ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN SPECIALIZED TRANSLATION

.....

CLUES IN THE TEXT: STYLISTIC SIGNALS OF SUCCESS IN EARLY DETECTIVE FICTION

Tesi di laurea magistrale in Corpus Linguistics

Relatrice
Prof. Silvia Bernardini

Correlatrice
Prof. Maja Milicevic Petrovic

Presentata da
Carlo Alfonso Bellini

Sessione Marzo 2026

Anno Accademico 2024/2025

Table of Contents

Abstract

This thesis investigates whether measurable stylistic features are associated with literary "success" in late nineteenth- and early twentieth-century detective fiction, and whether those features can support prediction within a controlled corpus. The study analyzes 281 English-language detective fiction works (1880–1935) from Project Gutenberg. Success is operationalized as contemporary public-domain engagement using a fixed snapshot of download counts (20 December 2025): texts in the top decile are labeled "successful," and the remainder form the comparison group.

Situated in digital humanities and computational stylistics, the research adopts a compact, interpretable feature set capturing broad dimensions of narrative language: lexical concreteness, affective arousal, narrative stance (first- and third-person pronoun density), and syntactic packaging (subordination density). A lexical domain analysis further examines how concreteness is employed in semantic fields related to perception, inference, and procedural detection, contextualized through genre expectations often discussed via the Holmesian tradition.

Empirical evaluation combines regression modelling of download intensity with an exposure control for time since Project Gutenberg release and supervised classification (logistic regression, random forest, and a feed-forward neural network) with conservative cross-validation and held-out testing. Results indicate that stylistic properties tied to emotional activation, lexical abstraction, and syntactic complexity are meaningfully associated with the download-based proxy of success, and that simple, interpretable models can achieve above-chance predictive performance, though conclusions are tempered by class imbalance and the small number of successful texts.

The thesis concludes by summarizing contributions, limitations, and directions for extending genre-controlled studies of style and reader attention.

Introduction

In an age where cultural attention is increasingly quantified, through platform metrics, recommendation systems, and digital libraries, questions that once belonged primarily to literary criticism are now also being asked with data. Why do some books remain widely read while others fade from view? Which properties of a text are plausibly connected to sustained public attention, and can those properties be measured in a way that supports systematic comparison across many works? Within the digital humanities and computational stylistics, these questions have motivated a growing body of research that treats literary success not as a purely subjective label, but as a phenomenon that leaves observable traces in circulation, reception, and language.

At the same time, defining “success” in literature is notoriously difficult. Commercial sales, critical reputation, institutional endorsement, canon formation, and long-term cultural persistence often diverge, and each captures a different facet of what it means for a work to “matter.” Recent quantitative approaches have therefore adopted pragmatic operationalizations grounded in available evidence—bestseller lists, library holdings, reviews, or digital engagement—while also emphasizing that such measures reflect attention dynamics as much as intrinsic quality. This thesis builds on that line of work by asking a deliberately narrower version of the broader problem: within a controlled genre and historical window, do certain high-level stylistic properties tend to characterize texts that attract disproportionate present-day readership attention?

The study focuses on English-language detective fiction from 1880 to 1935, a period that covers the consolidation of detective fiction as a mass-market genre and closely relates with the publication window of Arthur Conan Doyle’s Sherlock Holmes. Restricting the dataset to a single genre and timeframe serves a methodological purpose: it limits diachronic drift and reduces the likelihood that observed differences are simply artifacts of broad genre contrasts. Conan Doyle plays a central role in the thesis not because the project is an author study in the traditional sense, but because Holmes provides a widely recognized prototype of the genre, one in which observation is transformed into inference within an evidential and institutional frame. This “Holmesian” baseline helps motivate the corpus delimitation and interpretive framing, while the feature set and modelling choices are grounded in prior work on style and reader engagement rather than derived from Doyle alone.

It is important to note that Arthur Conan Doyle functions as more than a historical reference point in this thesis: he provides a conceptual and operational baseline for the object of study. Holmes narratives offer a widely recognized prototype of late nineteenth-century detective

fiction, foregrounding detection as the transformation of observation into inference within an institutional and evidential frame. The analyses developed in this thesis take this “Holmesian” configuration as a motivated starting point for interpretive framing and corpus delimitation, while the statistical and modelling stages rely on a compact set of five variables motivated by prior computational and stylistic research rather than derived from Doyle alone. In this sense, Doyle anchors the thesis at the levels of genre salience, temporal scope, and interpretation, without functioning as the direct source of the feature set.

Empirically the thesis draws on a corpus of 281 detective fiction works by 142 authors obtained from Project Gutenberg, using its machine-readable catalog to ensure reproducibility. To approximate present-day engagement with public-domain texts, literary success is operationalized through Project Gutenberg download counts at a fixed catalog snapshot (20 December 2025). Given the heavy-tailed distribution of attention measures, success is defined using a percentile threshold as follows: works in the top 10% of the corpus by downloads are labelled “successful,” while the remaining works form the comparison group. This criterion is explicitly interpreted as a corpus-internal proxy for contemporary public-domain readership attention, not as a direct measure of historical sales, critical merit, or aesthetic value.

Rather than treating style as an open-ended inventory of thousands of features, the thesis adopts a compact set of theory-driven, interpretable variables intended to capture macro-level dimensions of narrative style. Lexical choice is represented through semantic concreteness, indexing the extent to which a text favors tangible, perceptually grounded language over abstraction. Affective engagement is captured through arousal, interpreted as emotional activation independent of positive or negative valence. Narrative stance is approximated through the density of first- and third-person pronouns, reflecting differences in focalization and voice. Finally, syntactic organization is represented through subordination density, used as a proxy for clausal embedding and structural complexity. While a wider inventory is extracted for transparency, these five variables serve as the main explanatory and predictive backbone of the analyses.

The thesis is organized around two main research questions:

Is it possible to systematically differentiate “successful” from “other” detective fiction texts in the period of Arthur Conan Doyle using only linguistic and stylistic features? And to what extent can supervised computational models trained on extracted stylistic features predict success in detective fiction from the same period?

To address these questions, the thesis combines three complementary components that move from exploration to testing to prediction. First, a corpus-based stage examines distributions and recurrent tendencies across the dataset to identify candidate stylistic differences between high-engagement and other texts. Second, these observations are evaluated through statistical modelling to test whether the differences are robust once key controls are introduced. Third, supervised classification models are trained on the extracted features to assess whether stylistic signals associated with success generalize beyond descriptive patterns and support out-of-sample prediction.

The remainder of the thesis is structured as follows. Chapter 1 situates the project within scholarship on literary success and computational approaches to style, clarifying why success is difficult to define and why controlled corpus designs are necessary. Chapter 2 focuses on Conan Doyle's role in detective-fiction criticism and on stylistic properties commonly associated with Holmesian narration, establishing a literature-based rationale for treating Doyle as a methodological anchor. Chapter 3 details corpus construction, preprocessing, and the operationalization of success via Project Gutenberg downloads, including biases and limitations inherent to public-domain repositories. Chapter 4 presents a corpus analysis that explores how semantic fields related to perception, inference, procedural detection, and other thematic clusters pattern across the corpus. Chapter 5 reports the main statistical analyses of stylistic predictors of success. Chapter 6 evaluates supervised NLP models trained to classify texts as "successful" vs "other" from the compact feature set, emphasizing both performance and interpretability. Chapter 7 synthesizes findings, outlines contributions and limitations, and proposes directions for future work on style, attention, and genre-controlled literary success.

Chapter 1: Literary Success and Computational Approaches to Style

1.1 Research questions and introduction

This thesis adopts a comparative and computational perspective to investigate the relationship between stylistic features and literary success in late nineteenth and early twentieth century detective fiction. Its central aim is to examine whether stylistic characteristics, operationalized through measurable linguistic variables, distinguish more successful literary works from others, and to assess the extent to which such distinctions apply across authors of the same genre that were active during the same period.

The study focuses on the detective fiction of the early 1900s, giving particular attention to Arthur Conan Doyle, whose works provide a particularly suitable case study not only because of their enduring popularity and canonical status, but also because of their methodological relevance for comparative stylistic analysis. Doyle's detective fiction forms a relatively coherent and well-defined body of texts within a single genre and historical period, while simultaneously occupying a position of exceptionally high visibility compared to that of most of his contemporaries. This combination makes it possible to examine stylistic distinctiveness under controlled genre and temporal conditions, while also assessing how stylistic patterns relate to present-day readership engagement. Doyle's central role in the consolidation of detective fiction as a modern literary genre further allows his stylistic practices to be used as a reference point to interpret results. This perspective allows the study to address not only questions of authorial distinctiveness, but also variation associated with present-day engagement and repository visibility, operationalized in this study through Project Gutenberg download metrics, which are used to gauge public-domain readership attention.

The study focuses on a small set of theoretically motivated linguistic variables, each intended to capture a distinct macro-level dimension of narrative style. While a broader feature inventory is extracted for transparency (Chapter 3), the analyses reported in this thesis focus on these five variables.

The impact of lexical choices is represented through semantic concreteness, which indexes the degree to which a text favors tangible, perceptually grounded language over abstract expression. Syntactic and morphosyntactic organization is approximated through measures of subordination density, serving as a proxy for structural complexity. Stylistic perspective and narratorial stance are captured through the distribution of first and third-person

pronouns, reflecting differences in focalization and narrative voice. Finally, affective engagement is operationalized through arousal, interpreted as a measure of emotional activation that captures the intensity of emotional language independently of its positive or negative valence.

This thesis aims to answer the following research questions:

RQ1. Is it possible to systematically differentiate “successful” from “other” detective fiction texts in the period of Arthur Conan Doyle using only linguistic and stylistic features?

Adopting a corpus-based perspective and statistical approach, this question seeks to identify linguistic and stylistic features that recur across successful books. This question allows for both statistical testing and exploratory corpus analysis, allowing indicators highlighted by previous studies, such as the prevalence of specific semantic fields or patterns of lexical usage, to be systematically evaluated.

RQ2. To what extent can supervised computational models trained on extracted stylistic features predict success in detective fiction from the same period?

