise couldn't have been explored. Those are still not rightfully attributed as such due to the distorting veil of fiction, which hides overt commentary behind symbolism, metaphors or indirect language.

Figure 10: Main contributing authors to a given topic

Figure 10 examines the topics previously identified as essential to the corpus. For each topic, the top 5 contributions from a single author were summarized and examined.

Table 2: main contributors to core topics

Name	Attributes	Topic Numbers
H. James	Archaic Atmosphere	1
Wharton	Gloom and Longing, Blasphemy, Battles & Nature	11, 16, 65, 6
W. Scott	Gloom and Longing	11
Corelli	Emotions, Status, Convictions, Institutions, medieval, Mystery, Dances, Social Discontent	2, 3, 8, 15, 18,19, 51
Radcliffe	Emotions, Conflict, Madness, Social Discontent	2, 5, 14, 51
Poe	Gossip, Gloom, Undead, Mystery, Animals	4, 7, 9, 10, 17
Bierce	Ferocity & Tragedy	10
Le Fanu	Madness & Romanticism, Longing, repulsive Intimacy, archaic atmosphere	5, 6, 11, 19, 70, 45
Blackwood	Madness, Adventure, Conviction, Dreams and Mystery, Societal Discontent, Identity	5, 9, 18, 51, 45
Wilde	Conviction & Death	8
Keats	Home Invasion & Mystery and Conflict	12
V. Lee	Social Pleasantries and Scheming	7
Stoker	Home Invasion, Desecration, Dreams & Mystery, Castles & Myth	16, 12, 18, 70
Hawthorne	Home Invasion, Witchcraft, Status & Individuality, Deceit & Institutions, Mystery, Merriment	3, 12, 13, 15, 16, 18, 19 70, 65
Rymer	Rituals	13
Coleridge	Conflict, Emotions	14
Machen	Undead	17
La Spina	Festivities, Intimacy, and Disgust	19
Kipling	Battles & Royalty	65
Byron	Intimacy & Identity	45
C. Smith	Psychology and Trauma	38
M. Shelley	Psychological Trauma, Madness & Aggression, Trickery and Science, Science & Animalistic Violence	15, 10, 38, 5
Coleridge	Secrets and Demons	34

The same comparison offers a very similar picture when repeated at the level of individual texts, but some new associations could be gained. The underlying graph is not included here; it can be accessed via the notebook or the provided HTML.

Table 3: Additional authors of main contributing texts to core topics

Name	Attributes	Topic Numbers
Wharton	Frightful Dialogue, Myth and Trials, Chivalry & Faith, Rituals & Magic	2, 4, 8, 13
Walpole	Aggression & Madness	5
Kipling	Chivalry & Faith	8
M. Shelley	Gloom, Doom, and Longing	11
V. Lee	Undeath & Grief	17
Hawthorne	Myth, Wealth, and Castles has three entries by Hawthorne	70
T. Moore	Intimacy, Identity, and Emotions	45
Parson	Psychological Trauma	38
Lytton	Psychological Trauma	38

With regard to the authors previously explored in detail—Machen, Le Fanu, Blackwood, Hawthorne, Wilde, Stoker, and Poe—the overall associations confirm the expectations that were previously put forward. A similar pattern emerges when considering the writers of early female Gothic romance, including Radcliffe, Parson, Smith, and Lytton's homage to Goethe's Werther. The topic model also identified associations with war, violence, and death in Bierce's work, as well as the loneliness and longing inherent in Wharton's texts. Of particular interest here is the breadth of associations allotted to Mary Shelley, whose range of expression would define the content of several genres. She contributed to the genre through her philosophical investigations, the scientific and moral discourse of her time, and her deft unraveling of the meaning of humanity.[219]

As has become apparent in Tomashevsky's terminology, the modeling captured not only the base static motifs of the essential plot situations, characters, and landscapes as unique contributions, as well as a swath of dynamic motifs within the higher numbered topics that denote human interactions, but also many of the free motifs that make up the more intangible topics of personal convictions and philosophical or political leanings. In contrast to the prevailing opinion, these topics are distributed to a more diverse array of authors than is commonly assumed.

[219]Cf. eg. Punter, The Literature of Terror - Volume 1, 106.

7.3 Feature Distribution

The following section will examine the distribution of topics across genders, nationalities, and sentiment, in order to elucidate the particularities in style or topics and to discuss the implications arising from them or the reasons for a specific configuration within the corpus.

Gender

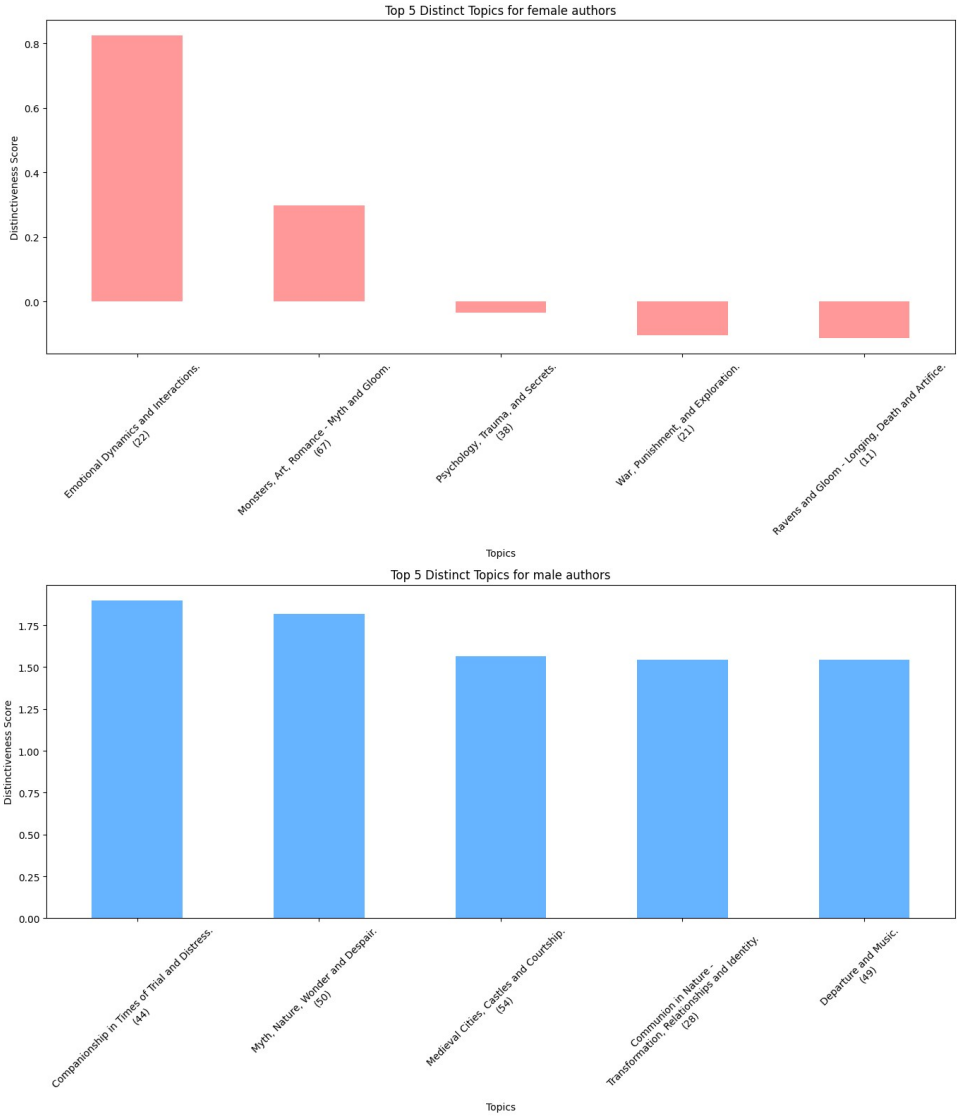


Figure 11: Distinctly gendered topics

This distinctiveness score in use here is the ratio of the specific contribution of a given gender to the total contribution, subtracted from the contribution of all others. A higher score indicates greater distinctiveness. It aims to represent the degree to which a topic is associated with one gender to the degree that it outways the contribution of opposing attributes. A negative score for a topic indicates that the topic has a less distinct association with the category of investigation than with the alternative groupings. For example, the association with war for female authors is so small in comparison to the male association with the topic that the presence of the topic becomes negatively associated with a female author.

It is challenging to make any definitive claims about these themes, particularly those that are easily gender-coded, as they appear to reflect each other in terms of general content. Both groups share themes related to entertainment, travel, mythology, fantasy, and hardship. Additionally, both have themes associated with romance and emotion. There is a slight difference in that the masculine emotional themes are more strongly associated with trials, honor, and courtship, and offer a more formal and restrained type of interaction. Topic 43: 'Companionship in Times of Trial and Distress' employs terms like 'brood, firmness, accommodate, acceptance, conducted, equilibrium,' which are outliers that obfuscate the general pattern. Meanwhile, Topic 22: 'Emotional Dynamics and Interactions,' with words such as 'breathless, hug, vociferating, moan, ruffled,' and 'brazen,' has a more immediate and passionate tone. While none of these gender-coded themes are among the most defining of the entire corpus, 38: 'Psychology, Trauma and Secrets' is prevalent enough to be among the 20 most influential topics, showing up as a defining element for not only Mary Shelley and Charlotte Smith, but also Charles Brockden Brown, Eliza Parson, and Edward Bulwer Lytton, whose work deal with grave emotional calamities. Among the most influential texts for this theme are those by Ann Radcliffe, Marie Corelli, Lee Sophia, and many other female authors in the corpus. Topic 28: 'Communion in Nature—Transformation, Relationships, and Identity,' can be considered prevalent among Romantic and Decadent writers. Texts by Poe, Byron, Wilde, and Hawthorne contribute the most to this theme.

The relationship between authors of the Romantic period and Gothic fiction was previously discussed in Chapter 2. However, it is worthwhile to briefly consider Decadence. The movement is regarded as a successor of Romantic fiction, insofar as it shares some of its attributes and some of the authors in question are the same. However, the focus is, in many ways, shifted to the negative end of the spectrum of associations and can be regarded as the most direct precursor to 20th-century Modernism and some of the avant-garde movements that arose in tandem with it.

For the purposes of this investigation, it is sufficient to mention the tendency towards self-indulgence, introspection, aestheticization, extravagance, heightened emotions, and the heavy use of symbolism.[220]

As previously stated, Hoppenstand posits that the authors of the predominantly female Gothic romance tradition catered to an audience seeking romance, while offering an outlet for repressed emotions. Those authors who followed in the wake of Walpole, however, conveyed more escapist notions, as well as violent and historic themes. Furthermore, the Romanticist tradition also had a leaning towards male writers. These findings would support these aforementioned notions.

Nationality

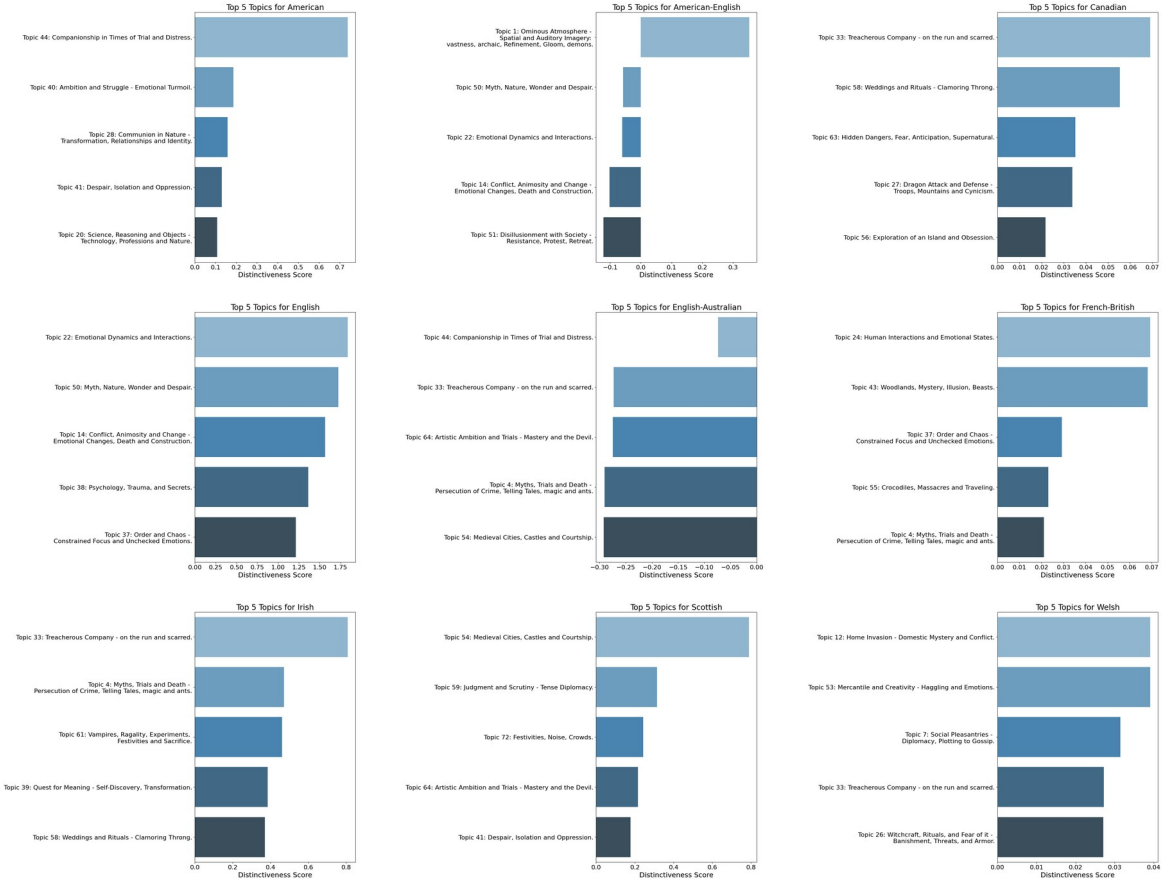


Figure 12: Distinct contribution by nationality

[220]For more in-depth insights: eg Murray, Decadence.

The calculation of national contribution is analogous to the variant for gender, as discussed above. The majority of contributions from American authors focus on masculine topics such as poise, as well as Romanticism, with Poe, Chambers, Brown, and Hawthorne having the most influence. It is interesting to note that the strongest contributing text is 'The Tell-Tale-Heart,' which subverts the expectations. The distinctly British voices carry a much stronger weight than any of the other nationalities, with two of them arising from the list of distinctly female topics: 22: 'Emotional Dynamics and Interactions' and 38: 'Psychology, Trauma, and Secrets.' While topic 38 has a very dense rate of Mary Shelley and Ann Radcliffe texts, topic 22 is very diverse in terms of authors contributing to it, and the topic carries a strong heterogeneity concerning nationality. As mentioned above, it utilizes highly passionate vocabularies like 'breathless, hug, vociferating, moan, ruffled, brazen' with the highest contribution by Vernon Lee's 'Hauntings' or Godwin's 'The Adventures of Caleb Williams.' This highlights the tendency of many works of Gothic romance to deal heavily in topics 22 and 38, while on average more women within the corpus are British. Topics 40, 41, and 44 align with the early American Gothic texts by Brown, which address themes of plague, urban squalor, and the isolation and dread of plantation life in the dark. This reinforces the notion that many of the early texts exert a significant influence on the corpus, while also underscoring that the Gothic Romance tradition was distinctly British.