The aim of this question is not to reduce literary value to a numerical score, but to evaluate whether stylistic signals associated with success are sufficiently consistent to support classification beyond chance levels.

To address these mentioned research questions, the study employs three complementary analytical components that together move from corpus exploration to statistical testing and modelling. For this purpose, the analyses focus on five variables: concreteness (lexical-semantic specificity), arousal (affective activation), first- and third-person pronoun density (narratorial stance), and subordination density (syntactic embedding/complexity).

First, a corpus-based analysis is conducted to explore broad linguistic and stylistic patterns across the dataset, focusing on analyzing how concreteness is distributed across corpora. This stage adopts an exploratory perspective, examining distributions, relative frequencies, and recurrent tendencies of these variables to identify potential stylistic differences between successful detective fiction books and other literary works in the analyzed period.

Second, these corpus-level observations are subjected to statistical analysis. This stage evaluates whether the stylistic patterns identified in the exploratory analysis are systematic and robust, and whether specific linguistic features significantly differentiate successful from

other detective fiction texts. Together, corpus-based and statistical analyses address the first research question by combining descriptive insight with formal testing.

Third, a supervised computational modelling component uses the examined stylistic features as input to classification models to assess their predictive capacity. This final stage addresses the second research question by evaluating whether stylistic signals associated with success are sufficiently consistent to support prediction model training. Taken together, these three analytical components provide an integrated framework for examining both the explanatory and predictive relationship between narrative style and literary success within the selected genre and period.

Together, these research questions and methodological choices establish a structured framework for investigating the stylistic dimensions of literary success in detective fiction. The remainder of this chapter situates the study within existing scholarship on literary success, institutional reception, emotional engagement, and stylistic analysis, providing theoretical and empirical contexts for the analyses that follow.

1.2 Defining literary success

The question of what determines the success of literary works has long captured the interest of scholars across multiple disciplines, including linguistics, literary studies, publishing, and marketing. Understanding this phenomenon is not only of theoretical interest but also of practical significance for publishers, authors, and cultural institutions. Literary success influences canon formation, shapes publishing strategies, and affects how texts circulate across generations and cultural contexts. As a result, researchers have employed a wide range of methodologies to investigate why certain works achieve widespread recognition while others may remain relatively obscure.

Early investigations into literary success tended to adopt predominantly qualitative and interpretive perspectives (Forsyth, 2000). These studies often relied on literary critics' evaluations, anecdotal evidence, or expert judgment to explain popularity and influence. While such approaches have generated valuable insights into aesthetic merit and cultural relevance, they are inherently subjective and difficult to generalize. In contrast, more recent research has increasingly leveraged large-scale datasets and computational methods, enabling systematic, replicable analyses of textual features, publication patterns, and reader engagement. This shift reflects a broader trend within the digital humanities toward combining traditional literary scholarship with quantitative analysis (Ciotti, 2021).

Lately, data-driven approaches have also started considering the book not merely as an isolated object, but as part of a broader cultural and material system. In this perspective, texts are embedded within ecosystems of genre conventions, publication practices, and reader attention, which shape both texts' reception and long-term survival (Moretti, 2000; Underwood, 2019). Consequently, large-scale textual analysis is increasingly complemented by awareness of selection effects: the fact that what is preserved, digitized, and widely available is often a function of cultural visibility rather than representativeness (Bode, 2012; Jockers, 2013).

This insight is particularly relevant for historical corpora, where canon formation, reprinting practices, and institutional endorsement play a decisive role in determining which works remain accessible for analysis.

However, a foundational challenge in the search for what determines a work's success has always been the definition of "success" itself. Literary success is a multifaceted concept encompassing commercial performance, readership, critical reception, institutional recognition, and long-term cultural impact. Due to the various interpretations of the concept of "success", previous studies have operationalized it in markedly different ways, depending on their disciplinary orientation and available data.

Given the multifaceted nature of literary success and the absence of a single agreed-upon definition, previous research has operationalized success in a variety of ways, reflecting different disciplinary priorities and data constraints. The following sections review major strands of this literature, focusing on institutional and market-based indicators, audience engagement and emotional mechanisms, long-term impact and persistence, and stylistic features. This overview provides the conceptual and empirical context for the operational definition of success adopted in the present study.

1.3 Institutional and market-based explanations of success

This section encompasses work that has examined success primarily through institutional and market variables. A study conducted by Yucesoy et al. (Yucesoy et al., 2018) defined a book's success through its inclusion in The New York Times Bestseller List and sought to predict the sales of newly published works based on variables such as genre, author prominence, publisher reputation, and the temporal and geographic context of publication. Beyond identifying predictors of short-term commercial performance, this study also emphasized the highly unequal distribution of cultural attention in literary markets. Most books experience only brief visibility, while a small minority enjoy sustained success, often

reinforced by cumulative advantage mechanisms in which early recognition increases the likelihood of future exposure. The authors claim that such dynamics are not fully explained by intrinsic textual quality alone, but emerge from interactions between genre expectations, author reputation, and external shocks such as media coverage or institutional recognition. Although this study focuses on bestseller data, findings have clear implications for historical literary analysis, suggesting that long-term prominence may reflect structural attention dynamics as much as stylistic distinctiveness. Their findings also indicate that fiction works are more likely to achieve bestseller status, a result consistent with prior research highlighting the enduring market dominance of fiction (D. Wang et al., 2013). Yucesoy et al. (2018) also observe that a book's lifespan on the bestseller list was typically short unless it reached the top position, underscoring the volatility of commercial success. Additionally, their analysis reveals notable gender patterns across genres: while the fiction dataset was balanced overall, female authors were more successful in romance, whereas male authors demonstrated a relative advantage in mystery fiction. These findings illustrate how success is shaped not only by textual factors but also by broader social and market dynamics.

These results were further confirmed by a recent study (X. Wang et al., 2019) that investigated the early prediction of book sales using large-scale data from the U.S. publishing market, focusing exclusively on features available prior to publication, such as author visibility, previous sales, genre, topic, publication month, and publisher prestige. Analyzing over 170,000 hardcover books using NPD BookScan data, they demonstrate that book sales follow a heavy tailed distribution and show that publisher imprint is the strongest predictor of success across both fiction and nonfiction. Author visibility and prior sales play a particularly important role in fiction, especially in Thrillers and Mystery & Detective, where serial authorship is common. These parameters were accounted for by examining the number of views the Wikipedia article for each author received. The authors explicitly acknowledge that their model fails to capture intrinsic literary quality, as several major bestsellers are significantly underpredicted despite favorable institutional conditions. This limitation highlights the explanatory gap left by metadata-driven approaches and motivates further investigation into whether textual and stylistic features correlate to literary success beyond institutional and market factors.

A similar quantitative approach was adopted in a recent study carried out by da Silva et al. (da Silva et al., 2024), on a dataset consisting of works published between 1895 and 1924. In this work, books were categorized into "success" and "other" groups, with success defined by at least one appearance on the Publishers Weekly Bestseller List. This was chosen as a

criterion since, although it was not fully specified what a literary work's inclusion in the list meant, bestsellers were understood by the researchers to have sold at least 2,000,000 paperbound copies or 750,000 hardbound copies, only including books distributed through trade channels such as bookstores and libraries. After preprocessing (tokenization, stopword removal, lemmatization, and removal of non-narrative paratext), the authors represented each book using bag-of-words and doc2vec embeddings and applied both visualization (SemAxis, LDA) and supervised classification. Their best performing configuration, standardized bag-of-words with logistic regression, reaches an accuracy of 0.75, which they interpret as evidence that high-accuracy prediction of success is not feasible using full-text content alone. Nonetheless, they identify lexical tendencies associated with class separation: their findings suggest that less successful works contain more frequent references to body parts, whereas more successful texts exhibit a broader and less conventional lexicon, implying a positive association between lexical diversity, stylistic richness, and commercial performance. The study also reports no evidence that subject (genre) headings explain the observed separation, suggesting that content-based differences persist beyond broad categorical metadata. Lastly, the study concludes that while not sufficient by itself to generate an adequately accurate model, the textual content of each book itself would seem to play a central role in determining success, corroborating earlier observations by researchers such as Ashok et al. (Ashok et al., 2013).

1.4 Audience engagement and emotional mechanisms

Other strands of research have conceptualized success in terms of audience engagement and emotional response rather than sales alone. Berger and Milkman (Berger & Milkman, 2012) offer a psychologically grounded account of textual "success" operationalized as social transmission. Their studies focused on 6,956 New York Times articles and predicted whether an article makes the newspaper's "most e-mailed" list. Using LIWC-based measures of valence and affect-ladenness, where LIWC (Linguistic Inquiry and Word Count) is a widely used lexicon-based text analysis tool for quantifying psychological and emotional categories in language, alongside human ratings of discrete emotions. Their results indicate that positive content is more likely to be shared but also demonstrate that virality is not explained by valence alone. Instead, emotions characterized by high physiological arousal, such as awe, anger, and anxiety, are positively associated with sharing, while low-arousal emotions such as sadness reduce transmission. Importantly, these effects persist even after controlling confounds that mechanically increase exposure (e.g., prominent homepage

placement). Their findings suggest that emotional “activation” can function as a measurable textual dimension linked to diffusion and audience engagement, offering a transferable conceptual model for studying how textual features relate to popularity beyond purely institutional or market factors. Although this study focused on journalism rather than fiction, its findings underscore the broader relevance of emotional resonance as a driver of textual success.

This distinction between emotional valence and arousal is also supported by psycholinguistic and neurocognitive research, which shows that these dimensions interact during reading and affect processing effort even at the single word level, with high arousal stimuli eliciting greater cognitive and neural engagement than low arousal stimuli (Citron et al., 2014).

Recent computational work has also extended the analysis of emotional content beyond isolated sentiment measures to the level of narrative structure. Reagan et al. (Reagan et al., 2016) analyzed the emotional trajectories of 1,327 fiction texts written in English and downloaded from Project Gutenberg. Sentiment time series were extracted across each narrative, and six core “emotional arcs” were identified using matrix decomposition, clustering, and machine learning methods. Emotional arcs refer to characteristic patterns in the rise and fall of a story’s emotional valence over time, capturing how a narrative’s affective tone evolves from beginning to end. In their study, Reagan et al. identify 6 main types of emotional arcs:

• ‘Rags to riches’ (a constant rise in Valence).
• ‘Tragedy’, or ‘Riches to rags’ (a constant fall in Valence).
• ‘Man in a hole’ (fall-rise).
• ‘Icarus’ (rise-fall).
• ‘Cinderella’ (rise-fall-rise).
• ‘Oedipus’ (fall-rise-fall).