Sentiment

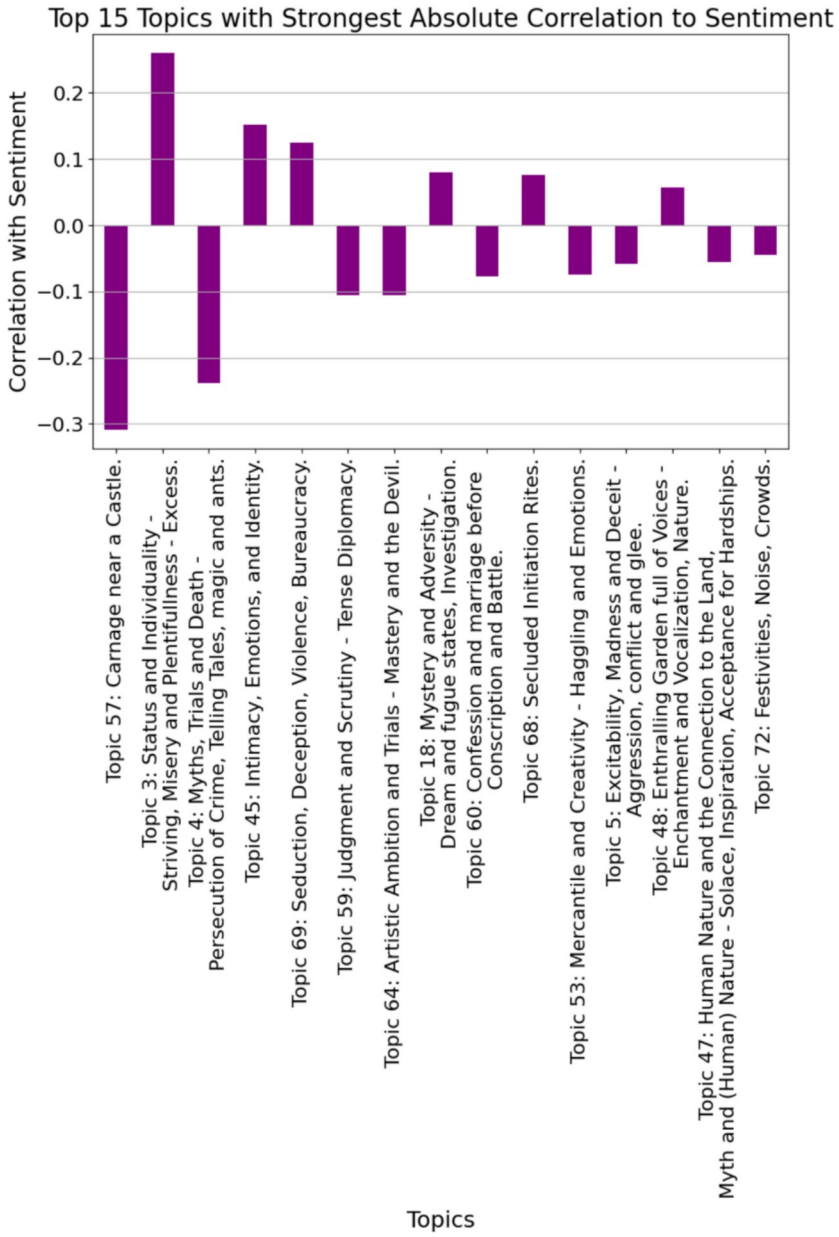


Figure 13: Topics correlated with a certain sentiment

The relationship between sentiment and various topics is not particularly strong. However, in cases where a connection is present, it appears natural and intuitive. Texts that strongly engage with topics such as carnage, crime, death, and tense judgments tend to have a negative sentiment. On the other hand, texts that focus on self-expression, ambition, intimacy, and seduction tend to have a positive sentiment. However, since only three entries have a value greater than 0.1, the

correlation is weak in almost all cases. The weakness of the sentiment association could hint at Punters association of a high density of polarizing dichotomies and emotional ambivalence within Gothic fiction.[221] These include contrasts such as the sacred and the unclean, attraction and disgust, and dread and pleasure. Such contrasts when used in tandem could result in the effects of sentiment analysis cancelling each other out.

[221]Punter, *The Literature of Terror - Volume 2*, 189f.

Cluster Analysis – PCA and K-Means

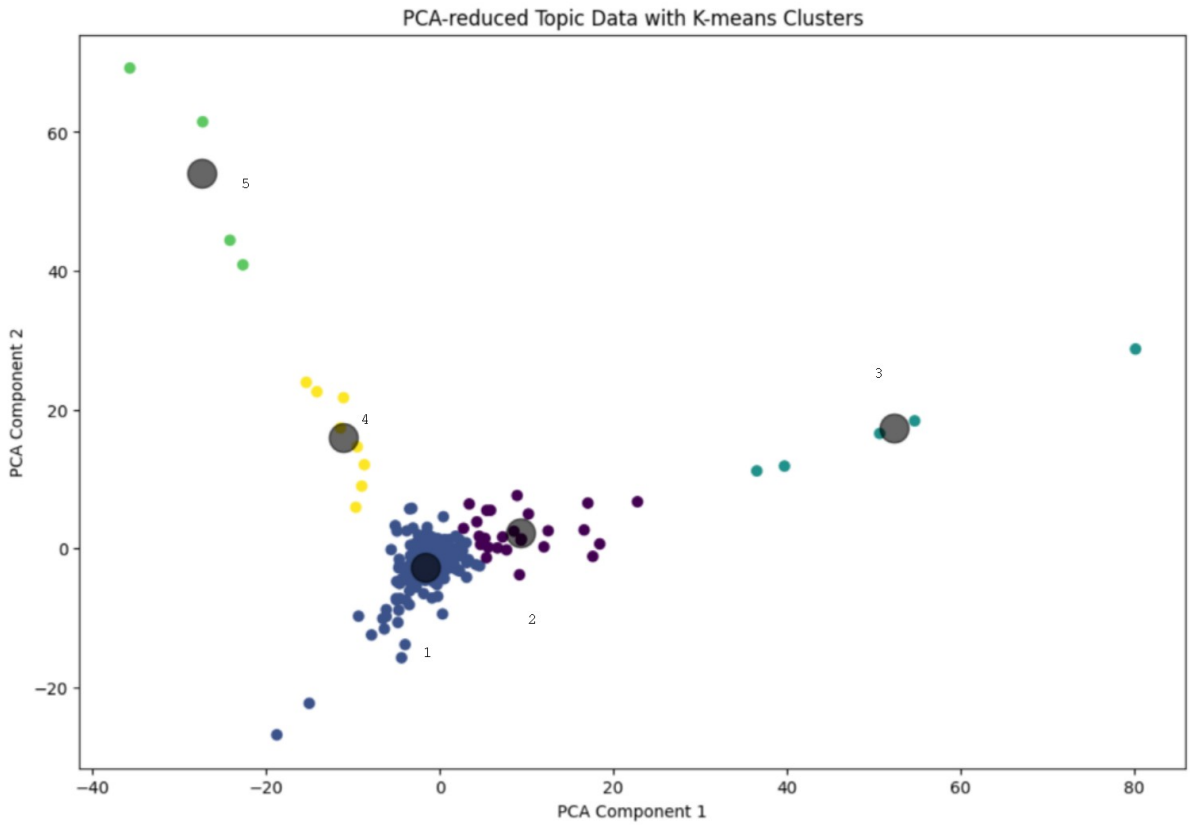


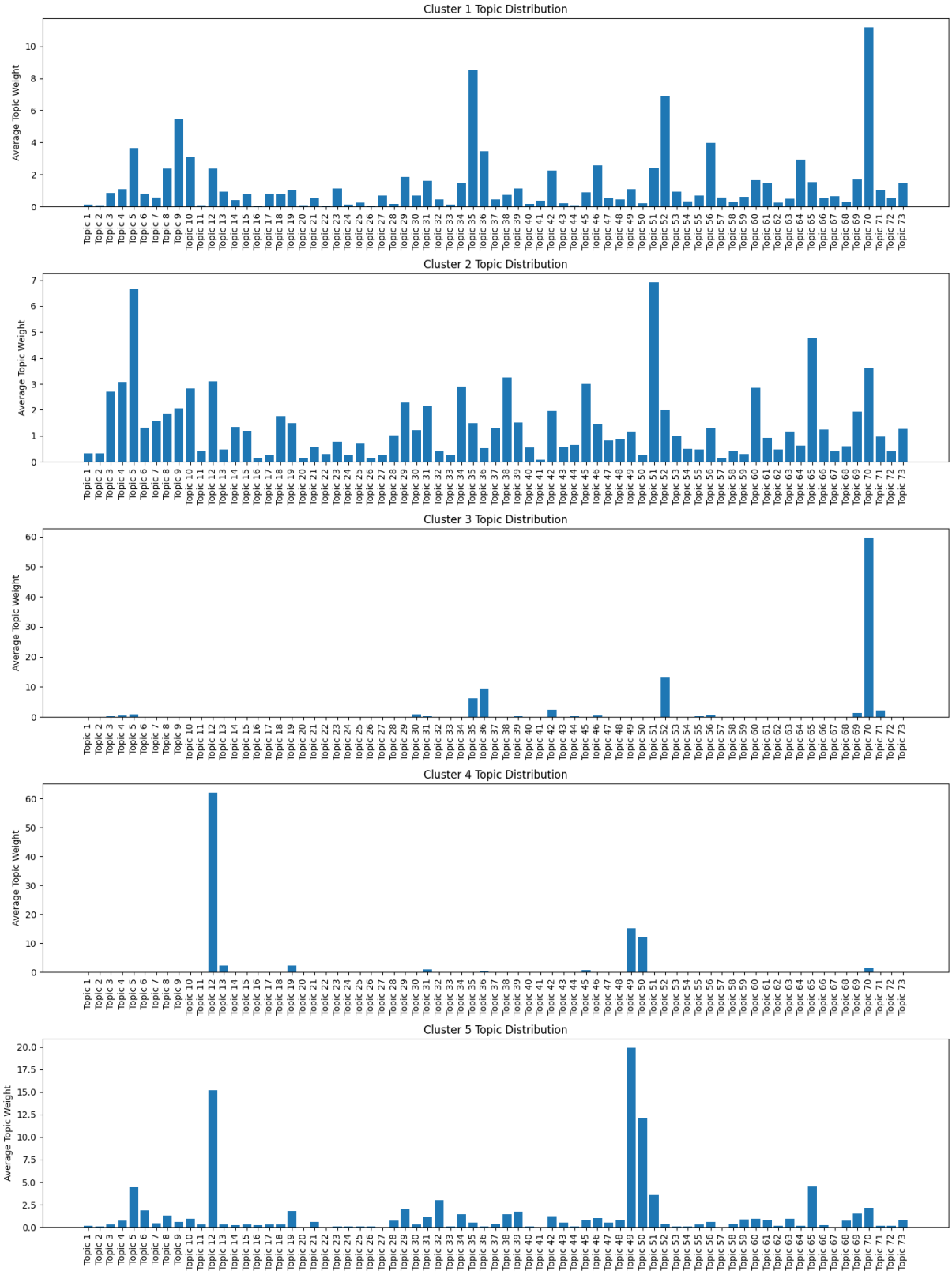
Figure 14: Principal component analysis of the topic distribution within all texts

As stated in Chapter 5, principal component analysis (PCA) is a technique for reducing dimensionality that aims to minimize the greatest amount of variation along a set of orthogonal axes. It is commonly used to explore the distribution of multidimensional features to identify outliers and interpret the composition of a dataset. Here, it is combined with K-means clustering, which is an efficient method for partitioning observations into a set of groups based on the shortest distance between each data point and the center of their assigned group.

PCA reveals a core group of texts with a balanced distribution of topics for the two central clusters and two peripheral stretches of topic distributions. The affected clusters, namely cluster 3 on the right and clusters 4 and 5 on the left, exhibit a narrower distribution that centers around a few specific topics. These selections focus on specific topics that have a significant impact on the corpus as a whole. The affected topics include 70 : 'Myth and Splendor—Wealth and Castles' for cluster 3, and 12: 'Home Invasion—Domestic Mystery and Conflict', along with the smaller

topics 49 and 50 for clusters 4 and 5. Topic 12 has a particularly strong influence on a select few influential authors who stand apart. This will be discussed in detail in the following sections.

Figure 15: Detailed look into the cluster composition



The significance of topic 70 for cluster 3 is due to its weight on some of the major voices within the corpus, such as Hawthorne and Marie Corelli, who are considered influential for the genre, but are situated later in its timespan. Further influencing factors are 35: 'Mental Illness, Law, and Outcasts—Fear, Suspicion and Struggles', 36: 'Individualism vs. Conformity—Rebellion and Social Norms,' and 52: 'Adventure, Splendor, Power and Challenges, History'. These topics are highly defining for the oeuvre of Hawthorne, but only partially so for the distribution fo the genre as a whole.

Clusters 4 and 5 additionally include topics 49: 'Departure and Music', 50: 'Myth, Nature, Wonder, and Despair', and to a lesser degree, 51: 'Disillusionment with Society—Resistance, Protest, Retreat'. While topic 49 is largely absent within the most prominent authors apart from Bierce, 50 is a topic highly relevant for the group of Romantic authors, while 51 is essential to the corpus as a whole.

Table 4: Texts comprising the outlier clusters

Key	Outlying Texts
5	HawthorneSundayatHo, Jacobs_TheMonkeyS, Jacobs_TheMonkeys, Keats_LaBelleDam
3	Aikin_SirBertran, Hawthorne_LittleAnni, Hawthorne_TheLilysQu, Hawthorne_TheMiniste, Hawthorne_TheWhiteOl
4	Bangs_GhostsIHav, Bierce_AnOccurren, Holcroft_AnnaStIves, Lewis_AlonzoTheB, Machen_TheHouseof, Stagg_TheVampyre, Stoker_TheLairOfT, unsigned_CountRoder

Table 4 shows the affected texts and provides further clarity into the division. Both Jacobs' *The Monkey's Paw* and Keats' *La Belle Dame sans Merci* recount stories of seductive encounters with a mystical entity or object, and describe the suffering that these desires caused. Similar themes are explored in Machen's collection of short stories, *The House of Souls.* The most prominent of these are *The Inmost Light, The Great God Pan,* and *The White People.* They deal with humans who cross the veil of what was previously perceivable and experience disturbing or corrupting events. Meanwhile, Hawthorne's *The Minister's Black Veil* tells the story of a man of faith who turns away from life in his community and his old life, only to rise in esteem, influence, and power through his renouncement of personal connection. *Sunday at Home,* also by Hawthorne, is an ambiguous text about worship and community, expressing a mixture of longing and contempt for a church congregation.

These selections appear to heavily focus on societal retreat, solitude, personal autonomy, and rebellion for the sake of one's convictions. However, there remains a split in the interpretation of these topics, as the model grouped them into one category focused on adventure, exploration, marveling at discoveries, and forgotten splendor. Meanwhile, the other category includes texts that are also disillusioned and oppose or depart from society, but do not find satisfaction and are troubled by external forces that cause conflict and sorrow.

Punter agrees with Machen's focus on veiled intrusions of tempting and corruptive forces at the borders of society.[222] Keats' Gothic texts, on the other hand, can be seen as a search for the horrific lurking behind depictions of great deeds. Punter compares this association stylistically to Stoker.[223] Punter supports the associations here presented. Hawthorne's writing is characterized by a focus on isolation, obsessions, mental illness, guilt, and spatial storytelling surrounding his hometown of Salem.[224] This aligns well with topics 35 and 36, but to a lesser degree with topics 70 and 52. While these topics carry historical and mythical associations, their general context does not fully fit this description.

In summary, the identified associations can largely be traced in the literature. Although Machen's primary points of comparison, Lovecraft, Stevenson, and Stoker, are only represented through the latter, his topical profile is highly isolated and unique. As the following section will explore, Stoker is also very isolated within the corpus despite his large representation. Thus, Machen is fittingly represented. Hawthorne's core topics are reflected in the literature, but it is not possible to verify his isolated position within the PCA or the subsequent network analysis through the literature. Hawthorne's and Corelli's topical makeups are particularly distinct, and they are the two largest contributors to the corpus. Thus, it is possible that the LDA model fit their contributions uniquely, singling them out in the process.

[222]Cf. Punter and Byron, *The Gothic*, 146ff.
[223]Cf. Punter and Byron, 18.
[224]Cf. Punter and Byron, The Gothic, 123f.

7.4 Network Analysis

> Until then I had thought each book spoke of the things, human or divine, that lie outside books. Now I realized that not infrequently books speak of books: it is as if they spoke among themselves. In the light of this reflection, the library seemed all the more disturbing to me. It was then the place of a long, centuries-old murmuring, an imperceptible dialogue between one parchment and another, a living thing [...][225]

While the previous sections in this chapter were focused on individual texts, authors, or their demographic characteristics, this section employs network analysis to uncover the hidden connections between key works of Gothic fiction. This approach enables the identification of the intricate web of influence that has shaped the beginnings of the genre. By mapping the relationships between texts, authors, and themes, this section seeks to identify patterns and trends that illuminate the transmission of ideas, motifs, and stylistic features. The following network graphs were created with the assistance of the Python package NetworkX. They employ degree centrality as a measure of node size, indicating that the number of interactions a specimen engages in determines its size.

Similarity between authors

The first network in figure 16 employs authors as its nodes, wherein each node represents the mean distribution across all text segments associated with it. It shares an edge or connection with each other author whose averaged topic distribution surpasses a cosine similarity of 0.85. Only the most central of authors are maintained for legibility, and unconnected or self-referential authors have been removed.

The largest cluster, comprising Mary Shelley, William Godwin, Frances Burney, and Charles Brockden Brown, represents the foundation of the Gothic genre. This group is characterized by their influence on the overall network and their particularly virulent representation of social issues in their work. Their position at the heart of the network suggests that they played a significant role in shaping the genre's early development. Mary Shelley, in particular, is often credited with creating the modern Gothic novel with her iconic work, *Frankenstein,* a text deeply entrenched in identity formation and questions of ethics and the boundaries of scientific curiosity. Godwin's philosophical and humanitarian works influenced the genre's exploration of

[225]Eco, *The Name of the Rose*, 286.

morality and social issues and had shaped the format as a political and interpersonal vessel capable of grappling with Romantic ideas on human autonomy and community. Godwin, Shelley, Brown, Smollet, and Hogg all tackle grievances with cultural developments either in religion, scientific rational discourse, or governmental institutions.

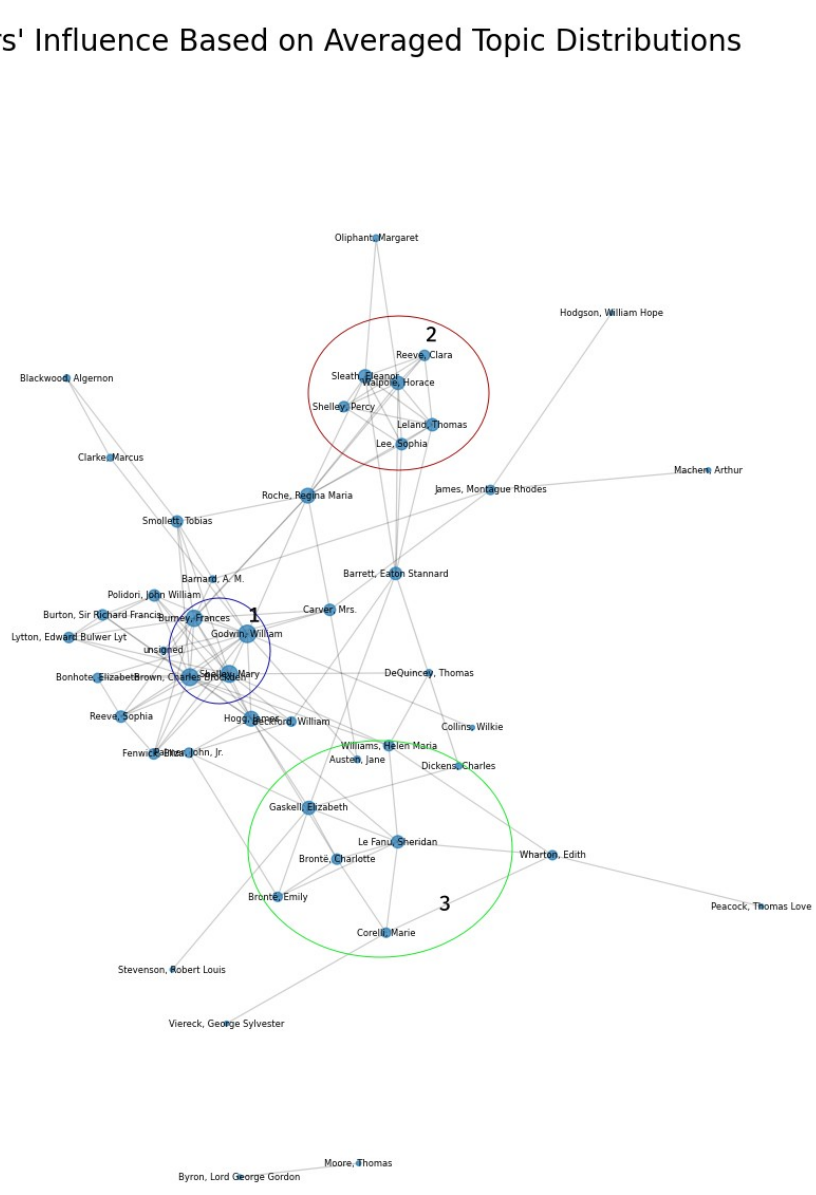


Figure 16: Author's influence based on averaged topic distribution

The second cluster, consisting of Percy Bysshe Shelley, Horace Walpole, Eleanor Sleath, and Thomas Leland, represents an early group of authors who laid the groundwork for the Gothic genre. Their presence before 1800 underscores their pioneering role in establishing the genre's

foundational themes and settings. Walpole's *The Castle of Otranto* is often cited[226] as the first English Gothic novel, while Sleath and Leland were influential writers of the time, establishing the mood, atmosphere, and core motifs in use.

The smaller, less densely connected cluster featuring Sheridan Le Fanu, the Brontë sisters, Jane Austen, Charles Dickens, Elizabeth Gaskell, DeQuince and Marie Corelli represents a focus on detailed character portraits, as well as deeply psychological and interpersonal fiction. This group highlights the genre's shift towards a wider reach and hybridization, which brought to light more nuanced character studies and introspective themes. These authors exhibit a strong correlation with the character- and narration-driven traits that Punter attributes to texts that facilitated the evolution of the modern novel. To use Barthes' terminology, these characters are less agents that serve the themes and objectives of a plot, and more characters for their own sake.

Interconnections and Influences

The connections between these centers reveal key influences and relationships:

Regina Maria Roche bridges clusters 1 and 2, suggesting that her work influenced both groups. Given that Roche's works were incredibly well-received and became famous bestsellers,[227] it stands to reason that her contribution to popularizing the format would establish the subset of functions frequently employed by subsequent authors. Eaton Stannard Barrett links clusters 2 and 3, implying that his writing had an impact on tying the psychological focus of Gothic fiction into its set of tropes. Given the remarkable popularity of *The Heroine,*[228] its distinctive integration of Gothic narrative with humorous political commentary, and themes of female emancipation, he facilitated the convergence of the Gothic Romance texts, those that established classical tropes, and those of a more psychological nature. James Hogg connects clusters 1 and 3, indicating that his satirical work on the moral misgivings of religion influenced the evolution of psychological themes as well as social themes in Gothic fiction.

The absence of authors like Hawthorne, Stoker, and Ann Radcliffe from this network is notable. This may be due to the unique styles or idiosyncrasies of these writers, which don't align with the dominant topics or relationships in this specific aggregation.

[226]Eg. Punter, *The Literature of Terror - Volume 1*, 43.
[227]See Botting, *Gothic*, 46.
[228]Cf. Kelly, "Unbecoming a Heroine."

In conclusion, this network analysis provides valuable insights into the early development of Gothic fiction. The three centers represent distinct groups that contributed to the genre's evolution:

1. The foundational group, comprising Mary Shelley, William Godwin, Frances Burney, and Charles Brockden Brown, shaped the genre's intellectual, political or social points of contrition, more akin to free motifs. Veselovsky's conceptualization of the evolution of generic motifs in parallel with one another due to shared social circumstances allows for the understanding that these were not only active acts of reception, but also individual co-creations of similar motifs due to similar conditions. Punter argues against that assertion, instead claiming that, as traditional, rural patterns of life in America and Britain were changed and disrupted in favor of industrialization, the subsequent urban squalor became increasingly apparent, significantly influencing the development of the genre.

2. The pioneering group, consisting of Percy Bysshe Shelley, Horace Walpole, Eleanor Sleath, and Thomas Leland, laid the groundwork for the genre's bound motifs of settings, tropes, and style of presentation.

3. The psychological group, featuring Sheridan Le Fanu, the Brontë sisters, Elizabeth Gaskell, and Marie Corelli, as well as Jane Austen and Charles Dickens, focused on internal struggles and character development, highlighting the genre's entanglement with the development of modern plot and character driven novels.

Groups 2 and 3 similarly support Veselovsky's argument that the periphery is the primary source of innovation, leading to integration from within. While clusters 1 and 2 feature a multitude of authors who were actively engaged at the advent of the genre's development, the texts of those active in cluster 3 offer a markedly novelistic, introspective, and psychological perspective. This cluster encompasses numerous authors who were not central contributors to the genre but whose writing exhibited traits that later proponents like Lovecraft, Blackwood, and the later works of Machen would integrate into their representation of genre fiction. Many of the authors in this group were not core contributors to the genre, but rather took up its form for individual texts and enriched the repertoire for the discourse at large. Regina Maria Roche's commercially successful bridging of groups 1 and 2 enforces Hoppenstand's assessment that popularization solidifies thematic elements to a core selection of motifs that an audience subsequently is expected to co-create through the reception process. The chronological dimension to the influence of these

authors will become apparent in Figure 18, which depicts connections traversing from younger to older texts to which they refer back.

Similarity between texts

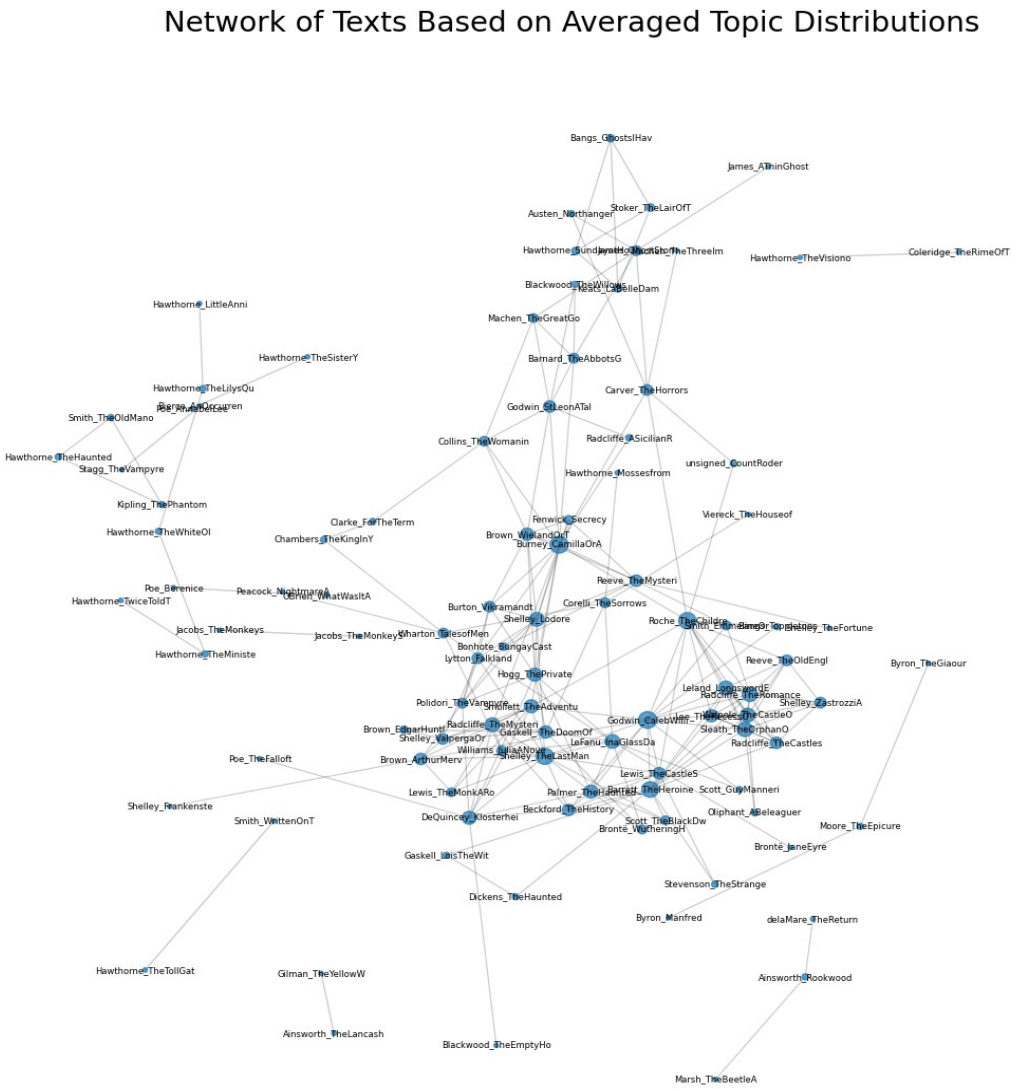


Figure 17: Influence of individual texts based on their averaged topic distribution

The second network employs texts as its nodes, wherein each node represents the mean distribution across all text segments associated with it. It shares an edge or connection with each other text whose averaged topic distribution surpasses a cosine similarity of 0.85. Only the most central of texts are maintained for legibility, and unconnected or self-referential texts have been removed. In contrast to the previous network, this view enables a more detailed analysis, allowing the impact of individual texts to be more clearly discerned while reducing the influence of authors with a broader and more diverse topical scope across numerous texts.

This network has many similarities with the previous one, but still provides unique insights:

This grouping puts more emphasis on Mary Shelley and William Godwin as a thematic bridge between different groupings of Gothic fiction authors, highlighting their thematic sway over multiple aspects of the genre formation as a whole. Ann Radcliffe, often considered the mother of Gothic fiction, has taken on a central position within the early pioneers of the genre bound motifs, or tropes. The collection of texts that are more strongly affiliated with the genre's free motifs and intellectual composition also occupy a central position.

The smaller hub that previously focused on psychological exploration and internal struggles has largely fractured and been reabsorbed into other groupings. Francis Burney's *Camilla* (1796) has taken center stage, surrounded by a few other texts, highlighting the complexity inherent in his investigation of social, emotional and mental challenges in his intergenerational exploration of interpersonal relationships.

Machen and Blackwood, previously isolated from the main network, have reunited at the periphery, reaffirming their shared exploration of themes of forbidden knowledge and the intrusion of outsiders into societal norms. Their work is connected to the main clusters through Godwin's *St. Leon* (1799) and Carver's *Horror of Oakendale Abbey* (1797). Carver's and Godwin's texts connect Faustian deals with supernatural forces and depictions of grave robbery and gore with classic Gothic settings, such as mad aristocracy and castles. This bridges the gap between Machen and Blackwood, who share similar topics but embed them in more contemporary surroundings.

As before, Hogg continues to play a significant role in connecting different groupings. Beckford is absent this time, but Lytton has taken his place; this association puts more emphasis on the interaction between classical Gothic texts covering social and religious themes and texts of other romantic authors.