The authors relate these emotional arc types to measures of success, operationalized as Project Gutenberg download counts. Their results indicate that emotional arcs are not equally successful: while the most common arcs are not featured in the most downloaded works, some of them, such as “Icarus”, “Oedipus”, and more complex “Man in a Hole” structures, exhibit substantially higher median download counts and greater variance.

The authors interpret these findings as evidence that the emotional experience evoked by a narrative’s emotional arc influences how stories are shared and engaged with, while also emphasizing the high variability within each category and the limitations of downloads as a

proxy for success. These results suggest that a complex emotional structure, alongside local stylistic features, may contribute to differential readership engagement.

1.5 Impact dynamics and long-term persistence

Other researchers have focused their studies on determining how long a work's success can endure. Long-term success has been examined in terms of the role played by institutions such as publishers, journals, and evaluative bodies, as well as by the accumulated visibility and prestige of authors and outlets, across different domains. In the context of scientific publishing, Wang, et al. (D. Wang et al., 2013) show that commonly used short term indicators of impact, such as early citation counts or journal impact factors, are poor predictors of long-term influence. Analyzing large-scale citation data, they demonstrate that publications with similar early trajectories can diverge substantially over time, and they therefore introduce the concept of an abstract "fitness" parameter to capture latent qualities such as novelty and perceived importance. In the context of scientific publishing, their results emphasize the relevance of institutional validation, showing that works released by highly reputed journals and publishers are more likely to achieve enduring impact. While the mechanisms governing success in academic publishing are not directly equivalent to those operating in literary fiction, these findings nonetheless highlight, in a more general sense, the role of institutional endorsement, visibility, and reputational accumulation in shaping the long-term trajectories of cultural artifacts.

Taken together, the strands of research reviewed in this chapter—ranging from institutional and market-based explanations to audience engagement and emotional mechanisms, to studies of long-term impact and persistence—highlight that literary success is a complex and multidimensional phenomenon. Across these approaches, success is shown to be shaped by an interplay between external factors, such as author visibility, institutional endorsement, publisher prestige, and genre conventions, and intrinsic textual properties, including lexical diversity, emotional intensity, and narrative structure. While these studies differ in their operational definitions of success and in their methodological focus, they converge in suggesting that textual features cannot be fully reduced to, nor fully separated from, broader social and institutional dynamics.

Building on this literature, the present study shifts the focus more explicitly toward stylistic form as an independent explanatory dimension. Rather than examining success primarily through metadata, institutional indicators, or high-level narrative patterns, Chapter 3 introduces the specific linguistic and stylistic features extracted from the corpus and

motivates their relevance for analyzing both authorial distinctiveness and success-related variation within detective fiction. This transition marks a move from a theoretical background to the empirical definition of stylistic variables that form the basis of the statistical and computational analyses developed in the remainder of the thesis.

1.6 Style as an independent explanatory dimension

Ashok et al. (Ashok et al., 2013) provide one of the most influential computational studies linking writing style to literary success. Focusing on novels drawn from Project Gutenberg across multiple genres, including Detective and Mystery fiction, they define success primarily through download counts and frame their analysis as a comparison between more and less successful professionally published works. They control authorship effects by limiting the number of works per author and ensuring that the same authors do not appear in both training and test sets, thereby isolating stylistic signals rather than author-specific ones.

Using a range of stylistic features, including lexical choices, part-of-speech distributions, grammar-rule encodings, and phrasal and clausal constituent structures, they demonstrate that statistical stylometry alone can discriminate successful novels with accuracies reaching up to 84%. Their analysis reveals consistent stylistic differences between classes: more successful novels tend to exhibit higher proportions of prepositions, determiners, nouns, and discourse connectors, while less successful works rely more heavily on verbs, adverbs, and action-oriented vocabulary. Lexical analyses further indicate that less successful novels contain more terms related to body parts and sentiment-laden or extreme language, whereas successful works favor cognitive verbs and reporting structures.

One of the study's most counterintuitive findings concerns readability. Contrary to conventional assumptions that readability is a hallmark of good writing, Ashok et al. (2013) find that more successful novels are, on average, less readable according to standard metrics such as the Flesch index and Gunning FOG score. They interpret this as evidence that syntactic and stylistic complexity may be positively associated with literary success in fiction as they are employed to portray more complex plots or characters. While the authors emphasize that their findings demonstrate correlation rather than causation, their work establishes a strong quantitative precedent for investigating stylistic features, such as part-of-speech distributions, as meaningful indicators of literary success.

Recent work in consumer psychology and computational linguistics further supports the claim that linguistic style constitutes an independent dimension shaping the success of

cultural artifacts. Boghrati et al. (Boghrati et al., 2023) examine the impact of writing style on the success of academic articles, explicitly distinguishing style from content through the analysis of function words, grammatical elements such as articles, prepositions, conjunctions, and pronouns that convey little semantic information but reflect how ideas are expressed. Analyzing nearly 30,000 articles across multiple disciplines, the authors demonstrate that stylistic features explain a significant proportion of variance in citation counts, even after accounting for topical content, author prominence, journal prestige, and readability. Notably, function-word usage accounts for between 4–11% of the total variance explained and up to 27% of the explanatory power attributable to linguistic content.

Although focused on academic writing rather than literary fiction, this work is methodologically relevant in that it provides a principled framework for separating stylistic form from semantic content and demonstrates that low-level grammatical patterns show systematic correlations with collective success outcomes. By showing that stylistic choices, such as syntactic complexity, temporal framing, and personal voice, systematically correlate with impact in a highly content-constrained domain, the study lends further support to the broader hypothesis that stylistic features captured through part-of-speech and function-word distributions may similarly contribute to differential success in literary texts.

1.7 Historical and genre-specific motivation

While extensive literary scholarship has examined the distinctive styles of individual authors (McGann, 1998) , and while other studies have analyzed content-related features such as plot structure, character development, emotional arcs and genre conventions, (Hall, 2012; Harvey, 1953), these investigations are largely qualitative. They rely on expert interpretation, intuition, and close reading, making it difficult to assess whether consistent stylistic patterns exist across successful works more broadly.

Historical examples from detective fiction illustrate both the importance and the ambiguity of style in determining literary success. A case in point is S.S. Van Dine. Van Dine (1888–1939), born Willard Huntington Wright, was among the most commercially successful American mystery writers of the 1920s and 1930s, primarily through his detective character Philo Vance, whose novels enjoyed wide circulation and sustained popularity during his lifetime (Knight, 2003).

Crucially, Van Dine explicitly sought to formalize the elements of successful detective fiction. In his 1928 essay *Twenty Rules for Writing Detective Stories*, published in *The American Magazine*, he articulated a set of prescriptive principles intended to codify the narrative,

stylistic, and structural features he believed were essential for reader satisfaction and genre legitimacy (Van Dine, 2015). These rules represent an early attempt to theorize literary success in explicitly stylistic and formal terms, linking narrative clarity, structural balance, and reader engagement.

In the present study, Van Dine's rules are not treated as normative prescriptions or as an evaluative framework for individual texts. Instead, they serve as a historically grounded source of stylistic hypotheses that inform the corpus-based and statistical analyses. Several of Van Dine's principles explicitly concern narrative content and stylistic emphasis—such as the marginalization of romantic subplots, the prioritization of rational problem-solving over emotional development, and the avoidance of digressive descriptive excess—which can be operationalized as measurable textual features.

Accordingly, the corpus analysis examines whether stylistic patterns consistent with these prescriptions are systematically associated with successful texts, for example by testing the relative prominence of thematic domains related to romance or emotional involvement, the distribution of descriptive versus action-oriented language, and other stylistic correlates suggested by Van Dine's guidelines. The operationalization of these hypotheses and their empirical testing are described in detail in Chapter 4.

This chapter has outlined the major approaches to defining and studying literary success, highlighting both their contributions and limitations. While prior research has advanced our understanding of content, emotion, and institutional factors, the quantitative investigation of stylistic markers remains comparatively underdeveloped. By integrating insights from historical exemplars of detective fiction with modern computational linguistics, this study aims to contribute a novel perspective on the stylistic foundations of literary success within the genre and period under consideration.

Chapter 2: Conan Doyle in the scholarship on detective fiction and stylistic analysis

2.1 Arthur Conan Doyle's background and its relevance for stylistic analysis

Arthur Conan Doyle (1859–1930) was trained in medicine at University of Edinburgh, qualifying in 1881 (Bachelor of Medicine and Master of Surgery) and later completing an M.D. in 1885; he practiced as a physician before turning fully to writing, and his medical formation is repeatedly identified as foundational to his narrative preoccupation with observation, inference, and professional credibility (Wilson, 2019; Conan Doyle Estate, n.d.). Although this chapter concentrates on Doyle's detective fiction—especially the Holmes stories narrated through Dr. John Watson—it is important to acknowledge that Doyle was a prolific writer beyond the detective mode: he produced historical fiction (e.g., The White Company), adventure and science fiction (e.g., The Lost World), and a substantial body of non-fiction and polemical writing, including works connected to the paranormal and spiritualism (Wilson, 2019; Doyle, 2006)

These parts of his works are not the main focus of the present analysis, but they clarify that Holmes emerges from a wider authorial career shaped by late-Victorian popular print culture and by a professional training that foregrounded diagnostic reasoning, an interplay that remains highly relevant when interpreting Doyle's stylistic choices in relation to narrative authority, evidential detail, and reader alignment (Wilson, 2019; Conan Doyle Estate, n.d.).

This chapter places Conan Doyle within major critical traditions on detective fiction and within the subset of scholarship that links Holmesian narration to analyzable linguistic patterns. In line with the thesis's existing framing, the focus here is not to restate the project's general aims, but to sharpen the literature-based rationale for using Doyle as a key case and to describe the stylistic properties that scholarship most consistently associates with his detective writing.

2.2 Conan Doyle's centrality to the consolidation of modern detective fiction

A durable claim in crime-fiction scholarship is that Doyle's Holmes stories helped stabilize a "classical" model of detective narrative: a staged problem of knowledge, followed by a solution that reorganizes scattered details into a coherent causal explanation. A canonical formulation of this model is Todorov's description of detective fiction as a structure built from two interlocking stories—the story of the crime (whose events already unfolded but are

partially hidden) and the story of the investigation (the reconstruction of that earlier story) (Todorov, 2019).

Later genre histories build on this structural insight to argue that “classical” detection becomes recognizable precisely because it is repeatable: it relies on recurrent roles (detective, assistant, client, police), an iterative plot-cycle (initial problem, inquiry and final revelation), and a distinctive rhetoric of proof (Scaggs, 2005). In formula-theory terms, detective fiction’s durability is not a by-product of simplicity but a feature of how conventions generate intelligibility and pleasure through variation within constraints (Cawelti, 2014).

This matters for the present thesis because Doyle is not merely an “influential author” in a general sense he is frequently treated as a cornerstone of the genre, through which its formal architecture and readability mechanisms become unusually visible. That visibility makes it easier to interpret why specific stylistic variables—lexical concreteness, syntactic packaging, narrative perspective, and affective activation—are likely to co-vary with the kind of engagement patterns the thesis later models.

2.3 Publication ecology, seriality, and the shaping of a mass-readable detective voice

Scholarship on late-Victorian print culture, and on Sherlock Holmes in particular, often emphasizes that the Holmes phenomenon is inseparable from the periodical ecosystem that disseminated it, especially the illustrated monthly magazine environment in which the stories circulated. Studies of Doyle’s relationship with media highlight how magazine formats incentivized clarity, momentum, and re-entry: narratives needed to be legible, suspenseful, and episodically satisfying, while also supporting a continuing reader relationship across installments (O’Gorman, 2007). Work on *The Strand Magazine* similarly underlines its “cultural readability”: the magazine’s mixed contents, visual framing, and community-building practices encouraged broad access and habitual consumption (Liggins & Vuohelainen, 2019).

Because the Holmes stories circulated within a serial print economy, some of their stylistic patterns could be interpreted as responses to the practical demands of that medium: maintaining momentum, supporting quick comprehension, and delivering periodic payoffs. This perspective does not assume that such patterns cause success, but it provides a historically plausible framework for explaining why certain measurable features might co-occur with enduring readership and later forms of engagement, keeping the computational

findings tethered to literary-historical conditions rather than treating them as free-floating statistical signals.

2.4 Watson as narratorial interface: focalization, information control, and stance

One of the most consistently discussed stylistic features of Doyle's Holmes canon is its narratorial arrangement: the detective is rarely the default narrator, and the case is most often mediated through Dr. Watson. Narratological work on the Holmes stories stresses that Watson's mediation is not incidental; it is a functional design choice that structures suspense and interpretation.

First, Watson provides limited focalization, enabling systematic delay: the reader sees much, but will not fully see how the pieces fit together until the solution phase. In their analysis of evidential presentation in Holmes, Emmott and Alexander (Emmott & Alexander, 2024) show how descriptions can appear credible while remaining incomplete or misleading, allowing the plot to bury key clues without breaking the internal logic of the story and world.

Second, Watson's voice supplies evaluative stance that calibrates reader alignment with the detective. Recurrent admiration, surprise, and retrospective endorsement help legitimate the eventual explanatory reconstruction, particularly when the reasoning chain is withheld or compressed earlier in the narrative.

Third, Watson also functions as a stabilizing "ordinary" baseline, an apparently commonsensical, socially legible viewpoint through which Holmes's exceptional inferences can be narrated as credible rather than merely spectacular (Jann, 1990). This narratorial design is inseparable from the text's information management: by restricting access to full explanation and controlling when interpretive closure arrives, the stories can sustain suspense and then legitimate reconstruction at the moment of revelation (Krasner, 1997).

2.5 Observation, evidence, and the language of inference

Another influential strand of scholarship links detective fiction, and Holmes in particular, to nineteenth-century epistemologies of evidence. Thomas's account of the rise of forensic thinking in the nineteenth century frames detective narratives as cultural sites where bodies and objects are rendered readable through trace-based interpretation (Thomas, 1999).

For the present thesis, this connection is significant because an evidence-centered narrative stance tends to encourage stylistic regularities that can be operationalized: lexical attention to concrete entities and observable traces, sequencing that supports procedural reconstruction, and inferential language that turns perception into explanation. In the Holmes

stories, this often takes the form of a rapid alternation between perceptual reporting—a catalogue of visible details such as stains, footprints, handwriting, odors, or the state of a room—and inferential commentary, where those details are explicitly linked through causal or evidential connectives (e.g., therefore, suggesting that, it follows that, which implies). This alternation aligns with Todorov's "two-story" architecture, but here it becomes visible at the level of linguistic texture: a text must move between observation and retrospective explanation to deliver both puzzle and payoff.

2.6 Characteristic prose tendencies in criticism: clarity, pacing, and descriptive economy

Across criticism and genre history, Doyle's Holmes prose is often characterized not by maximal ornamentation but by a balance between vividness and control; an equilibrium consistent with magazine readability and the demands of puzzle plotting (O'Gorman, 2007; Scaggs, 2005). For instance, scenes are frequently built through short, functional descriptive units that identify what matters for inference (a mark on a sleeve, an object out of place, a fragment of dialogue), while non-essential atmospheric detail is kept brief or postponed. This produces prose that feels clear and fast without sacrificing vividness.

Three tendencies recur across these discussions and are especially useful when interpreting computational features. First, the Holmes stories tend to maintain narrative velocity by advancing through conversational exchange, including client briefings, interrogations, and the continual dialogue between Holmes and Watson, often punctuated by brisk transitions to investigative sites, a rhythm that supports pace and sustains curiosity (Scaggs, 2005). Second, description is frequently selective and function-driven: rather than distributing atmosphere evenly, the narration foregrounds details that are likely to matter for reasoning—marks, objects, and other traces—so that concreteness can be high while remaining economical, with descriptive density concentrated where it serves evidential work. Third, affect is commonly managed rather than melodramatically saturated; even when the subject matter is sensational, Watson's reportorial stance tends to contain emotional language and reserve peaks of intensity for moments of threat, discovery, or final revelation. Read through a computational lens, this pattern supports the possibility, consistent with this thesis's feature design, that engagement may correlate with strategic arousal patterning rather than with uniformly elevated emotional diction throughout the text. These claims are best read as "interpretive priors": theory-motivated expectations that guide what patterns we look for, without predetermining the outcome. They provide literary reasons to expect certain

aggregate patterns while leaving space for the thesis's empirical results to confirm, qualify, or complicate them.

2.7 Doyle and computational work: a practical bridge between criticism and measurement

Doyle's Holmes is also methodologically interesting due to its large, internally coherent, and widely digitized qualities that have made it a recurring object of corpus stylistics and computational literary study. For instance, Vezzani & Di Nunzio (Vezzani & Di Nunzio, 2019) adopt a mixed quantitative–qualitative approach to linguistic analysis of the Holmes corpus (in their case, medical terminology), explicitly arguing for the interpretive value of computational extraction paired with close reading.

This type of work supports a key methodological point for the present thesis: Doyle sits at a productive intersection where (1) traditional criticism has articulated stable claims about narration, evidence, and readability, and (2) computational methods can test whether those claims correspond to measurable distributional tendencies across texts. The significance is not that computational studies "prove" Doyle's importance, but that the Holmes canon reliably enables theory-driven operationalization without forcing the analysis into purely impressionistic description. While a broader critical map of Holmes's cultural persistence and interpretive flexibility is also available in resources such as The Cambridge Companion to Sherlock Holmes (Allan & Pittard, 2019), computational studies have also used the Holmes canon for reproducible text analysis: Vezzani & Di Nunzio (Vezzani & Di Nunzio, 2019) process the complete set of 60 narratives through automatic term extraction and collocation-based corpus analysis to map the distribution and contextual behavior of medical terminology.

2.8 Implications for this thesis: literature-grounded, interpretable hypotheses

Taken together, the scholarship reviewed above supports a set of interpretable expectations that align closely with the thesis's analytic dimensions. Because the Holmes stories are so often mediated through Watson, the narration provides a principled basis for anticipating systematic signals in first-person perspective markers and in evaluative stance—signals that can shape reader alignment and help regulate suspense (Emmott & Alexander, 2024). At the same time, critical accounts that link Holmesian detection to a trace-based epistemology suggest that Doyle's style should show a strong orientation toward concrete objects, perceptual detail, and procedural sequencing, since observation and evidential reasoning

are central to how the narratives generate plausibility and payoff (Thomas, 1999). The genre's requirement to delay explanation while ultimately delivering a legible causal reconstruction also makes it reasonable to connect stylistic effects to syntactic packaging—particularly patterns that support inferential framing and the explicit staging of causal chains, such as explanatory subordination and evidential constructions (Scaggs, 2005; Todorov, 2019). Finally, work on periodical culture and "readability" provides a framework for understanding why emotional intensity may be distributed strategically: rather than remaining uniformly heightened, affect can be patterned to maintain curiosity and concentrate payoff at moments of threat, discovery, and revelation (Liggins & Vuohelainen, 2019; O'Gorman, 2007).

On this basis, it is clear why Doyle is particularly important for the present thesis: not simply because he is canonical, but because literary scholarship repeatedly attributes to his detective writing a set of mechanisms that can be meaningfully connected to measurable stylistic features and then interpreted back through genre theory and media history.

This chapter has argued that Doyle's centrality to this work is not only historical or canonical, but methodological: the Holmes–Watson configuration makes mechanisms of narration, evidence, and reader alignment unusually legible, and thus suitable for operationalization. However, translating such interpretive claims into testable evidence requires that the textual material, the success criterion, and the measurement procedure be defined with precision. Chapter 3 therefore details the construction of the Project Gutenberg corpus, the rationale and limitations of using download-based thresholds as a proxy for present-day engagement, and the feature-extraction pipeline that converts each work into document-level measures of concreteness, subordination, pronoun-based perspective, and arousal. In doing so, the thesis shifts from literature-grounded expectations to the empirical framework through which those expectations can be evaluated.