Hawthorne's works remain disconnected from most other pieces, indicating that his self-referential themes and stories may have been too unique or isolated. Stoker and Byron are largely absent from this network, but their presence is felt in their mutual influence on each other.

In conclusion, this network offers a more nuanced perspective on the interconnections between the aforementioned divided spheres, while still maintaining a cluster of thematic reformers and social critics, with Godwin and Shelley frequently acting as bridges or transgressors of

boundaries. It is noteworthy that Roche, Corelli, and Burney, authors who popularized the genre and reached a much wider audience than their contemporaries, are closely associated with one another. This suggests that the popularization of the genre led to a distillation of characteristics that set them apart from other, more closely knit groupings. This positioning would lend support to Moretti's assertion that market demands result in the emergence of generic forms from the center of production, a waterfall of outward facing unification that facilitates the process of rapid adoption. Hoppenstand would argue that this integration of readers' expectations into the formal requirements is a fundamental aspect of this process, allowing for a more recognizable format that can reach a wider audience.

7.4.1 Network of Influence

The following network is an adaptation of Matt Erlin's investigation into the prevailing thematic groupings found in German Romantic texts and their relations to one another.[229] Erlin employed the results of topic modeling to create a network between text chunks by linking works that share at least one one-thousand word passage on the same topic with a participation strength of 20% or higher.

In this comparison, the cosine similarity between all sections of a text was calculated independently. Only the ten most prominent topics per section were retained for comparison with other segments in the network. To estimate the impact of influence between individual texts, edges indicating similarity were established unilaterally from older to newer texts. Degree centrality was once again employed as a measure of node size, indicating that the number of interactions a text has is what determines its size.

This chronological grouping introduces a few previously unidentified texts, yet the larger picture reinforces the impression derived from previous networks. The fact that the labeling has been conducted in order of degree centrality allows for a more precise ranking than any of the preceeding attempts.

The arrangement of the network has undergone a significant transformation, resulting in the emergence of a single, expansive, loosely integrated cluster comprising a multitude of sub-groupings. Additionally, a select few highly influential outliers have emerged, seemingly independent of prior works in the genre, which served as a rich source of inspiration for subsequent texts.

[229]Erlin, “The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864.”

The clusters are as follows:

Table 5: IDs of the texts that comprise the individual clusters

Cluster	Text IDs	Color
Upper Outliers	22, 69, 24, 44, 39, 40, 41	light green
Upper Half	38, 29, 32, 77	dark green
Upper Half Lower	42, 36, 46, 53, 58	light blue
Lower Outliers	19, 28, 52, 33, 34	magenta
Bottom Half Up	13, 45, 31, 23, 30	dark purple
Bottom Half Upper	20, 17, 47	orange
Center Bottom Right	25, 6, 74, 18	yellow
Dead Center	2, 9, 10, 27, 35, 37	light red
Center Left	12, 8, 4, 5, 15, 7, 21	dark blue
Center Top Right	1, 3, 11, 26, 14, 16	black

The initial analysis was based on a version of the subsequent graph devoid of any additional features; only the placement and relationship of the nodes were considered. However, for the purpose of subsequent evaluation, additional features inherent to the corpus, such as the period and role that Caroline Winter assigned to a given text, were also consulted. In order to avoid redundancy, only the graph used for evaluation is included here; however, both versions are included in notebook three. In order to interpret the topical composition of each cluster, the 15 most prevalent topics for each cluster have been visualized and analyzed for their shared features. This visualization is provided in the appendix.

The following section will interrogate each cluster individually, comparing the composing texts by their shared plot elements and shared topical composition, as well as period and role in the development of the genre. It should be noted that any references to Romantic or Victorian texts in the following section pertain to the period in question, rather than to a specific movement.

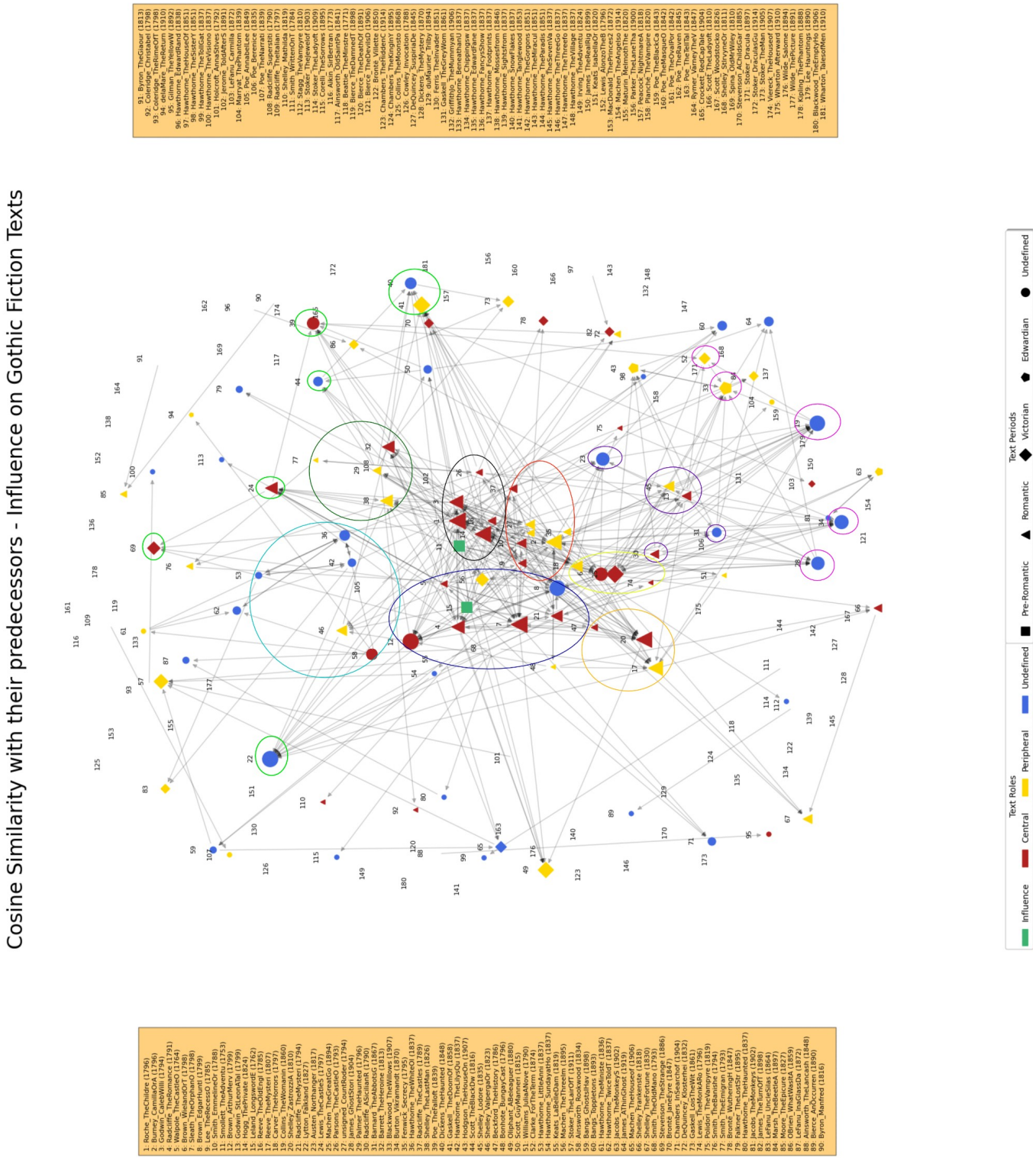


Figure 18: A detailed network of the influences within early Gothic fiction

Upper Outliers:

Texts:
• 22: Edward Bulwer-Lytton's Falkland • 69: Mary Shelly's The Fortunes Of Perkin Warbeck • 24: Mathew Lewis' The Castle Spectre • 44: Walter Scott's The Black Dwarf • 39: Edgar Allan Poe's The Fall of the House of Usher • 40: Charles Dickens' The Haunted Man and the Ghost's Bargain • 41: Elizabeth Gaskell's The Doom of the Griffiths:

The Upper Outliers cluster explores themes that delve into the darker aspects of human experience, including psychological terror, fanaticism, transgressive indulgence, and moral debasement, as well as politics. This constellation of topics is centered on tumultuous emotional experiences, where the sacred and profane collide, and societal discontent simmers just below the surface. The three texts in this cluster—39, 69, 24 from Poe, M. Shelley, and Lewis—are high-impact texts that stem from the pivotal moments of either the first or second wave of high Gothic productivity. These texts exemplify characteristic features of inter- and intrapersonal conflicts in high Gothic texts. This cluster's contribution to the genre as a whole is a set of ephemeral free motifs of putting a character's psyche under duress via inhospitable circumstances and all the trappings this entails.

Upper Half:

• Texts:
• 38: Mary Shelley's The Last Man • 29: John Palmer's The Haunted Cavern. A Caledonian Tale • 32: Eaton Barrett's The Heroine • 77: Charlotte Smith's The Emigrants

The Upper Half cluster is characterized by a shared plot structure of societal disarray and alienation, as well as a playful subversion of expectations. The thematic focus is centered on clamor, grief, social and emotional upheaval, and identity crises. The four texts in this cluster—three early Romantic works and one influential text by M. Shelley—are notable for their significance to the genre. Eaton Barrett's *The Heroine* stands out as a pivotal work that not only popularized Radcliffian Gothic fiction, but also cleverly satirized and self-referentially critiqued

its own narrative conventions. This text was widely acclaimed by fellow authors within the genre and enjoyed immense popularity among a broader audience. Its node is nigh on central. This cluster bridges both free motifs that drive the leitmotif of a genre, and static motifs establishing archetypical settings and plot patterns.

Upper Half Lower:

- Texts:
- Nathaniel Hawthorne's The Lily's Quest
- Nathaniel Hawthorne's The White Old Maid
- Nathaniel Hawthorne's Little Annie's Ramble
- William Ainsworth's Rookwood
- Mary Shelley's Valperga

The Upper Half Lower cluster explores themes of life, death, and the pursuit of happiness at great personal risk. The topical focus centers on an entirely different makeup than other clusters, with a focus on human desires, yearning, health, atmospheric settings, and fantastical elements. The texts in this cluster are dominated by Hawthorne, yet both *Valperga* and *Rookwood* share the same traits of ill-fated convictions that lead to a preemptive death. Hawthorne's role within the genre was seemingly not recognized as such by the creators of either of the corpora made use of here. The other graphs allotted him little space, which is seconded here, even if his texts fall quite squarely within the second wave of increased production within the corpus. This evaluation remains peculiar, as these texts undoubtedly occupy a significant space within the Gothic literary tradition. This cluster's addition to the genre stems, in large degree, from its addition in the realm of free motifs, reclaiming the moral legacy of Puritanism, the past's weight on the present, and the questions of redemption from history and individual identity.

Lower Outliers (similar to Upper Outliers):

- Texts:
- 19: Wilkie Collin's The Woman in White
- 28: Montague James' Ghost Stories of an Antiquary
- 52: Marcus Clark's For the Term of his Natural Life
- 33: Algernon Blackwood's The Willows
- 34: Richard Francis Burton's Vikram

The Lower Outliers cluster is very similar to the Upper Outliers cluster and explores themes that address systemic violence and supernatural othering encounters. The topical focus centers on extreme societal disenfranchisement and punishment. Of the texts in this cluster, only Blackwood's *The Willows* (33) and Clark's *For the term of his natural life* (52) are recognized for their influence on the genre. Notably, *The Willows* is often regarded as one of the finest pieces of supernatural fiction and an important predecessor to the Weird Fiction surge within the 20th century.[230] This text appears well-connected to the lower sections of the center, particularly Arthur Machen and Godwin's *St. Leon*, which share a common atmosphere of ominous foreboding and anxiety. This connection is in line with the previous network. Like the cluster Center Bottom Right, the contribution of this cluster to the genre is focused on a set of free motifs that lead from the core themes of the genre to shape future genres that arose out of it.

Bottom Half Up:

• Texts:
• 13: William Godwin's St. Leon
• 45: Mary Shelley's Lodore
• 31: AM Bernard's The Abbot's Ghost
• 23: Jane Austin's Northanger Abbey
• 30: Ann Radcliffe's A Sicilian Romance

The Bottom Half Up cluster is defined by a shared plot structure that explores fallen nobility, emotional distress and struggles of love, and satire. The cluster's topical focus centers on domestic and social conflicts, nobility, conflict, and intimacy. The four texts in this cluster are notable for their unique characteristics. Godwin's *St. Leon* (13) is a central text that explores Faustian bargains, fallen nobility, and madness. M. Shelley's *Lodore* (45) is a peripheral text that offers a feminist and egalitarian perspective. Jane Austen's *Northanger Abbey* (23) is renowned for its delicate parody of the genre, while remaining a prime example between the first and second spikes. Ann Radcliffe's *Sicilian Romance* is a central text with very classical Gothic Romance themes from the first wave of genre production. This cluster's input on the development of the genre provides both defining character traits and distilled interaction patterns that shape the conventions of Gothic fiction. Hoppenstand argues, that, for a genre to be parodied, its conventions and constraints already need to be firmly established. A closer look at

[230]Cf. Punter, The Literature of Terror - Volume 1, 3; Punter, *The Literature of Terror - Volume 2*, 81.

the works that influenced *Northanger Abbey,*[231] show which texts would have constituted the frame of reference Austen built on. Among them are 11, 1, 16, 3, 35, 2 and 37: Roche, Godwin, Smollet, Fenwick, Burney, Reeve and Radcliffe. This is a wide selection of texts with many examples of a more popular nature, among them texts that are actively passed on as endorsed by Austen and pieces of social critique. This reinforces the Hoppenstand's argument and reaffirms that by 1817, the core genre features had firmly solidified. This assessment falls in line with Moretti's argument that many genres of this period would have solidified in their core topics and only diversified and specialized past 1820, as discussed in Section 7.2.

Bottom Half Upper:

- Texts:
- 20: Percy Shelley's Zastrozzi
- 17: Clara Reeve's The Mysterious Wanderer
- 47: William Beckford's Caliph Vathek

The Bottom Half Upper cluster explores themes that revolve around dark desires and dire consequences, pacts, and the struggle between good and evil. The topical focus centers on battle, death, aggression, conflict, resistance against demands and mystery. All three texts in the Bottom Half Upper cluster are Romantic and very early works. Shelley's *Zastrozzi* is regarded as a central text in the genre, influential for its iconic villain and exploration of cruel self-indulgence, revenge, and amorality. Beckford's *Caliph Vathek* is an early Orientalist adaptation of Walpole's seminal text, highly successful and esteemed for its portrayal of amorality, devil's pacts, and powerful use of architecture. The text successfully combined elements of earlier Gothic texts with influences from the *Arabian Nights,* which exerted a significant influence on the English Romantic movement.[232] *The Mysterious Wanderer* was a popular book during its time, reaching a large audience. However, it has since been largely forgotten, ceding recognition to the other two texts. Like the preceding cluster, this cluster's influence on the genre resides more in the realm of bound motifs, providing defining input on archetypical villains and the moral trappings they contribute to the plot at large.

[231]See Punter and Byron, *The Gothic*, 81.
[232]Punter and Byron, 181.