Chapter 3: Corpus construction and operationalization of the concept of literary success

3.1 Introduction

This chapter describes the construction of the textual corpora used in the present study and outlines the operational definitions of literary success adopted. Given the quantitative and corpus-linguistic orientation of the thesis, particular attention is paid to transparency, reproducibility, and methodological justification. The dataset is designed to support a comparative stylistic analysis of successful detective fiction in relation to that of other detective writers active in the early 1900s, combining statistical analysis, corpus-based investigation, and computational modelling to examine systematic differences in linguistic and stylistic patterns.

This section focuses on extracting a comprehensive set of linguistic and stylistic features intended to capture complementary macro-level dimensions of narrative style in a transparent and reproducible manner. At the lexical-semantic level, the analysis includes measures of semantic concreteness, as well as affective lexicon–based features such as valence and arousal, which characterize the emotional and perceptual properties of word choice.

At the grammatical and morphosyntactic level, features include normalized part-of-speech distributions and morphological indicators such as tense, grammatical person, and number, and syntactic complexity measures derived from dependency parses. These include noun phrase density, verb phrase density, prepositional phrase density, and subordination density, operationalized through clausal embedding relations.

Stylistic perspective and narratorial stance are captured through the relative frequency of first- and third-person pronouns, providing an index of narrative focalization and point of view. Additional stylistic measures include adjective density and sentence-level approximation of descriptive density, designed to identify passages characterized by high degrees of modification and stative predication.

Finally, readability-related features are computed at the document level, including established metrics such as Flesch Reading Ease, as well as auxiliary measures based on sentence length, word length, and syllabic complexity.

All features are aggregated at the document level and normalized to ensure comparability across texts of varying length and format. The same feature set is used consistently across

the corpus-based analysis, statistical testing, and supervised modelling components of the study.

Although the extraction pipeline computes a broad inventory of grammatical, syntactic, semantic, affective, and readability measures, the core analyses in this thesis deliberately focus on a restricted, theory-driven subset of five document-level variables: concreteness, arousal, first-person pronoun density, third-person pronoun density, and subordination density. This restriction is motivated by interpretability and hypothesis alignment: these features operationalize the macro-level stylistic dimensions introduced in Chapter 1 (lexical-semantic specificity, affective activation, narratorial stance, and syntactic embedding) and allow both statistical testing and predictive modelling to remain directly linked to the study's research questions. All other extracted features are retained in the dataset to ensure transparency and reproducibility and to support potential extensions, but they are not used as predictors in the supervised models reported in following chapters.

3.2 Data source and selection criteria

All texts analyzed in this study were obtained from Project Gutenberg, an open digital library that provides free access to public-domain books, using Project Gutenberg's machine-readable catalog rather than ad-hoc scraping to ensure reproducibility. Metadata were drawn from the official Project Gutenberg catalog feeds (CSV/RDF), and raw text files were downloaded using the corresponding e-book identifiers. The catalog snapshot used for this study was retrieved on December 20th, 2025, and the snapshot date is treated as part of the dataset specification because Project Gutenberg metadata (including download counts) is updated over time.

Texts were selected by their subject headings as recorded in the Project Gutenberg catalog. Specifically, each record can include multiple subject labels. A book was included only if its full set of subject labels, considered jointly, satisfied two conditions:

1) at least one subject label contained "detective" or "private investigator",
2) at least one subject label included "fiction"

These conditions were evaluated across the complete set of subject labels for each book, rather than requiring both criteria to appear within a single label. (case-normalized string match on subject headings).

Only English texts were included. Works tagged as translations (or where the catalog metadata clearly indicated a translator or non-English source language) were excluded to minimize translator-driven stylistic artifacts. The reason for this is the fact that, as shown by

previous studies (Koppel & Ordan, 2011), translated texts often present significant differences in the language they employ, thus including them in the corpus would have introduced inconsistencies in the study.

The corpus spans the period between 1880 and 1935, covering the consolidation of detective fiction as a mass-market genre and encompassing Arthur Conan Doyle's Sherlock Holmes publication window from the first appearance of Holmes in *A Study in Scarlet* (originally published in 1887; cited here in the 1892 edition) to the final Holmes collection *The Case-Book of Sherlock Holmes* (1927; cited in the 1993 edition). This window is used to reduce diachronic stylistic drift while retaining a sufficiently large comparison set. This period also ensures that Doyle's Holmes texts are represented within the same genre and publication context as the comparison set, allowing the thesis to treat Doyle as a methodological anchor rather than a separate object of study.

Text identity and deduplication. To reduce duplication and edition effects, the corpus excludes: exact duplicate texts (byte-identical after header/footer removal), and re-issues of the same work where Project Gutenberg provides multiple e-book IDs for substantially identical content. When duplicates were identified, the earliest-added or best-documented Project Gutenberg record was retained (see Section 3.6 for limitations).

3.3 Corpus composition and text type considerations

The final corpus consists of 281 detective fiction works authored by 142 different writers active during the selected period and was employed in a different configuration for each of the analyses (see Chapter 3.7). To mitigate potential distortions arising from differences in text length or form, all text features were normalized during preprocessing, and analyses focused on relative frequencies and distributions rather than absolute counts. By analyzing stylistic variables as proportions or normalized measures, the study minimizes the impact of document length and supports comparability across texts of varying sizes.

Because the corpus consists of book-length narratives drawn from Project Gutenberg, documents vary substantially in length even after applying the minimum-length threshold. The final corpus has the following values: median = 48,782, mean = 51,694, SD = 20,415. To mitigate distortions arising from these differences in text length and narrative form, all stylistic features were computed as normalized measures (rates, proportions, or densities) rather than raw counts. This normalization supports comparability across documents and ensures that observed differences reflect stylistic tendencies rather than the sheer amount of text.

All analyses were conducted at the level of whole documents, rather than chapters or smaller segments. Treating each work as a single analytical unit preserves authorial and narrative coherence and aligns with the study's focus on global stylistic patterns characteristic of book-length detective fiction in this period.

3.4 Preprocessing and text normalization

Prior to analysis, all texts were subjected to a standardized preprocessing and normalization pipeline designed to ensure comparability across documents and to minimize the influence of material not present in the original works. As the source texts were drawn from Project Gutenberg, each file contained substantial amounts of paratextual content unrelated to the narrative itself, including headers, footers, legal notices, and metadata. To address this, the following Python script employing regular expressions was developed to systematically remove superfluous text from both the beginning and the end of each file, isolating the narrative content produced by the author:

```
START_PATTERN = re.compile(
    r"(?im)^[^\S\r\n]*\*{3}[^\S\r\n]*START OF (?:THIS|THE) PROJECT GUTENBERG(?: EBOOK)?\b.*?\*{3}[^\S\r\n]*$"
)

END_PATTERN = re.compile(
    r"(?im)^[^\S\r\n]*\*{3}[^\S\r\n]*END OF (?:THIS|THE) PROJECT GUTENBERG(?: EBOOK)?\b.*?\*{3}[^\S\r\n]*$"
)
```

Figure 1: regular expressions employed in cleaning

This step ensured that subsequent analyses were based exclusively on literary text rather than on editorial or platform-specific additions.

Following text cleaning, the remaining content was processed using standard natural language processing procedures. All texts were tokenized and lemmatized (when necessary) using spaCy, an open-source Python library widely used for natural language processing (https://spacy.io). spaCy provides efficient and linguistically informed tools for tokenization, part-of-speech tagging, lemmatization, and syntactic analysis, relying on pretrained statistical models for English that incorporate lexical, morphological, and contextual information. Tokenization segments the text into individual linguistic units (tokens), while lemmatization reduces inflected word forms to their canonical base forms (lemmas), as in the mapping of "*running*" and "*ran*" to "*run*".

Lemmatization supports the study's focus on stylistic and functional features by reducing surface-level lexical variation and facilitating the aggregation of semantically comparable word forms. In the present pipeline, lemmatization is applied only for the corpus analysis (Chapter 4) and for lexicon-based semantic and affective measures, such as concreteness, valence, and arousal, where normalization improves coverage and comparability. Importantly, this normalization does not entail the loss of grammatical or inflectional information: tense, number, person, and syntactic relations are preserved and explicitly captured through part-of-speech tagging, morphological features, and dependency annotations. As a result, stylistically relevant grammatical preferences—such as tense usage, pronoun perspective, and patterns of clausal embedding—remain available for analysis alongside lemmatized lexical features. All preprocessing steps, including text cleaning, tokenization, and lemmatization, were applied uniformly across the entire corpus.

3.5 Operationalizing literary success

Literary success is treated as a binary variable throughout the corpus and NLP study, while the data analysis chapter considers it as a continuous variable. This choice reflects both the structure of prior computational research on literary success and the requirements of the statistical and modelling methods employed.

Literary success is operationalized using Project Gutenberg download counts as a proxy for readership engagement. Given the highly skewed and heavy-tailed distribution, success is defined using a percentile-based threshold rather than absolute download values. Specifically, works falling within the top 10% of the corpus by download count are classified as "successful", while the remaining texts are grouped as "other".

Given the highly skewed distribution of attention measures, the use of percentile-based definitions is a common strategy to obtain robust high-impact subsets that are less sensitive to extreme values (Leydesdorff & Bornmann, 2011; Waltman et al., 2012). Related computational work on literary success similarly operationalizes success via threshold-based binarization of download counts (Ashok et al., 2013).

This threshold is not intended to represent a categorical boundary between successful and unsuccessful literature in an absolute sense. Rather, it constitutes a deliberately pragmatic and corpus-internal criterion that isolates a subset of texts receiving disproportionately high levels of attention, thereby enabling a clear contrast between high-engagement works and the broader background of the corpus. Percentile-based classification reduces sensitivity to extreme values and mitigates the effects of uneven attention distributions.

This operationalization is applied uniformly across all authors, and serves as the sole success criterion used in the following analyses. By adopting a single, consistent definition of success at the corpus level, the study avoids introducing author-specific criteria and ensures comparability across texts and authors.

The use of Project Gutenberg downloads as a success proxy aligns with prior work in computational literary studies (see Chapter 1) and is explicitly interpreted as a measure of present-day engagement rather than as a direct indicator of historical sales or intrinsic literary quality. Therefore, success is treated as an analytical category that facilitates comparison between texts rather than as a definitive judgment of literary value.

3.6 Biases and limitations of the corpus

Several limitations inherent to the corpus construction process must be acknowledged. First, the reliance on Project Gutenberg introduces a selection and survival bias, as inclusion in the repository is contingent on copyright status, digitization priorities, and volunteer availability. As a result, the corpus may overrepresent works that have already achieved a degree of cultural longevity or institutional recognition.

As mentioned before, Project Gutenberg download counts reflect digital engagement rather than historical readership. While downloads provide a useful proxy for current interest, they cannot be interpreted as direct measures of original commercial success or historical popularity. It is also important to note that Project Gutenberg downloads are updated over time. The study therefore treats the download snapshot date as part of the dataset definition and reports it alongside the corpus release. Furthermore, texts can vary in formatting conventions, chapter headings, and residual paratext, all of which can influence NLP pipelines. While the preprocessing stage employs regular expressions to remove superfluous text (Section 3.4), the presence of some residual noise is unavoidable.

Despite these limitations, Project Gutenberg remains an appropriate and widely accepted resource for large-scale literary analysis, particularly for studies focused on stylistic and linguistic features rather than publication history alone.

3.7 Corpus configuration across analyses

The textual data described in this chapter constitutes a single underlying corpus that is employed across all subsequent analyses in the study. However, this corpus is organized into different analytical configurations depending on the specific goals of each methodological component. This distinction allows the study to address complementary

research questions while maintaining consistent preprocessing procedures, feature extraction methods, and operational definitions of success.
The thesis refers to these configurations using the following labels:

1) Main corpus
2) Statistical Analysis corpus
3) Predictive Modelling corpus

The Main corpus structures the texts into two subcorpora: a “Success” subcorpus and a comparison subcorpus consisting of detective fiction works by “Other” authors. This configuration supports comparative stylistic analysis and enables the assessment of authorial distinctiveness and systematic stylistic differences within the genre. In this configuration, Doyle’s Holmes texts are part of the success/other contrast based on where they fall according to the download threshold and are not isolated as a separate author-specific subset.
The Statistical Analysis corpus refers to the same set of 281 texts pooled into a single collection, rather than split into “Success” vs “Other”. Operationally, this corresponds to the unified JSON output produced by the extraction pipeline, which aggregates the document-level representations for all 281 texts into a single file.
The Predictive Modelling corpus pools texts into a single modelling set for supervised prediction but applies additional balancing constraints to reduce author dominance and improve generalizability. The number of texts per author is limited to ten, resulting in a corpus of 246 texts. In addition, authors represented in the “Success” category are constrained to appear at least once in the other set, ensuring that predictive performance is not trivially driven by author identity. See Section 6.4.1 for the evaluation protocol and split strategy.
Across all configurations, the same feature extraction pipeline and success operationalization are used, ensuring that results from corpus-driven exploration, statistical testing, and supervised modelling remain methodologically comparable. This flexible but principled corpus design allows different analytical approaches to be applied to the same textual material, while tailoring data organization to the assumptions and requirements of each method.
It is important to note that download counts are extremely right-skewed: values range from 155 to 26,832 downloads at the snapshot date (median = 282; mean ≈ 759), implying that a small number of titles account for a disproportionate share of total attention. By contrast, the

lower tail contains texts with only around 150–200 downloads. These examples clarify that "success," as operationalized here, does not reflect incremental differences but rather a strongly unequal attention distribution, which motivates the percentile-based definition used throughout the study.

Within the works by Arthur Conan Doyle, download counts span from high values (e.g., The Disappearance of Lady Frances Carfax with 3,437 downloads; His last bow with 3,013) to relatively low values near the bottom of the overall corpus (e.g., The Valley of Fear with 171 downloads). This internal spread suggests that, even for a canonical author, the Project Gutenberg attention signal is heterogeneous across titles and editions, reinforcing the importance of modelling success at the level of individual texts rather than assuming uniform author-level popularity.

Finally, looking specifically at the "successful" subset defined as the top 10% of the corpus by download count, author recurrence is concentrated in a small number of names. In the top 28 titles, Agatha Christie appears most often (5 titles), followed by Doyle (4 titles) and Van Dine (3 titles); Arthur Morrison also appears more than once (2 titles). Most other authors in the top-decile set occur only once. This pattern is consistent with the broader motivation for later balancing choices in the predictive modelling configuration (capping texts per author), because a small number of prolific or high-attention authors can otherwise dominate the "success" signal.

All resources required to reproduce the analyses in this thesis are publicly available in a GitHub repository. The repository includes the main JSON file containing the extracted feature representations for all books in the study, together with an aligned metadata table (CSV) keyed by the same TextID and reporting Type, Issued date, Title, Language, Authors, Subjects, LoCC, Bookshelves, and Downloads. To ensure full methodological transparency, the repository will also provide the complete codebase used for the statistical analyses and for the supervised NLP/predictive experiments, along with documentation describing how to rerun each stage of the pipeline. Finally, the repository includes a corpus folder containing the full texts analyzed in the thesis; these texts are sourced from Project Gutenberg and are included to facilitate replication and long-term reproducibility, in addition to the corresponding identifiers and metadata.

3.8 Feature extraction pipeline

After preprocessing, each text was transformed into a structured, array-like representation through an automated feature extraction pipeline implemented in Python. The goal of this

pipeline was to generate a consistent set of quantitative features that could support statistical analysis, corpus-level exploration, and supervised classification while remaining comparable across texts of different lengths.

All books were processed using spaCy, an open-source natural language processing library that provides efficient and linguistically informed annotation. The pipeline relied on spaCy's transformer-based English model *en_core_web_trf*, with named entity recognition disabled to reduce computational overhead. This model produces detailed token-level annotations, including lemmas, part-of-speech tags, morphological features such as tense and grammatical person, and dependency relations. Sentence boundaries were identified using spaCy's dependency-based sentence segmentation provided by the *en_core_web_trf* model.

From these annotations, a set of stylistic and syntactic features extracted from the document level was obtained. All features were normalized to facilitate comparison across texts of varying lengths. Density-based measures were computed as proportions relative to the total number of tokens, excluding punctuation and white spaces.

Grammatical preferences were captured through normalized part-of-speech profiles. For each document, the relative frequency of each part-of-speech category assigned by spaCy (e.g., nouns, verbs, adjectives, and adpositions) was calculated. These values are treated as a compact grammatical signature intended to capture stylistic tendencies in lexical and syntactic choice and to support systematic comparison across texts.

To approximate phrase structure and syntactic complexity in a scalable manner, the pipeline extracted several density measures derived from dependency parses. Noun phrase density was operationalized as the number of noun chunks identified by spaCy, normalized by token count, while prepositional phrase density was approximated by counting dependency relations labeled as "prep" in spaCy's dependency scheme, normalized by token count. Verb phrase density was approximated by identifying verbs whose dependency children included core verbal attachments, such as auxiliaries, objects, or adverbial modifiers. Subordination density was computed as the proportion of tokens participating in clausal embedding relations, operationalized via spaCy dependency labels: advcl (adverbial clause), relcl (relative clause), ccomp (clausal complement), and xcomp (open clausal complement). The label set is fixed and reported to ensure reproducibility. Together, these measures are meant to provide interpretable proxies for syntactic packaging, degrees of clausal embedding, and the balance between phrasal and clausal constructions.

Readability was assessed using the Flesch Reading Ease score, calculated at the document level using the standard formula based on average sentence length and syllables per word, which is the following:

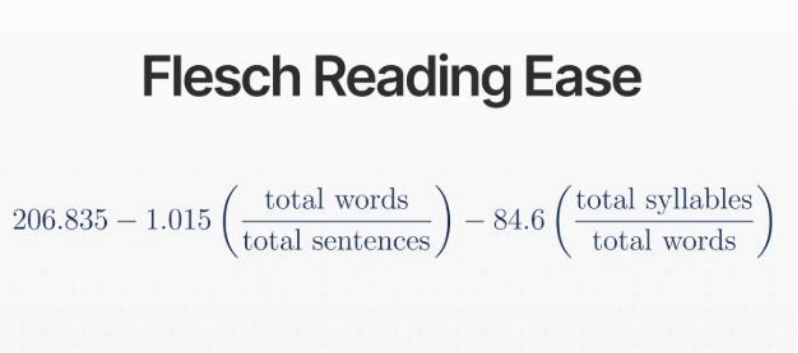

Figure 2: Flesch reading ease formula

Sentence counts were derived from spaCy's sentence segmentation, while syllable counts were estimated using Pyphen, a hyphenation-based library that approximates syllable boundaries in English. Syllable counts were aggregated across alphabetic tokens only.

The pipeline also incorporated lexicon-based semantic and affective features by matching tokens, using both surface forms and lemmas, against established psycholinguistic resources. This was employed to extract concreteness, valence and arousal. For each lexicon, document-level values were computed as the mean score across all matched tokens, yielding a single summary measure per text.

Semantic concreteness was computed using the concreteness ratings provided by Brysbaert et al. (Brysbaert et al., 2014), which offer human judgments of how concrete or abstract English lemmas are. The concreteness feature corresponds to the average concreteness score of all tokens in the document that were matched to the lexicon.

Affective dimensions were derived from the NRC Valence/Arousal/Dominance lexicon (Mohammad, 2018), focusing on valence and arousal. For each document, mean valence was computed across all matched tokens, while arousal was operationalized as the mean arousal score, reflecting emotional intensity irrespective of direction.

Narratorial stance and grammatical perspective were captured through pronoun use and auxiliary tense distribution. First-person and third-person pronoun usage was quantified using spaCy's morphological annotations for grammatical person, restricted to pronouns and normalized by the total number of pronouns in the document. This normalization reflects relative preference rather than absolute frequency. Temporal framing was approximated

through the distribution of auxiliary verb tense, computed as the proportion of auxiliary tokens marked as past or present tense relative to all auxiliary tokens.