Center Bottom Right:

• Texts:
• 25: Arthur Machen's The Great God Pan
• 6: Charles Brockden Brown's Wieland
• 74: Mathew Lewis' The Monk
• 18: Mrs. Carver The Horrors Of Oakendale Abbey

The Center Bottom Right cluster explores themes that encompass highly grotesque, physical depictions in religiously coded settings, preceded by moral transgressions and body horror. The topical composition is characterized by a focus on the invasion of the sacred, religion, death, and ferocity. The texts in this cluster are all considered central or influential works. *The Great God Pan* (25) is a central Romantic text within the first peak of the genre, which was highly influential for its exploration of Christian symbolism, scientific amorality, pagan rites and crossing of boundaries. The text drew heavily from the works of Poe, Le Fanu, and Stevenson, and its influence can be seen in the works of authors such as Charles Brockden Brown, H. P. Lovecraft, Oscar Wilde, and Bram Stoker, as well as in early 20th-century Weird Fiction. Brown's *Wieland* is a gruesome early American text that explores Christian symbolism, madness, and supernatural elements, making it another central work in this cluster. Punter would categorize these texts as 'degeneration based' works that dwell on the essence of humanity and its inevitable boundaries under strain.[233] He notes that these texts explore the boundaries of taboo and human experiences alike, with their transgressions carrying knowledge and corruption alike.[234] Lewis' *The Monk* (74) is surprisingly peripheral to this cluster, despite being widely regarded as one of the most central texts. This is due to its unique topical composition: 70, 63, 51 (highest), 38, 34, 30, 9, 4, 5. It still shares 70, 5 and 38 with the other texts in the cluster. Its topical composition emphasizes persecution, treacherous company, murder, and temptation by devils. While it shares some elements with the essential texts in this cluster—including as a focus on clerical elements and transgression—its overall makeup falls out of line with the others due to heavy romance and clergy elements. Finally, *The Horrors Of Oakendale Abbey* (18) is a peripheral Romantic text that is highly gruesome and grotesque, but not regarded as highly influential due to its uncouth themes. Despite this, it was a highly popular and influential work in its time. Like the cluster Lower Outliers, the addition of this cluster to the genre itself is more focused of static motifs that establish locale and characters, but it also contributed a set of free

[233]Punter, The Literature of Terror - Volume 2, 1.
[234]Punter, 2.

motifs that are more a divergence from the core themes of the genre, but would still grow to shape future genres that arose out of it.

Dead Center:

• Texts:
• 35: Eliza Fenwick's Secrecy
• 2: Frances Burney's Camilla
• 9: Sophia Lee's The Recess, Or A Tale Of Other Times
• 10: Charlotte Smith's Emmeline, Or The Orphan Of The Castle
• 27: the anonymous Count Roderic's Castle
• 37: Ann Radcliffe's The Castles Of Athlin And Dunbayne

The Dead Center cluster's shared plot structure is defined by female protagonists struggling for control, their place in a changing environment, or love, also known as the Gothic Romance. The shared topical composition is characterized by a focus on a very high occurrence of emotional and interpersonal conflicts, Gothic settings, and quests for identity. The texts in this cluster feature strong female representation, and most of them are regarded as central and Romantic works by Winter. Fenwick's *Secrecy* (35) is a peripheral Romantic text that was an early influence on Radcliffe's *The Italian,* exploring morbid, feminist and transgressive themes of female independence. Burney's *Camilla* (2) is an immensely popular generational coming-of-age story, blending comedic elements with gothic episodes, praised by Jane Austen. Sophia Lee's *The Recess, Or A Tale Of Other Times* (9) is a very early, Romantic, and central text that features political intrigue, heroic female leads, seafaring, warfare, gothic castles, and marriage; it was a highly popular title. Charlotte Smith's *Emmeline, Or The Orphan Of The Castle* is a Romantic, central, and early text that tells a Cinderella story of female emancipation, gaining property and standing, where ownership over masonry and bodily autonomy blend. This text was also highly financially successful.[235] The position of Radcliffe's text *The Castles of Athlin and Dunbayne* (37) reinforces the association that popularity, financial success, and a consistent shared form went hand in hand for Gothic Romances, leading to the type of rapid adoption of patterns both Moretti and Hoppenstand would expect. The Dead Center cluster unites a number of highly popular texts featuring heroic women, romance, and terror, some of which are largely overlooked. The contributions of this cluster reside more in the realm of static motifs, establishing critical character constellations and themes.

[235]Cf. Punter and Byron, *The Gothic*, 166.

Center Left:

- Texts:
- 12: Charles Brockden Brown's Arthur Mervyn
- 8: Charles Brockden Brown's Edgar Huntly
- 4: Ann Radcliffe's The Romance Of The Forest
- 5: Horace Walpole's The Castle of Otranto
- 5: Thomas Leland's Longsword, Earl Of Salisbury
- 7: Eleanor Sleath's The Orphan Of The Rhine
- 21: Ann Radcliffe's The Mysteries Of Udolpho

The Center Left cluster plot composition explores themes of heightened emotions, insanity, and misery. The core foundational early British texts and their first adaptations across the pond carry a clear distinction in their plot structure; the British texts focus on the sublime, ghosts, and castles, while their early American adaptations grapple with disease and the wilderness. The former reclaimed and reprocessed parts of their history, while the latter still had much more immediate survival concerns. The cluster's shared topical composition encompasses social and emotional distress, more interpersonal intimacy and interactions, Gothic settings, supernatural impressions, and violence.

The cluster is composed of central early texts of the genre, including Radcliffe's works and Eleanor Sleath's *The Orphan Of The Rhine.* These picturesque novels feature scoundrels, half-ruined castles, strong emotions, romance, terror, and are praised for their adaptations of Radcliffe's style.[236] Indeed, Jane Austen herself acknowledged the influence of some of these novels on the composition *Northanger Abbey.* The cluster also includes Horace Walpole's *The Castle of Otranto,* a haunted medieval castle narrative that features horrific supernatural happenings, violence, surprising humor, sexuality, and questions of identity. This work is retrospectively considered one of the earliest, most formative, and exemplary of the Gothic fiction genre. While many of the British texts in this cluster deal with love, sexuality, royalty, and a great deal of historical fiction, the American branch has a distinctly more individualist and contemporary touch. Brown draws more heavily from Godwin than the others, creating a distinctive voice of wilderness, social disarray, and abandonment in his works such as *Arthur Mervyn* and *Edgar Huntly.* These texts defined the American branch of the genre as early highly influential adaptations. In contrast, *Wieland* (Center Bottom Right cluster) is characterized by a

[236]See Punter and Byron, 81.

more carnal approach, while *Edgar Huntly* is more philosophical and closely aligned with both the social and motif focused side of the tradition. Meanwhile, *Arthur Mervyn* stands out for its expansive narrative and darker tone as it addresses the topic of yellow fever. All texts in this cluster are regarded as central early texts of the genre by Winter. This cluster bridges both free motifs that drive the leitmotif of a genre and static motifs that establish archetypical settings and plot patterns.

Center Top Right:

- Texts:
- 1: Regina Maria Roche's The Children Of The Abbey
- 3: William Godwin's Caleb Williams; Or, Things as They Are
- 11: Tobias Smollett's The Adventures Of Ferdinand Count Fathom
- 26: Eliza Parsons' The Castle Of Wolfenbach
- 4: James Hogg's The Private Memoirs And Confessions Of A Justified Sinner
- 16: Clara Reeve's The Old English Baron

The Center Top Right cluster's shared plot structure is occupied with turmoil within power structures, whether through family intrigue, religious power struggles, or the quest of individuals against corrupt institutions. The shared topical composition is characterized by royalty, institutions, tragedy, obsession, individualism, and revolt. Tobias Smollett's T*he Adventures Of Ferdinand Count Fathom* stands out as the only pre-Romantic and non-central text in this cluster, but its influence is undeniable. It shares many similarities with *Zastrozzi. The Castle Of Wolfenbach* (26), written by Eliza Parsons, is an important early work that predates many of the Radcliffe texts and is praised by Jane Austin as essential.[237] This Gothic romance features royalty, frenzied expressions of emotion, fainting, weeping, and struggles with identity formation. It is an early outlier in the cluster. Clara Reeve's *The Old English Baron* (16) pays homage to Walpole's *Castle of Otranto,* but was intended to be a more realistic and streamlined Gothic template that was thus widely adopted. This text features horror, mystery, ghost stories, and castles. Regina Maria Roche's *The Children Of The Abbey* (1) was one of the best-selling novels of the 19th century and deals with wicked relatives, a languishing quest for inheritance, love, royalty, castles, and heightened emotions.

[237]Cf. Punter and Byron, 81.

This cluster includes three very popular texts (1, 26, 16) that continue the trend of the Center Left and Dead Center. They are highly popular Gothic Romances in the style of Radcliffe or even preceding it, or sharing in Walpole's gruesome, sexual, and supernatural style. William Godwin's *Caleb Williams; Or, Things as They Are* (3) can be considered an outlier in this group, as it provides the core underlying philosophical side of the genre. This text features treachery, persecution, political oppression, corrupt hierarchies, and psychological obsession. Godwin was a major influence on Brown, as mentioned above, or more generally, influenced the works of his daughter M. Shelley, including *Frankenstein,* as seen here in the network. It explores individualism and the constraints of institutions. James Hogg's *The Private Memoirs And Confessions Of A Justified Sinner* deals in a castle haunted by a young woman wrongfully accused of murder. This text grapples with corruption and moral degradation of institutions, both within gruesome settings. While *Caleb Williams* attempts to actively advocate for a cause from the perspective of the downtrodden, *The Private Memoirs And Confessions Of A Justified Sinner* was almost unnoticed until the 20th century, likely due to the deep religious criticism put forth by its anti-hero. This cluster can be considered a collection of the core free motifs, comprising style, theme and an intellectual set of references, as well as the core static motifs of landscapes and character constellations.

Conclusion

To conclude the evaluation of the clusters within the network, additional features of the dataset taken from the investigation of Caroline Winter were overlaid onto Figure 18 to add additional dimensions. This was done after the interpretation to prevent the bias of past research from influencing the patterns found. In general, the rise of the genre explored within this network is strongly associated with authors classified as Romantic, and almost all the core texts central to the network are considered central, though some peripheral texts are included. Nearly all aforementioned texts within the network come from Caroline Winter's selections, although in cases where several sources covered the same text, her variant was preferred due to better textual quality and additional metadata. The central clusters exhibit a pronounced prevalence of female authors, comprising over 50% of the total. Many of the texts included are authored by individuals who would be classified as genre or popular fiction writers. They are interlaced with works by Brown, Lewis, Poe, Dickens, and Austen, which are perceived as occupying a more elevated position within the literary canon. Additionally, the collection features contributions from Godwin and Shelley, who played a pivotal role in shaping the foundational Romantic

elements of the genre. Many popular texts are by female authors dealing in romance, or the gruesome pulp variants, including those in the Walpolian tradition. Gothic fiction is distinguished by the prominence of female authors, who played a pivotal role in popularizing the genre and achieving significant financial success. Notable examples include Roche, Smith, Lee, and Reeve. Additionally, female authors contributed to the evolution of the style through conscious departures from conventional norms, as evidenced by Radcliffe, Austen, and Shelley. This reinforces Hoppenstand's conceptualization of fundamental static motifs that establish a popular canon that unifies settings and characters, which disseminates structure outward and was financially viable. However, it also reinforces a picture in line with Veselovsky's core beliefs, that innovation trickles inward. Ultimately, it shows that core Romantic-era authors outside the genre itself provide some of the intellectual foundation for it and expanded the genre in its early conception in terms of developing internal themes to grapple with.

7.4.2 Contextual Comparison

The comparisons on the level of influential topics and authorial contribution to a given topic put more emphasis on the uniqueness of certain voices towards a specific theme and association. This put a lot of focus on authors with a characteristic style that carried a wide reach, such as Poe, Le Fanu, and Lewis. The network analysis showed that while their contributions were influential and formative for a certain aesthetic, they did not invite the same kind of imitation and provide the basis for the formation of a movement that the likes of Walpole, Radcliffe, Brown, or Shelley did. In the topical analysis, Walpole's voice was not represented, while his influence carried exceedingly far on the level of textual similarity at the start of a movement. This means that while he was influential and formative, he was not distinct in the same manner. The influential, exceptional voices which combined both analyses are Godwin, Mary Shelley and to a lesser degree Ann Radcliffe.

Regarding the points of contentions in the formation of genre a more nuanced picture takes shape. To reiterate, for Moretti, the main source of innovation comes from the center of cultural production and is disseminated outward, while Veselovsky argues that the periphery is the main source of innovation, leading to integration inward. For Veselovsky, hybridization is a result of innovation at the periphery, and while he agrees with Hoppenstand that periods of social unrest lead to more hybrid forms, they differ in their assessment of the quality of these forms. Hoppenstand would attribute the rise of parody to a form that was firmly established and

thoroughly rooted in a state of mass production. The preceding picture of the development of the genre shows that its roots are indeed firmly planted in popular fiction. Many of the texts most central to the network and most densely connected can be likened to penny dreadfuls, or the likes of *Vathek* written in a fugue state. While texts by Godwin, Brown, and Smollet are central, they arrive on the scene 40 years after Walpole's inciting text and more than 10 years after Radcliffe's early predecessors. This thesis therefore argues for a duality in the development of the genre. While a core selection of popular texts disseminated scenes, settings, characters, and constellations at a rapid pace, the genre was enriched by authors who adopted the mantle of the genre only for individual texts and as vehicles for other purposes. Todorov's assessment rings true: a genre that threads in the distinctly unreal allows for the expression of the true nature of literature, as a transgression of laws, whether social or narrative, to allow for a transition between different states.[238] External changes of state, brought about in times of social upheaval, have shaped a collection of modern texts deeply concerned with transitions—social, political, moral, stylistic, mental and often hypnagogic. While Godwin, Brown, and Mary Shelley took up the mantle of the unreal to express their ideas about the nature of the social decay they saw, they brought tropes, stylistic devices, and Romantic ideas of individualism, the primacy of passion, and self-determination to a genre that readily embraced them. Whereas the right to bodily autonomy had previously been the subject of murderous family feuds and expressed by actors in a constellation, as many texts in the tradition of Radcliffe or Walpole feature, it could now be openly proclaimed by individuals with conviction. The movement of development is thus twofold: outward from a core group of popular authors and inward from authors outside the tradition. Just as renowned writers of great character portraits like Austen and Dickens introduced the trends of a plot- and character-focused form, so too did the likes of Poe and Le Fanu introduce the form of short manic psychograms into a roster of actors struggling to integrate the psychological content they were mouthing. These examples thus support a duality between the movements in genre development proposed by Veselovsky and Moretti. While generic forms and static motifs are disseminated from the popular center outward, individuals at the periphery enrich the center with new free motifs.

[238]Todorov, *The Fantastic*, 166.

8 Conclusion:

This master's thesis has aimed to explore the structural patterns present in 182 Gothic Fiction texts. In order to address the potential hesitation with which more orthodox literary scholars might regard such an undertaking, this endeavor has not only contextualized the results of quantitative investigations within qualitative studies on the subject, but it has also traced various approaches to studying literary form in Formalism and Structuralism, as well as different conceptions of genre as a means of categorizing texts based on distinct attributes over time. Further light has been shed on the current tradition of literary investigations with computational means since the onset of Distant Reading, including the aim of rising to the challenge of addressing the hesitation they are presented with in the hopes of allowing for a conciliatory, shared body of research.

The initial chapters establish the methodological and contextual framework for the subsequent analysis. They outline various conceptions of the structure of fiction as a sum of interdependent elements. These conceptions share the belief that the composition of individual texts can be understood as a set of functional components. The frequency and characteristics of these components can serve as a basis for comprehending the essence of the texts as a whole.

> The 'reality' of a sequence lies not in the 'natural' succession of the actions composing it but in the logic there exposed, risked and satisfied. […] [T]he origin of a sequence is not the observation of reality, but the need to vary and transcend the first form given man, namely repetition.[239]

The likes of Yarkho, Tomashevsky, and Barthes have argued for the search for recurring textual patterns as a way of revealing meaning inherent in a given text. Lévi-Strauss goes so far as to call for the humanities to synchronize their methodology with the natural sciences. On these grounds it can be argued that mathematical methods of pattern recognition based on correlations, co-occurrences, and repetitions can reveal characteristics and frequencies in the structure of a text. It follows that the distribution of those attributes throughout a collection of texts allows for comparisons and an investigation of influence and relatedness throughout time, given a connection through similarity, proximity, and/or antecedence.

The formulaic nature of many works of Gothic fiction, with its consistent cast of archetypical characters and recurring themes, makes it well-suited for Structuralist investigations and Distant

[239]Barthes, "An Introduction to the Structural Analysis of Narrative," 271.

Reading analysis. This has also led to its frequent examination in these fields. The corpus in use in this study builds on existing research. One body of research, conducted by Caroline Winter and Eleanor Stribling, focuses on the color space used in texts of Gothic fiction. The other body of research, conducted by Ted Underwood, aims to differentiate Gothic fiction from science fiction and crime fiction. Other texts are drawn from Project Gutenberg's Gothic fiction shelf.

The Python programming language and its standard data and text processing packages, along with gutenbergpy, were used to retrieve and normalize the texts for the corpus. The corpus was stripped of all elements except for nouns, verbs, adjectives, and adverbs, and then chunked into 5000-word segments. In order to evaluate the model's fit, a numeric representation using word2vec was created. Extensive manual enrichment of the data with additional features was undertaken in order to later enable text clustering based on shared features. The subsequent modeling focused primarily on latent dirichlet allocation (LDA) in its Gensim implementation, The model was optimized using Word Embedding Coherence Centroid as the primary metric, with Word Embedding-Based Inverted Rank-Biased Overlap as an additional metric to balance a coherence metric against a diversity based metric. This was done to ensure optimal interpretability for subsequent human evaluation of the results. While comparisons with similar model types were made, LDA was the clear winner, in terms of lucidity and reusability.

The final exploration and interpretation were based on the topic distribution for each text segment, enriched with sentiment scores and the previously created features. Once the topics had undergone a lengthy process of evaluation to provide them with a human-readable label, numerous aggregations and comparisons of these features, with regard to their distribution of topics, were undertaken in order to examine the structural patterns inherent in the composing texts. The aggregations and visualizations allowed for the grouping of authors into segments of shared topical characteristics and changes in them throughout time. In Section 7.1, the overall composition of the topics is explored. Section 7.2 tracks the progression of the topic with the greatest variance over time. Section 7.3 looks for thematic uniqueness of certain features.

Last but not least, the topic distributions of the individual texts were used to create a set of network visualizations. Section 7.4 aggregates similarities in the topic distribution first bidirectionally and then unidirectionally from the oldest to the youngest texts. This is done in order to investigate how the trends of pioneers of the genre are passed on and which groups they have formed among themselves. A text that shared a cosine similarity above a certain threshold

with an older text was considered to be influenced by it. In order to mitigate the curse of dimensionality, only the ten most influential topics for each text were taken into account, while all others were discarded. The resulting clusters were interpreted and evaluated using the additional features at hand, and then compared to canonical opinions on the texts in question.

The movement of genre development was found to be twofold and supports a duality between the motions proposed by Veselovsky and Moretti. While generic forms, also known as static motifs, are disseminated from the popular center outward, individuals at the periphery enrich the center with new free motifs—themes, styles and topics—from the outskirts.

Ultimately, the results proved fruitful and promising, opening up new avenues of investigation for subsequent studies to come. The endeavor of combining 20th century perspectives on literature as atomic textual features with attributes of modern interpretations of a given genre and a budding new research tradition that strives to employ computational and statistical means has allowed for a conductive quantitative comparison of numerous competing conceptions on how genres are formed, while offering new insights into a collection of texts that served as the origin point of Gothic fiction. Future applications could include investigations into the directed links between genre clusters or comparisons of author-function signals.

Distant Reading has the potential to introduce quantitative measures into the study of literary form. Particularly in the fields of literary history, narratology, and genre studies, the object of research is so vast that it exceeds the amount of text readable by a single scholar. While traditional investigations in these fields have been limited to small samples, single authors, or canonical texts, efforts in recent decades have allowed for the formulation of new questions and investigations that challenge these constraints in response to technological developments. Prudent future digital and analog researchers will need to collaborate, prioritize data quality and validity of results, and redefine their beliefs about technical methodology and perspective on the object of inquiry to achieve ideal practices that invite further readings that enrich the discipline of literary criticism as a whole. Questions about the defining characteristics of literary genres are an integral part of literary criticism that can benefit from bridging the gap between what lies within the scope of what a single scholar can read in a lifetime and what they deem important to the object of study and can interpret with quantitative means. Reservations about the integration of quantitative and qualitative scholarship offer an opportunity for vigilance about academic rigor and the positioning of future interdisciplinary contributions. However, they also provide an

opportunity for future generations to reevaluate what has long been thought to be true about their subject, and to move from subjective and anecdotal findings on the shoulders of scholarship to an assessment that can include verification and frequency distributions of attributes. Statistical frameworks and computational resources are more accessible than ever, inviting researchers to broaden their horizons, challenge their beliefs, and expand their toolkits to explore perspectives that have long been impossible.

Sources:

Digital Sources

- Ted Underwood's metadata on texts of genre fiction – 'finalmeta.csv': https://github.com/tedunderwood/fiction
- The Gothic color space exploration of Caroline Winter, both in metadata and texts: https://github.com/CarolineWinter/gothic
- The gutenbergpy project: https://pypi.org/project/gutenbergpy/
- A number of texts directly from https://www.gutenberg.org/
- Python 3.9 and two separate environments for it, one for the modeling and one for everything else.
- The complete Python code inside 3 ipynb files, html renderings of it, all the raw, preprocessed or later enriched data, the created embeddings, models and graphs used, the lists of packages contained in 'requirements.txt' and 'requirements_02.txt' for the modeling section can be found on GitHub: https://github.com/f-klement/gothic-fiction-pattern-detection

Literature

Algee-Hewitt, Mark. "Distributed Character: Quantitative Models of the English Stage, 1550–1900." *New Literary History* 48, no. 4 (2017): 751–82. https://doi.org/10.1353/nlh.2017.0038.

Allison, Sarah, Heuser, Ryan, Jockers, Mathew, and Moretti, Franco. *Quantitative Formalism: An Experiment*. Stanford Literary Lab Pamphlets, 1, 2011.

Alt, Peter-André. "Theorien literarischer Evolution bei Šklovskij, Tynjanov und Mukařovsky" 21, no. 1–3 (January 1, 1986): 1–22. https://doi.org/10.1515/arca.1986.21.1-3.1.

Barthes, Roland. "An Introduction to the Structural Analysis of Narrative." Translated by Lionel Duisit. *New Literary History* 6, no. 2 (1975): 237–72. https://doi.org/10.2307/468419.

———. "The Imagination o f the Sign." In *Critical Essays*, 205–12. Evanston: Northwestern University Press, 1972.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks," 2009.

Bawarshi, Anis. "The Genre Function." *College English* 62, no. 3 (January 2000): 335–60.

Beausang, Chris. "A Brief History of the Theory and Practice of Computational Literary Criticism (1963-2020)." *Magazén*, no. 2 (December 22, 2020): 181–201. https://doi.org/10.30687/mag/2724-3923/2020/02/002.

Bergstra, J, D Yamins, and D D Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures," n.d.

Bergstra, James, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. "Algorithms for Hyper-Parameter Optimization." In *Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc., 2011. https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Bianchi, Federico, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. "Cross-Lingual Contextualized Topic Models with Zero-Shot Learning." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, edited by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, 1676–83. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.eacl-main.143.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (April 2012): 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David M, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

Botting, Fred. *Gothic*. London, New York: Routledge, 2005.

Broadwell, Peter M., David Mimno, and Timothy Tangherlini. "The Tell-Tale Hat: Surfacing the Uncertainty in Folklore Classification." *CA: Journal of Cultural Analytics* 2, no. 1 (February 18, 2017).

Brooke-Rose, Christine. "Historical Genres/Theoretical Genres: A Discussion of Todorov on the Fantastic." *New Literary History* 8, no. 1 (1976): 145–58.

Burke, Edmund. *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*. Columbia University Press, 2019.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information*

Processing Systems, Vol. 22. Curran Associates, Inc., 2009. https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html.

Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. "Termite: Visualization Techniques for Assessing Textual Topic Models." In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74–77. Capri Island Italy: ACM, 2012. https://doi.org/10.1145/2254556.2254572.

Csárdi, Gábor, Tamás Nepusz, Kirill Müller, Szabolcs Horvát, Vincent Traag, Fabio Zanini, and Daniel Noom. "Igraph for R: R Interface of the Igraph Library for Graph Theory and Network Analysis." Zenodo, February 20, 2024. https://doi.org/10.5281/zenodo.10681749.

Day, William Patrick. *In the Circles of Fear and Desire - A Study of Gothic Fantasy*. Chicago, London: University of Chicago Press, 1985.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. "Topic Modeling in Embedding Spaces." Edited by Mark Johnson, Brian Roark, and Ani Nenkova. *Transactions of the Association for Computational Linguistics* 8 (2020): 439–53. https://doi.org/10.1162/tacl_a_00325.

Eads, Alicia. "Separating the Wheat from the Chaff: A Topic and Keyword-Based Procedure for Identifying Research-Relevant Text." *Poetics (Amsterdam)* 86 (June 2021): 101527-.

Eco, Umberto. *The Name of the Rose*. Translated by William Weaver. London: Vingage, 2005.

Eder, Maciej. "Visualization in Stylometry: Cluster Analysis Using Networks." *Digital Scholarship in the Humanities* 32, no. 1 (April 2017): 50–64. https://doi.org/10.1093/llc/fqv061.

Eichenbaum, Boris Mikhailovich. "The Theory of the 'Formal Method.'" In *The Norton Anthology of Theory and Criticism*, 1062–80, 2001. https://www.semanticscholar.org/paper/The-theory-of-the-%E2%80%9CFormal-Method%E2%80%9D-Mikhailovich/553cc9b53b6a76d03354992e323919760333965f.

El-Assady, Mennatallah, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. "Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework." *IEEE Transactions on Visualization and Computer Graphics* 24, no. 1 (January 2018): 382–91. https://doi.org/10.1109/TVCG.2017.2745080.

Erlich, Victor. "XI. Literature and 'Life' - Formalist and Structuralist Views." In *Russian Formalism*, 192–211. De Gruyter Mouton, 2012. https://doi.org/10.1515/9783110873375.192.

Erlin, Matt. "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731–1864." In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 55–90. Boydell & Brewer, 2014.

Etherington, Ben. "World Literature as a Speculative Literary Totality: Veselovsky, Auerbach, Said, and the Critical-Humanist Tradition." *Modern Language Quarterly* 82, no. 2 (June 1, 2021): 225–51. https://doi.org/10.1215/00267929-8899139.

Fishelov, David. "The Strange Life and Adventures of Biological Concepts in Genre Periodization." Accessed July 9, 2023. https://www.academia.edu/17874660/The_Strange_Life_and_Adventures_of_Biological_Concepts_in_Genre_Periodization.

Francis, Elizabeth. "Appropriating the Theory: Structuralism and Children's Literature." *Children's Literature Association Quarterly* 7, no. 3 (1982): 52–58. https://doi.org/10.1353/chq.0.0153.

Free, William J. “Structuralism, Literature, and Tacit Knowledge.” *Journal of Aesthetic Education* 8, no. 4 (1974): 65–74. https://doi.org/10.2307/3332029.

Garcia Landa, Jose Angel. “The Structure of the Fabula (II): Boris Tomashevski, ‘Thematics’; Mieke Bal, ‘Narratology’ (Narrative Theory, 2).” *SSRN Electronic Journal*, 1990. https://doi.org/10.2139/ssrn.2693261.

Gasparov, Mikhail. “Boris Yarkho’s Works on Literary Theory.” Translated by Lavery, Michael. *Studia Metrica et Poetica* 3, no. 2 (December 31, 2016): 130–50. https://doi.org/10.12697/smp.2016.3.2.05.

Gius, Evelyn, and Janina Jacke. “Are Computational Literary Studies Structuralist?” *Journal of Cultural Analytics* 7, no. 4 (December 1, 2022). https://doi.org/10.22148/001c.46662.

Goodlad, Lauren M. E. “A Study in Distant Reading: Genre and the Longue Durée in the Age of AI.” *Modern Language Quarterly (Seattle)* 81, no. 4 (December 2020): 491–525.

Greenwood, David. *Structuralism and the Biblical Text*. Mouton Publishers, 1985. https://www.biblio.com/book/structuralism-biblical-text-greenwood-david/d/951947757.

Grün, Bettina, and Kurt Hornik. “Topicmodels: An R Package for Fitting Topic Models.” *Journal of Statistical Software* 40 (May 9, 2011): 1–30. https://doi.org/10.18637/jss.v040.i13.

Hagberg, Aric A, Daniel A Schult, and Pieter J Swart. “Exploring Network Structure, Dynamics, and Function Using NetworkX,” 2008.

Hammond, Adam. “The Double Bind of Validation: Distant Reading and the Digital Humanities’ ‘Trough of Disillusionment.’” *Literature Compass* 14, no. 8 (August 2017): 1–13.

Hettinger, Lena, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. “Genre Classification on German Novels.” In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 249–53. Valencia, Spain: IEEE, 2015. https://doi.org/10.1109/DEXA.2015.62.

Heuser, Ryan. “Mapping the Emotions of London in Fiction, 1700–1900: A Crowdsourcing Experiment.” In *Literary Mapping in the Digital Age*, 43–64, 2016.

Holland, Kate. “Narrative Tradition on the Border: Alexander Veselovsky and Narrative Hybridity in the Age of World Literature.” *Poetics Today* 38, no. 3 (September 1, 2017): 429–51. https://doi.org/10.1215/03335372-4166647.

Hoppenstand, Gary. “Genres and Formulas in Popular Literature.” In *A Companion to Popular Culture*, 101–22. John Wiley & Sons, Ltd, 2016. https://doi.org/10.1002/9781118883341.ch7.

Hou, Renkui, and Minghu Jiang. “Analysis on Chinese Quantitative Stylistic Features Based on Text Mining.” *Digital Scholarship in the Humanities* 31, no. 2 (June 2016): 357–67. https://doi.org/10.1093/llc/fqu067.

Hutto, C., and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1 (May 16, 2014): 216–25. https://doi.org/10.1609/icwsm.v8i1.14550.

Isbell, John Clairborne. *An Outline of Romanticism in the West*. Cambridge: Open Book Publisher, 2022.

Jacke, Janina. “Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology.” *Journal of Literary Theory* 8 (June 1, 2014). https://doi.org/10.1515/jlt-2014-0005.

Jannidis, Fotis. “Perspektiven empirisch-quantitativer Methoden in der Literaturwissenschaft — ein Essay.” *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 89, no. 4 (December 2015): 657–61. https://doi.org/10.1007/BF03396503.

Jauss, Hans Robert, and Elizabeth Benzinger. “Literary History as a Challenge to Literary Theory.” *New Literary History* 2, no. 1 (1970): 7–37. https://doi.org/10.2307/468585.

Jenkins, Tricia. “The History and Logic of Genre Study.” In *A Companion to Popular Culture*, 83–100. John Wiley & Sons, Ltd, 2016. https://doi.org/10.1002/9781118883341.ch6.

Jockers, Matthew. “Syuzhet,” 2015. https://github.com/mjockers/syuzhet.

Jockers, Matthew L. *Macroanalysis*. Urbana, Chicago: University of Illinois Press, 2013.