Finally, the pipeline included a sentence-level heuristic intended to approximate descriptive stylistic passages in a scalable and reproducible manner. For each sentence, adjective density was calculated as the proportion of adjectives relative to all tokens. Sentences were classified as descriptive if adjective density exceeded a fixed threshold of 0.2 and if at least one stative verb lemma (such as *be*, *seem*, or *appear*) was present.

This operationalization is not intended as a formal linguistic definition of description, but as a conservative estimate designed to identify sentences characterized by a relatively high degree of adjectival modification and stative predication, which are commonly associated with descriptive discourse. The adjective-density threshold was chosen to capture locally adjective-rich sentences while avoiding the inclusion of minimally modified narrative clauses, and the stative-verb constraint serves to distinguish descriptive passages from action-driven narration. Descriptive density was then defined as the proportion of such sentences relative to the total number of sentences in the document, providing a consistent and interpretable measure that can be compared across texts at scale.

For each processed text, all extracted features were stored in a structured JSON file. This output includes metadata describing corpus size and processing details, document-level feature records, and logs of any processing errors. This format was chosen as it supports reproducibility and enables straightforward integration with statistical software and machine learning workflows. The following section documents in detail the information contained by each part of the file.

It is important to note that while only a selected subset of these variables was employed in this thesis (concreteness, arousal, first and third-person pronouns and subordination density), all variables extracted are reported here as they are consistent with previous research and could therefore prove useful for future research.

3.9 Output data structure

The output obtained from the features extraction pipeline was a JSON document with two main keys:

1. meta: This field included metadata describing the corpus and extraction process This information was useful for logging and sanity checking, but it was not used as input for the model. This section of the JSON file included:
    - Corpus directory

- Number of files (books) found
- Number of files successfully analyzed
- Number of errors occurred
- An array storing book success labels aligned with the order in which books appeared in the “results” array
- Number of books tagged as “success”
- The ids of the books marked as success

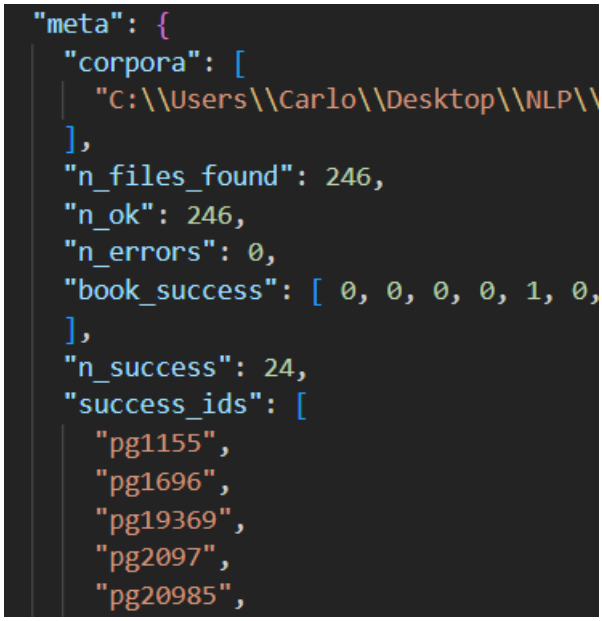

Figure 3: “meta” field information

2. results: an array in which each element corresponded to one book/document and contains all extracted features pertaining to it.

Each element of the “results” array corresponds to a single document and contains three types of information. First, a unique string identifier (e.g., "book": "pg10059") is used to reference the source text. Second, several nested JSON objects group together related linguistic and psycholinguistic features, such as part-of-speech distributions, phrase-level densities, readability measures, pronoun usage statistics, and auxiliary verb distributions. Each of these nested objects contains only numeric values. Third, the document representation includes scalar numeric features, such as concreteness, valence, arousal, and descriptive density, which are expressed as single numeric values and are not

embedded within higher-level structures. Each element has the following structure (as also represented in Figure 4):

1) book (string)
2) POS (dictionary of part-of-speech relative frequencies; numeric values)
3) Phrase (dictionary of phrase-related densities; numeric values)
4) Readability (dictionary containing readability metrics; numeric values)
5) Concreteness (numeric scalar)
6) Valence, Arousal (numeric scalars)
7) PPronoun (dictionary of pronoun-related proportions; numeric values)
8) AuxiliaryDistribution (dictionary; numeric values)
9) DescriptiveDensity (numeric scalar)

```
{
  "POS": {
    "pos_verb": 0.14185178771037674,
    "pos_propn": 0.05859848544484154,
    "pos_adj": 0.06702142733441441,
    "pos_part": 0.030684698584630885,
    "pos_adv": 0.05242690933136011,
    "pos_sconj": 0.02278130264338681,
    "pos_pron": 0.13404285466882881,
    "pos_cconj": 0.030212384086150165,
    "pos_num": 0.006187319930097454,
    "pos_intj": 0.0027079364579561376,
    "pos_x": 0.0006927279311050585,
    "pos_punct": 0.00023615724924036085
  },
  "Phrase": {
    "np_density": 0.2999826818017224,
    "pp_density": 0.10211439457153203,
    "vp_density": 0.06097580175386117,
    "subordinate_density": 0.07390147519561692
  },
  "Readability": {
    "flesch_reading_ease": 85.31757508639326
  },
  "Concreteness": 3.1921896202621576,
  "Valence": 0.15203371535736612,
  "Arousal": 0.23194905188580955,
  "PPronoun": {
    "first_person_pronouns": 0.21634954193093728,
    "third_person_pronouns": 0.4969462062485318
  },
  "AuxiliaryDistribution": {
    "past": 0.32895992366412213,
    "present": 0.35186068702290074,
    "future": 0.0
  },
  "DescriptiveDensity": 0.018496276723516693
},
```


Figure 4: “results” field information

This chapter has outlined the construction of the textual corpora and the operational choices underlying the analyses presented in the study. It has described the criteria for text selection, preprocessing, and normalization, and motivated the use of a deliberately pragmatic definition of literary success based on Project Gutenberg download counts. Throughout,

success has been treated as a corpus-internal analytical category rather than as a claim about intrinsic literary value or historical reception.

The chapter has also introduced the feature extraction pipeline used to derive grammatical, syntactic, semantic, affective, and stylistic measures from the texts. Together, these features provide a structured and interpretable representation of narrative style that supports both comparative analysis and supervised modelling.

On this methodological foundation, the following chapter turns to the corpus analysis stage of the study.

Chapter 4: Lexical domain analysis across the Main corpus

4.1 Analysis focus and lexical categories

This chapter operationalizes a set of lexical domains to test whether “successful” detective fiction works differ systematically from other books in the same genre by how they lexicalize investigation, cognition, perception, and romance. The categories are motivated by a convergence of historical genre prescriptions, Holmes-centred scholarship on detection as evidence-driven reasoning, and computational findings about success-related lexical tendencies, as will be explained in what follows. The Main corpus configuration described in Chapter 3 was employed for all analyses reported in this chapter.

Rather than claiming that these elements are exclusive to “Success”, the analysis asks whether “Success” and “Other” differ in how these functions are employed: for example, whether “Success” favors verbs of investigative noticing and reasoning over more purely sensory description, and whether romance is encoded more selectively.

The four categories were selected to provide an interpretable lens on concreteness-related stylistic patterning and genre-relevant thematic emphasis in the two subcorpora. Concreteness is later (Chapter 5) treated as one of the key variables in the quantitative analyses. Here, however, it is approached at the corpus level through a set of motivated lexical domains that capture different “concrete” components of detective narration: perceptual access to the world (perception verbs), explicit cognitive processing of evidence (thinking verbs), procedural and institutional investigative framing (forensic/police lexicon), and the presence/absence of romance-related language as a competing narrative focus. In other words, this chapter provides a transparent decomposition of concreteness-related stylistic choices complementing following stages by showing which semantic fields contribute to differences between the subcorpora. Importantly, this corpus-level analysis is not intended as a “compact” or one-to-one counterpart of the computational model of Chapter 6 or of the statistical regression in Chapter 5. The two stages do not rely on the same observations by design and therefore are not strictly reducible to one another. Rather, they are conceptually linked and mutually informative: the present chapter offers an interpretable, domain-motivated view of lexical patterning close to the textual surface, while the following analyses test whether broader stylistic variables derived from the texts capture systematic differences and predictive signal. Together, they provide complementary ways of approaching the same research object: one foregrounding explicit semantic fields, the other

capturing aggregate tendencies and deeper mechanisms that may not be visible to the naked eye.

Detective fiction is structurally anchored in the management of evidence, inference, and investigative procedure. As described in Chapter 2, the Holmes canon is particularly important in establishing an “investigative register” centered on observable traces and their interpretation. Scholarship on Sherlock Holmes and linguistic analysis further links Conan Doyle’s detective method to the cultural development of forensic reasoning and situates forensic linguistics more generally as the application of linguistic knowledge to legal and criminal contexts (Boucher & Perkins, 2020; Vezzani & Di Nunzio, 2019). This makes forensic/investigative vocabulary a principled category to test, not because greater lexical concreteness is expected to correlate with success, but because prior work suggests the opposite direction. As discussed in Chapter 1, studies such as Ashok et al. (2013) and Da Silva et al. (da Silva et al., 2024) report that more successful texts tend to exhibit lower concreteness, and shift toward more abstract lexical choices. From this perspective, foregrounding the investigative apparatus (clues, evidence, witnesses, interrogation, courts) is not necessarily a matter of producing vivid sensory depiction; it may instead reflect a preference for institutional, procedural, and inferential framing: language that organizes evidence, reasoning, and social/legal categories. In this sense, the forensic/investigative domain provides an interpretable lens for assessing whether “Success” aligns with a move away from concrete depiction and toward abstract, reasoning-oriented narration.