Jockers, Matthew L., and David Mimno. “Significant Themes in 19th-Century Literature.” *Poetics* 41, no. 6 (December 2013): 750–69. https://doi.org/10.1016/j.poetic.2013.08.005.

Johann Peter Eckermann. “Gespräche Mit Goethe in Den Letzten Jahren Seines Lebens.” Projekt Gutenberg, 09 1827. https://www.projekt-gutenberg.org/eckerman/gesprche/gsp1087.html.

Kelly, Gary. “Unbecoming a Heroine: Novel Reading, Romanticism, and Barrett’s The Heroine.” *Nineteenth-Century Literature* 45, no. 2 (September 1990): 220–41. https://doi.org/10.2307/3045125.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schutze. “Automatic Detection of Text Genre.” In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Madrid, Spain: Association for Computational Linguistics, 1997. https://doi.org/10.3115/976909.979622.

Kuhn, Jonas. “Computerlinguistische Textanalyse in Der Literaturwissenschaft? Oder: »The Importance of Being Earnest« bei Quantitativen Untersuchungen.” In *Quantitative Ansätze in Den Literatur- Und Geisteswissenschaften*, edited by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, 11–44. De Gruyter, n.d.

Lau, Jey Han, David Newman, and Timothy Baldwin. “Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality.” In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, edited by Shuly Wintner, Sharon Goldwater, and Stefan Riezler, 530–39. Gothenburg, Sweden: Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/E14-1056.

Lemon, Lee. “Boris Eichenbaum.” In *Russian Formalist Criticism: Four Essays, Second Edition*, 65–86. Lincoln: University of Nebraska Press, 2012. https://web-p-ebscohost-com.uaccess.univie.ac.at/ehost/ebookviewer/ebook/bmxlYmtfXzE4NzU0OTRfX0FO0?sid=dbb3129b-a965-467a-95c5-4c4f87caa7fe@redis&vid=0&format=EK&lpid=p02-ch01&rid=0.

———. “Boris Tomashevsky.” In *Russian Formalist Criticism: Four Essays, Second Edition*, 46–64. Lincoln: University of Nebraska Press, 2012. https://web-p-ebscohost-com.uaccess.univie.ac.at/ehost/ebookviewer/ebook/bmxlYmtfXzE4NzU0OTRfX0FO0?sid=dbb3129b-a965-467a-95c5-4c4f87caa7fe@redis&vid=0&format=EK&lpid=p02-ch01&rid=0.

Lévi-Strauss, Claude. “The Structural Study of Myth.” *The Journal of American Folklore* 68, no. 270 (1955): 428–44. https://doi.org/10.2307/536768.

Lvoff, Basil. “Distant Reading in Russian Formalism and Russian Formalism in Distant Reading.” *Russian Literature* 122–123 (May 2021): 29–65.

MacKenzie, Ian. “Narratology and Thematics.” *Modern Fiction Studies* 33, no. 3 (1987): 535–44.

Maslov, Boris, and Ilya Kliger. *Persistent Forms: Explorations in Historical Poetics*. New York: Fordham University Press, 2016.

Merrill, Jessica. “Distant Reading in Russia: Franco Moretti and the Formalist Tradition,” n.d.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” arXiv, September 6, 2013. http://arxiv.org/abs/1301.3781.

Mimno, David. “Computational Historiography.” *Journal on Computing and Cultural Heritage* 5, no. 1 (April 2012): 1–19.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. “Optimizing Semantic Coherence in Topic Models.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–72. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011. https://aclanthology.org/D11-1024.

Moretti, Franco. *Distant Reading*. London, New York: Verso, 2013.

———. *Distant reading*. London : Brooklyn, NY. Accessed February 20, 2023. https://usearch.uaccess.univie.ac.at/primo-explore/fulldisplay/UWI_alma21325986530003332/UWI.

———. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London, New York: Verso, 2007.

Murray, Alex, ed. *Decadence: A Literary History*. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108640527.

Newman, David J., and Sharon Block. “Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper.” *Journal of the American Society for Information Science and Technology* 57, no. 6 (April 2006): 753–67. https://doi.org/10.1002/asi.20342.

Nikitina, Natalia, and Natalia Tuliakova. “Genre Studies in Russian Literary Research: Achievements and Challenges.” *Interlitteraria* 25, no. 2 (December 31, 2020): 319–31. https://doi.org/10.12697/IL.2020.25.2.5.

Parisot, Eric. “The Aesthetics of Terror and Horror: A Genealogy.” In *The Cambridge History of the Gothic: Volume 1: Gothic in the Long Eighteenth Century*, edited by Angela Wright and Dale Townshend, 1:284–303. The Cambridge History of the Gothic. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108561044.014.

Piper, Andrew. *Enumerations - Data and Literary Study*. Chicago, London: University of Chicago Press, 2018.

Propp, V. “Morphology of the Folktale,” n.d.

Punter, David. *The Literature of Terror - A History of Gothic Fictions from 1765 to the Present Day - Volume 1*. London, New York: Routledge, 2013.

———. *The Literature of Terror - A History of Gothic Fictions from 1765 to the Present Day - Volume 2*. London, New York: Routledge, 2013.

Punter, David, and Glennis Byron. *The Gothic*. Blackwell Publishing, 2004.

Radcliffe, Ann Ward. *The Mysteries of Udolpho, a Romance*. Vol. 1. Dublin: printed by Thomas Burnside. For P. Wogan, No. 23, Old Bridge, 1800. https://link.gale.com/apps/doc/CW0113877622/ECCO?sid=bookmark-ECCO&xid=56792457.

Ramage, Daniel, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. “Topic Modeling for the Social Sciences.” *Advances in Neural Information Processing Systems* 22 (2009): 1–4.

Ramsay, Stephen. *Reading Machines: Toward and Algorithmic Criticism*. Champaign, UNITED STATES: University of Illinois Press, 2011. http://ebookcentral.proquest.com/lib/univie/detail.action?docID=3413843.

Řehůřek, Radim, and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora.” In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA, 2010.

Riddell, Allen Beye. “How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models.” In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 91–114. Boydell & Brewer, 2014.

Roque, Antonio. “Towards a Computational Approach to Literary Text Analysis.” In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 97–104. Montréal, Canada: Association for Computational Linguistics, 2012. https://aclanthology.org/W12-2514.

Rosner, Frank, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. “Evaluating Topic Coherence Measures.” arXiv, March 25, 2014. http://arxiv.org/abs/1403.6397.

Saccenti, Edoardo, and Leonardo Tenori. “Multivariate Modeling of the Collaboration between Luigi Illica and Giuseppe Giacosa for the Librettos of Three Operas by Giacomo Puccini.” *Digital Scholarship in the Humanities* 30, no. 3 (September 2015): 405–22. https://doi.org/10.1093/llc/fqu006.

Sason, Igal. *Divergence Measures: Mathematical Foundations and Applications in Information-Theoretic and Statistical Problems*. Vol. 24, 2022. https://www.mdpi.com/1099-4300/24/5/712.

Schmidt, Benjamin M. “Plot Arceology: A Vector-Space Model of Narrative Structure.” In *2015 IEEE International Conference on Big Data (Big Data)*, 1667–72. Santa Clara, CA, USA: IEEE, 2015. https://doi.org/10.1109/BigData.2015.7363937.

Schöch, Christof. “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.” arXiv, March 24, 2021. https://doi.org/10.48550/arXiv.2103.13019.

Schofield, Alexandra. “Comparing Apples to Apple: The Effects of Stemmers on Topic Models.” *Transactions of the Association for Computational Linguistics* 4 (December 2016): 287–300.

Schofield, Alexandra Kathryn. “Text Processing for the Effective Application of Latent Dirichlet Allocation.” Ph.D., Cornell University. Accessed April 18, 2023. https://www.proquest.com/docview/2242440138/abstract/F3B8757D33DD47CBPQ/1.

Sievert, Carson, and Kenneth Shirley. “LDAvis: A Method for Visualizing and Interpreting Topics.” In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/W14-3110.

Simpson, Daivd. “Romanticism, Criticism and Theory.” In *The Cambridge Companion to British Romanticism*, edited by Curran Stuart, 1–24. Cambridge: Cambridge University Press, 1993.

Slingerland, Edward, Ryan Nichols, Kristoffer Neilbo, and Carson Logan. “The Distant Reading of Religious Texts.” *Journal of the American Academy of Religion* 85, no. 4 (2017): 985–1016.

Smithson, Isaiah. “Structuralism as a Method of Literary Criticism.” *College English* 37, no. 2 (1975): 145–59. https://doi.org/10.2307/375060.

Tangherlini, Timothy R., and Peter Leonard. “Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research.” *Poetics (Amsterdam)* 41, no. 6 (2013): 725–49.

Terragni, Silvia, and Elisabetta Fersini. “OCTIS 2.0: Optimizing and Comparing Topic Models in Italian Is Even Simpler!” In *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-It 2021*, edited by Elisabetta Fersini, Marco Passarotti,

and Viviana Patti, 328–34. Accademia University Press, 2022. https://doi.org/10.4000/books.aaccademia.10863.

Terragni, Silvia, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. "OCTIS: Comparing and Optimizing Topic Models Is Simple!" In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–70. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.eacl-demos.31.

Terragni, Silvia, Elisabetta Fersini, and Enza Messina. "Word Embedding-Based Topic Similarity Measures." In *Natural Language Processing and Information Systems*, edited by Elisabeth Métais, Farid Meziane, Helmut Horacek, and Epaminondas Kapetanios, 12801:33–45. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-80599-9_4.

Todorov, Tzvetan. *The Fantastic - A Structural Approach to a Literary Genre*. Cleveland, London: Press of Case Western Reserve University, 1973.

Uglanova, Inna, and Evelyn Gius. "The Order of Things. A Study on Topic Modelling of Literary Texts," n.d.

Underwood, Ted. *Distant Horizons - Digital Evidence and Literary Change*. Chicago, London: University of Chicago Press, 2019.

———. "The Life Cycles of Genres." *Journal of Cultural Analytics*, May 23, 2016, 1–25.

———. "Topic Modeling Made Just Simple Enough." Blog. *The Stone and the Shell* (blog), April 7, 2012. https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/.

Underwood, William. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11, no. 2 (2017). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Walpole, Horace. *The Castle of Otranto: A Gothic Novel*. Waiheke Island: The Floating Press, 2009. https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=313943&site=ehost-live.

Wilkerson, John, and Andreu Casas. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20, no. 1 (2017): 529–44. https://doi.org/10.1146/annurev-polisci-052615-025542.

Yarkho, Boris. "The Elementary Foundations of Formal Analysis." Translated by Igor Pilshchikov and Michael Lavery. *Studia Metrica et Poetica*, January 1, 2016. https://www.academia.edu/30979087/Boris_Yarkho_The_Elementary_Foundations_of_Formal_Analysis_.

Yarkho, Boris I. "Speech Distribution in Five-Act Tragedies (A Question of Classicism and Romanticism)." *Journal of Literary Theory* 13, no. 1 (March 1, 2019): 13–76. https://doi.org/10.1515/jlt-2019-0002.

Zherebin, Alexej. "Aleksandr Veselovskijs Konzept der Historischen Poetik und sein Echo im kulturwissenschaftlichen Diskurs Russlands und Deutschlands." *Journal of Integrative Cultural Studies* 1, no. 1 (July 22, 2019): 5–14. https://doi.org/10.33910/2687-1262-2019-1-1-5-14.

Appendix:

List of Topics & Additional Graphs

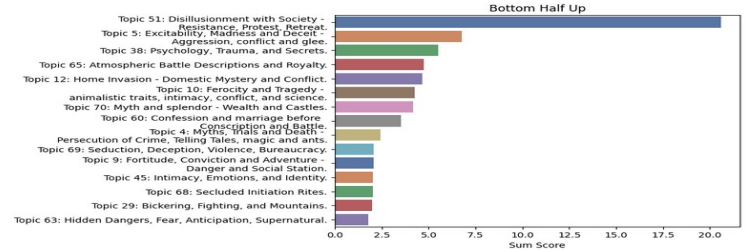
1: Ominous Atmosphere - Spatial and Auditory Imagery: vastness, archaic, Refinement, Gloom, demons.
2: Emotional Dialogue - Fear, Secrecy, Flattery, Arousal and Strife - Religion and Devils.
3: Status and Individuality - Striving, Misery and Plentifulness – Excess.
4: Myths, Trials and Death - Persecution of Crime, Telling Tales, magic and ants.
5: Excitability, Madness and Deceit - Aggression, conflict and glee.
6: Nature and Reasoning - Creativity, Understanding, mixed with Fauna.
7: Social Pleasantries - Diplomacy, Plotting to Gossip.
8: Faith, Convictions, Chivalry and Death - Erudition, Religion and Knights. Ants.
9: Fortitude, Conviction and Adventure - Danger and Social Station.
10: Ferocity and Tragedy - animalistic traits, intimacy, conflict, and science.
11: Ravens and Gloom - Longing, Death and Artifice.
12: Home Invasion - Domestic Mystery and Conflict.
13: Rituals and Festivities - Dance, Witchcraft and Coronations.
14: Conflict, Animosity and Change - Emotional Changes, Death and Construction.
15: Trickery and Science - Deceit, Reasoning and Institutions.
16: Desecrated Chapel - Confessions and Defilement - Devils and Maniacs.
17: (Un-)death, spectral bodies and judgment - human physicality, grief, emotions.
18: Mystery and Adversity - Dream and fugue states, Investigation.
19: Forlorn Carnival - Dances, Disgust and Intimacy.
20: Science, Reasoning and Objects - Technology, Professions and Nature.
21: War, Punishment, and Exploration.
22: Emotional Dynamics and Interactions.
23: War, dreams and demons.
24: Human Interactions and Emotional States.
25: Flattery, Clothing, Interactions.
26: Witchcraft, Rituals, and Fear of it - Banishment, Threats, and Armor.
27: Dragon Attack and Defense - Troops, Mountains and Cynicism.
28: Communion in Nature - Transformation, Relationships and Identity.
29: Bickering, Fighting, and Mountains.
30: Bureaucracy, Bargaining and Dissatisfaction.
31: Exploration, Gloom, Caverns.
32: Tranquility and Bustle - Terms of Relaxation, Calm and Action.
33: Treacherous Company - on the run and scarred.
34: Secrets and Suspense - Mystery, Devils and Assassinations.
35: Mental Illness, Law and Outcasts - Fear, Suspicion and Struggles.
36: Individualism vs. Conformity - Rebellion and Social Norms.
37: Order and Chaos - Constrained Focus and Unchecked Emotions.
38: Psychology, Trauma, and Secrets.
39: Quest for Meaning - Self-Discovery, Transformation.
40: Ambition and Struggle - Emotional Turmoil.
41: Despair, Isolation and Oppression.
42: Illusion, Enchantment and Betrayal.
43: Woodlands, Mystery, Illusion, Beasts.
44: Companionship in Times of Trial and Distress.
45: Intimacy, Emotions, and Identity.
46: Frustration, Society, Retreat into Nature - Society, Reason, Tension, negative Feelings, Forests.
47: Human Nature and the Connection to the Land, Myth and (Human) Nature - Solace, Inspiration, Acceptance for Hardships.
48: Enthralling Garden full of Voices - Enchantment and Vocalization, Nature.
49: Departure and Music.
50: Myth, Nature, Wonder, and Despair.

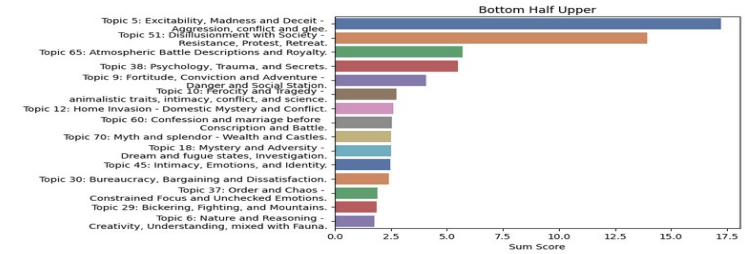
51: Disillusionment with Society - Resistance, Protest, Retreat.
52: Adventure, Splendor, Power and Challenges, History.
53: Mercantile and Creativity - Haggling and Emotions.
54: Medieval Cities, Castles and Courtship.
55: Crocodiles, Massacres and Traveling.
56: Exploration of an Island and Obsession.
57: Carnage near a Castle.
58: Weddings and Rituals - Clamoring Throng.
59: Judgment and Scrutiny - Tense Diplomacy.
60: Confession and marriage before Conscription and Battle.
61: Vampires, Regality, Experiments, Festivities and Sacrifice.
62: Dragons, Subterranean Lairs, Riddles and Lore.
63: Hidden Dangers, Fear, Anticipation, Supernatural.
64: Artistic Ambition and Trials - Mastery and the Devil.
65: Atmospheric Battle Descriptions and Royalty.
66: Hidden Knowledge, Learning and Secrets.
67: Monsters, Art, Romance - Myth and Gloom.
68: Secluded Initiation Rites.
69: Seduction, Deception, Violence, Bureaucracy.
70: Myth and Splendor - Wealth and Castles.
71: Haunted Castles and their Prophecies.
72: Festivities, Noise, Crowds.
73: Camps, Trenches and Weather.

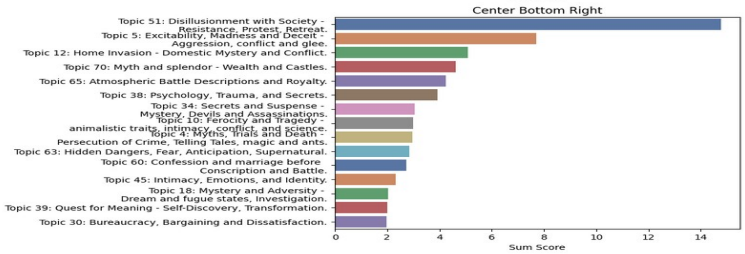
7.2 – Most important topics for the most influential authors, for more details, see the notebook

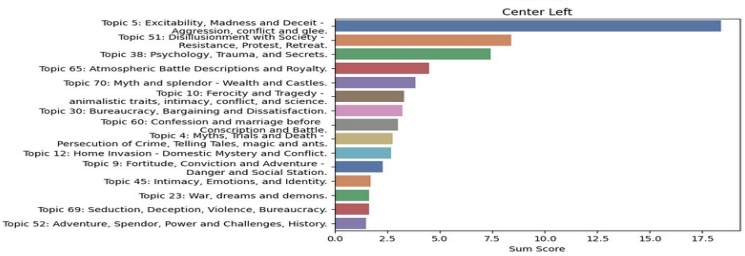
Author	Sum Topics	Median Topics
Hawthorne, Nathaniel	70, 3, 65, 56, 12	70, 52, 35, 3, 29
Corelli, Marie	51, 70, 29, 42, 31	51, 29, 42, 70, 31
Le Fanu, Sheridan	70, 5, 51, 60, 65	70, 51, 5, 12, 39
Poe, Edgar Allan	10, 28, 44, 4, 9	4, 70, 10, 9, 7
Wharton, Edith	9, 65, 5, 51, 34	65, 5, 70, 51, 34
Blackwood, Algernon	51, 5, 18, 9, 70	51, 5, 18, 12, 70
Radcliffe, Ann	5, 14, 51, 38, 67	5, 38, 51, 70, 9
Shelley, Mary	5, 51, 38, 65, 66	5, 38, 51, 65, 9
Stoker, Bram	12, 70, 65, 61, 5	70, 65, 51, 5, 12
Smith, Charlotte	3, 65, 38, 5, 51	38, 51, 5, 37, 34
Lee, Vernon	22, 70, 60, 51, 65	51, 52, 7, 12, 43
Bierce, Ambrose	49, 10, 39, 32, 69	10, 69, 29, 12, 5
Scott, Walter	54, 5, 65, 51, 12	5, 65, 51, 12, 60
Kipling, Rudyard	65, 31, 8, 18, 28	65, 5, 8, 18, 45
Machen, Arthur	51, 12, 5, 65, 70	51, 12, 5, 70, 65
Ainsworth, William Harrison	35, 34, 5, 70, 37	5, 51, 42, 35, 38
Chambers, Robert William	5, 46, 12, 51, 10	5, 51, 10, 46, 6
Gaskell, Elizabeth	5, 51, 73, 52, 70	5, 51, 70, 12, 52
Lytton, Edward Bulwer Lyt	51, 38, 5, 12, 66	51, 5, 38, 12, 65
Brown, Charles Brockden	51, 38, 5, 65, 10	5, 51, 38, 10, 65

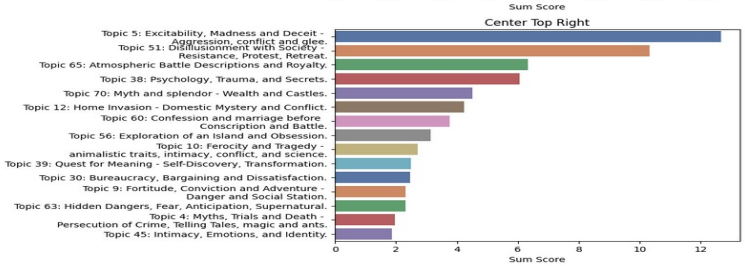
Top 15 Topics per Cluster

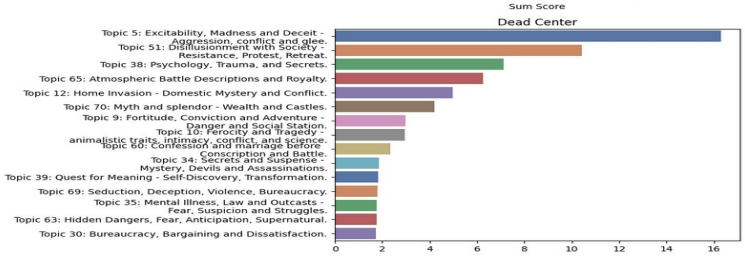


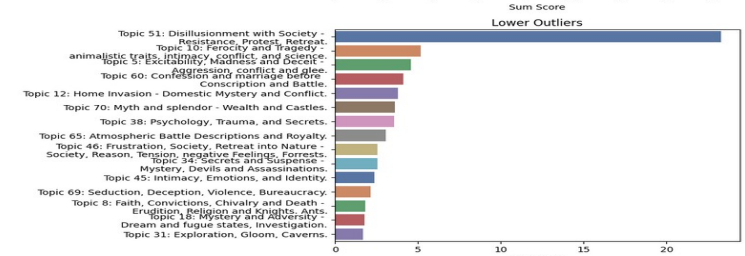


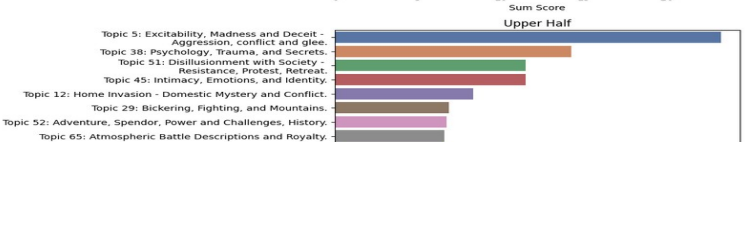


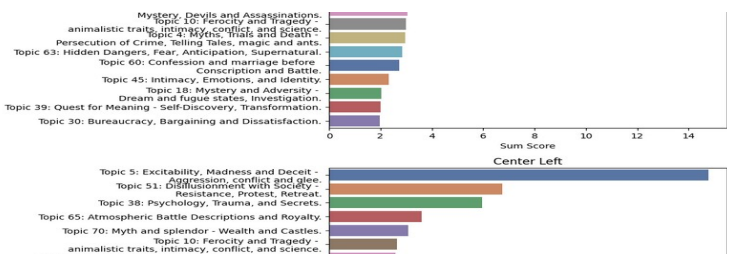


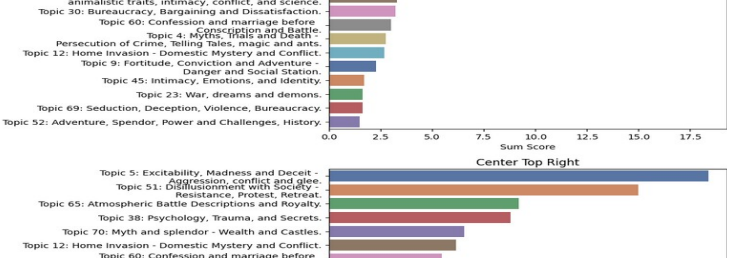


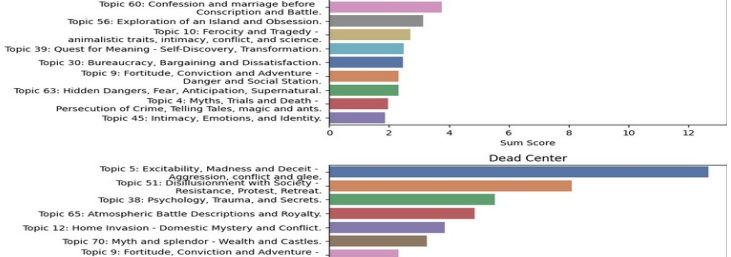


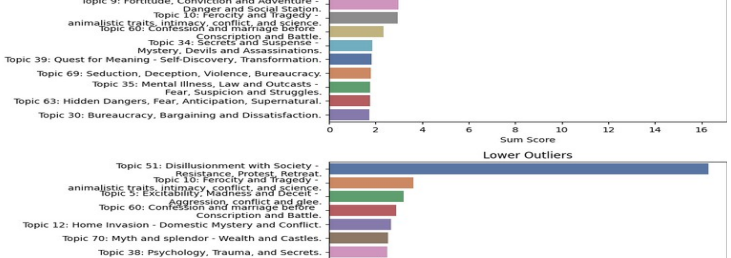


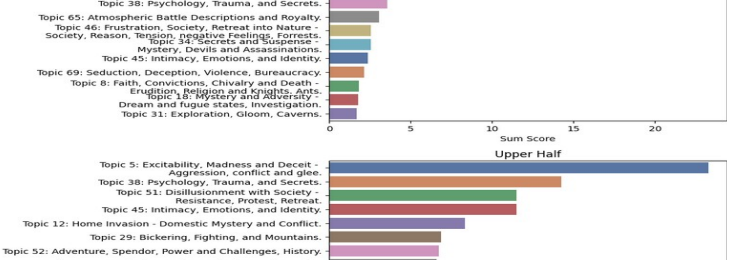


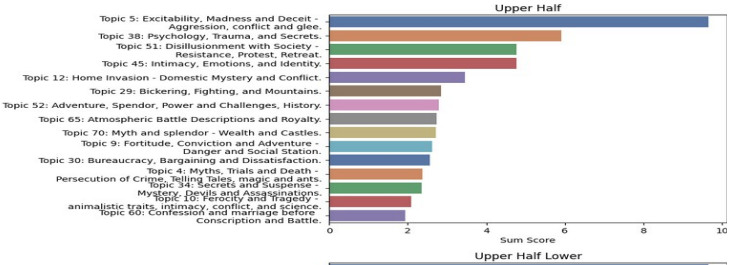


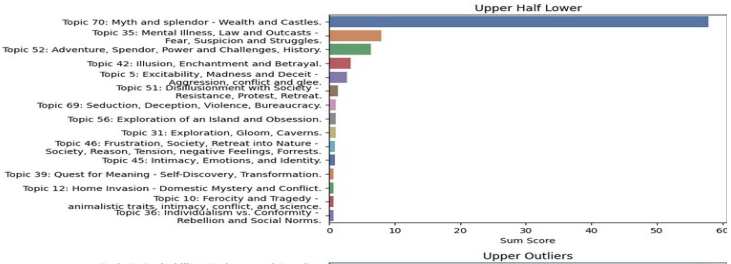


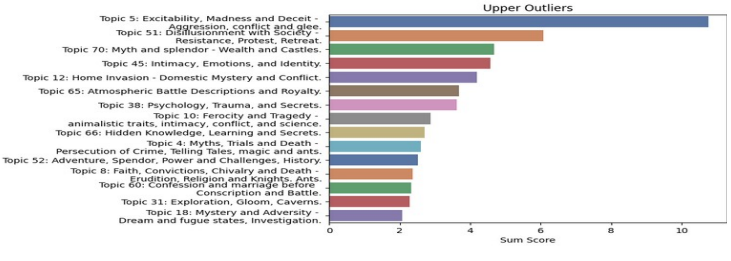












Abstract

The purpose of this master's thesis is to apply natural language processing (NLP) methods to a corpus of 182 Gothic fiction texts in order to gain insight into the genre's composition and explore its early influences in a network. To this end various computational approaches had been employed including machine learning models, exploratory data analysis and clustering, yet focusing on network analysis, and topic modeling. The results are aggregated and compared across different categories as well as throughout time. The consistent stock of characters and motifs within Gothic Fiction, lends itself well for an analysis of its recurrent composing elements. This quality had made it a frequent topic of explorations within traditional schools of literary criticism, such as Structuralism and Formalism. This thesis aims to expand on previous approaches with the tool set of the Digital Humanities and Distant Reading, taking a more quantitative perspective, while arguing for the necessity of closer future collaboration between digital and analogue research in the humanities, in order to enable new avenues of investigation.

Keywords: Distant Reading, genre studies, Gothic Fiction, natural language processing, computational literary studies, Structuralism, Formalism, network analysis, topic modeling

Abstract (German)

Die vorliegende Masterarbeit befasst sich mit der Anwendung von Natural Language Processing (NLP)-Methoden auf einen Korpus von 182 Gothic-Fiction-Texten. Ziel ist die Gewinnung neuer Einblicke in die Zusammensetzung des Genres und dessen Entstehungsgeschichte, dargestellt als ein Netzwerk der Beeinflussung. Zu diesem Zweck wurden diverse computergestützte Verfahren und maschinelle Lernmethoden, wie das Clustering und die explorative Datenanalyse, angewendet. Der Schwerpunkt liegt dabei auf Netzwerkanalysen und Topic Modeling. Die Ergebnisse wurden aggregiert und jeweils über einen Zeitverlauf sowie über verschiedene Kategorien hinweg verglichen. Das statische Inventar an Motiven und Charakteren des Genres Gothic Fiction begünstigt die Analyse wiederkehrender Elemente und Strukturen. Diese Eigenschaft führte dazu, dass das Genre Gegenstand zahlreicher traditioneller literaturwissenschaftlicher Untersuchungen wurde. Diese Masterarbeit fokussiert sich insbesondere auf die Distant Reading-Ansätze, welche als Verängerung und technsiche Erweiterung der Schulen des Strukturalismus und Formalismus dargestellt werden. Dabei werden sowohl die Ähnlichkeiten als auch die Unterschiede zwischen diesen Ansätzen herausgearbeitet. Zudem wird untersucht, inwiefern sich die Ziele dieser Ansätze mit den quantitativen Forschungsmethoden der Digital Humanities erreichen lassen. In diesem Zusammenhang wird auch die Frage erörtert, wie eine künftige enge Verzahnung analoger und digitaler literaturwissenschaftlicher Forschung mit neuen Denkansätzen den Diskurs zu makroperspektischen Strukturen bereichern kann.

Schlagworte: Distant Reading, Genreforschung, Gothic Fiction, Natural Language Processing, computergestützte Literaturwissenschaft, Strukturalismus, Formalismus, Netzwerkanalysen, Topic Modeling