A second, closely related domain concerns the language of cognition: verbs and nouns that explicitly encode inference, belief, memory, doubt, conclusion, and understanding. This category is motivated by previous studies conducted by Ashok et al. (2013) who, as discussed in Chapter 1, report that successful fiction seems to favor cognitive verbs and reporting structures, whereas less successful texts show different lexical emphases. In genre terms, this also aligns with classic “fair play” conceptions of detective fiction, where the reader is invited into an intellectual puzzle shaped by the detective’s reasoning. As Van Dine (Van Dine, 2015) argues:

The truth of the problem must at all times be apparent—provided the reader is shrewd enough to see it. By this I mean that if the reader, after learning the explanation for the crime, should reread the book, he would see that the solution had, in a sense, been staring him in the face—that all the clues really pointed to the culprit—and that, if he had been as clever

as the detective, he could have solved the mystery himself without going on to the final chapter.

Detective fiction also depends on the staging of observation: what can be seen, heard, noticed, watched, or sensed, and how such perceptions become “clues.” This connects with the Holmes tradition of trace-based epistemology and procedural observation (the idea that narrative plausibility and payoff are built from perceivable details and their interpretation). In this analysis, the perception category is used to test whether the language through which evidence is accessed is primarily anchored in physical sensation or in more mentally mediated, interpretive processes, and whether these tendencies differ systematically between the “Success” and “Other” corpora. As Van Dine noted: “The detective novel must have a detective in it; and a detective is not a detective unless he detects” (Van Dine, 2015). The inclusion of a category linked with perception may help in determining whether successful texts emphasize “access to evidence” through sensory reporting and noticing to a greater extent than less successful ones.

The final category aims to establish how romantic language is used across subcorpora. Again, Van Dine’s Twenty Rules for Writing Detective Stories argue that there should be “no love interest,” on the grounds that romance clutters “a purely intellectual experience” with irrelevant sentiment. This makes romance lexis an especially interpretable diagnostic: if the “Success” subcorpus aligns more closely with classical puzzle-oriented norms, it should show reduced prominence of overt romance vocabulary relative to the comparison corpus (or at least a different distribution of how romance is lexicalized). On this basis, the expectation is that the “Success” subcorpus will display a lower prominence of explicitly romance-related vocabulary than the comparison corpus. The purpose of including romance as a lexical domain here is therefore diagnostic: rather than assuming the rule holds uniformly, the analysis tests whether reduced romance marking is actually a distinguishing characteristic of the texts labelled “Success” in this dataset, or whether romance is present but simply showcased differently. This also resonates with how love is portrayed in Holmes’ adventures: in Conan Doyle’s canonical stories, romance is generally not developed as a sustained subplot, and even cases involving it are typically instrumental to the mechanics of the mystery rather than emotional fulfilment. Even in “A Scandal in Bohemia,” in which the narrative revolves around blackmail and a diplomatic marriage connected to Irene Adler’s possession of a photograph, the revolving plot focuses on solving the mystery rather than resulting in a romantic arc for Holmes himself.

Together, these categories form a coherent interpretive set. Forensic vocabulary indexes the institutional-investigative frame, cognition terms index explicit reasoning and epistemic stance, and perception verbs index access to evidence through observation. Importantly, the present analysis does not assume that these domains are unique to “Success” or absent from “Other”; rather, it examines whether the two subcorpora differ in the relative weighting and lexical realization of these domains. In line with the discussion in Chapter 1, the expectation is therefore framed in terms of how these functions are employed: “Success” may rely relatively more on abstract, inferential, and institutionally anchored cues (e.g., reasoning-oriented framing), whereas “Other” may give relatively more space to sensory/experiential rendering. Romance language is included as a diagnostic domain to test whether reduced overt romance marking, or a shift toward more abstract romance concepts, distinguishes “Success” from the comparison corpus.

4.2 Methods

This section describes the corpus analysis procedure used to compare the “Success” and “Other” subcorpora and to compute normalized frequencies for each lexical domain. The goal of this stage was descriptive and hypothesis-guided, and it aimed to document systematic differences in the prominence of theoretically relevant semantic domains. The analysis was motivated by genre-based expectations in early twentieth-century detective fiction, as explained in the previous section. The analysis compared two subcorpora derived from the Main corpus configuration described in Chapter 3: a “Success” subcorpus (download-based “successful” texts) and an “Other” subcorpus (the remaining detective-fiction texts in the same dataset). Since the two subcorpora differ substantially in size, all comparisons are reported as normalized frequencies. Token totals for the two subcorpora are:

- “Success”: 1,592,414 tokens
- “Other”: 12,841,689 tokens

Rates are expressed as occurrences per 10,000 tokens, allowing direct comparison across subcorpora. All texts were cleaned using the preprocessing pipeline described in Chapter 3, including the removal of Project Gutenberg boilerplate. For all analyses the files were also lowercased to minimize capitalization-driven variation. All analyses were carried out using AntConc (Anthony, 2023).

To improve replicability and reduce ambiguity caused by inflection and part-of-speech homography, the two subcorpora were also lemmatized and POS-tagged prior to analysis using spaCy. Each token was converted into a searchable compound form that combines its lemma with its POS label (e.g., *saw* → *see*POSVERB; *witness* as a noun → *witness*POSNOUN). In the tagged versions used for this chapter, lemma and POS were concatenated, so that AntConc treats each annotated token as a single word type. These tagged corpora were then loaded into AntConc as plain-text corpora and queried manually, with the POS labels used to restrict searches to the intended grammatical category.

Analyses were conducted separately for each semantic domain. For the perception and thinking domains, only verb tokens were counted. Operationally, this was implemented by searching for lemma+POS forms ending in *POSVERB* (e.g., *notice**POSVERB*). For the forensic/legal domain, both nouns related to institutions and procedural verbs were retained as part of the domain; therefore, searches targeted the relevant lemma+POS tokens for each item. The romance domain was treated as a broader lexical field and was not restricted to a single POS class, since salient items occur across nouns, adjectives, and verbs.

For each domain, candidate items were finalized through an iterative AntConc workflow consisting of frequency inspection, concordance validation, and manual selection of the most stable contributors. Specifically, after loading a subcorpus, a Word List was generated in AntConc and used as an empirical guide to which domain-relevant items were frequent enough to support stable normalized rates. From this list, the 15 most frequent domain-relevant items by raw occurrences in the target corpus were selected and reported in the Result section tables. Each candidate was then validated using AntConc's Concordance (KWIC) tool: items were retained only when KWIC inspection indicated that the dominant usage instantiated the intended semantic domain and did not introduce off-domain noise. Where a candidate term showed frequent idiomatic, discourse-structural, or non-domain uses, it was excluded. This exclusion criterion was applied consistently across all four domains.

Raw frequency counts were obtained from the concordance hit totals and then normalized to occurrences per 10,000 tokens using the token totals reported above. For each item, the direction of the difference and the magnitude of the difference were computed as a simple delta. In addition to within-domain frequency comparisons, Keyword List comparisons were also run with "Success" as target and "Other" as reference and then reversed, as a robustness check to identify items that are disproportionately associated with either subcorpus rather than simply frequent in both.

Four main semantic domains were operationalized and analyzed:

1) Perception/observation verbs
2) Thinking/inference verbs
3) Forensic/legal lexicon
4) Love/romance language

4.3 Results

4.3.1 Perception and observation language

As shown in Table 1, the perception domain divides into two broad patterns: verbs that primarily encode direct sensory experience and bodily perception (such as *see*, *hear*, *feel*), and verbs that more often signal investigative attention and interpretation (such as *notice*, *observe*, *witness*). Several high-frequency perception verbs occur at broadly similar rates in both corpora, as expected for narrative prose; however, the following subset of items diverges in a way that suggests different emphases in how evidence is accessed and processed. Because the corpora were lemmatized and POS-tagged prior to querying, the AntConc counts in this chapter are computed over single-word lemmas (e.g., LOOK_POSVERB), regardless of whether the verb is followed by a particle or preposition in the running text. As a result, multiword constructions such as *look at/into/for* are included in the frequency of *look* but not distinguished from non-phrasal uses.

Table 1: distribution of perception tokens in the subcorpora

lemma	S_raw	O_raw	S_per10k	O_per10k	direction	S_vs_O
see	5422	43064	34.3899	33.8063	S	+1.73
hear	1705	14007	10.8142	10.9958	O	-1.65
feel	966	9298	6.1270	7.2992	O	-16.06
watch	477	4664	3.0254	3.6614	O	-17.37
notice	438	2443	2.7781	1.9178	S	+44.86
listen	334	3452	2.1184	2.7099	O	-21.83
observe	316	2192	2.0043	1.7208	S	+16.48
stare	299	2521	1.8965	1.9790	O	-4.17
touch	171	1376	1.0846	1.0802	S	+0.41
peer	73	930	0.4630	0.7301	O	-36.58
grasp	73	699	0.4630	0.5487	O	-15.62
glare	50	356	0.3171	0.2795	S	+13.45

witness	46	332	0.2918	0.2606	S	+11.97
perceive	45	328	0.2854	0.2575	S	+10.83
detect	41	306	0.2600	0.2402	S	+8.24

Overall, perception language is present in both subcorpora, so the difference is not simply that "Success" contains more perception. Rather, what differs is how perception is represented and what narrative function it serves. In the "Other" subcorpus, perception is more often realized through bodily and sensory experience (e.g., *feel*), consistent with a more concrete, experiential mode of depiction. In contrast, the "Success" subcorpus places greater weight on perception as evidential uptake and investigative attention (e.g., *notice, observe, witness*), and this emphasis aligns with the broader tendency discussed in Chapter 1 for successful texts to rely more on abstract, inferential framing than on vivid sensory concreteness. In this sense, "Success" does not increase perception per se, but foregrounds perception as a route to interpretation and reasoning.

4.3.2 Thinking and inference language

Cognitive language is present in both corpora, but the relative weighting of epistemic stance differs. The most frequent cognition verbs behave differently depending on whether they encode certainty/knowledge, controlled inference, or open-ended speculation (See table 2).

Table 2: distribution of lemmas detailing cognitive functions in the subcorpora

lemma	S_raw	O_raw	S_per10k	O_per10k	direction	S_vs_O
know	5494	40700	34.8465	31.9505	S	+9.07
think	3684	29489	23.3663	23.1496	S	+0.94
believe	801	6985	5.0805	5.4834	O	-7.35
suppose	794	6068	5.0361	4.7635	S	+5.72
remember	772	4585	4.8965	3.5993	S	+36.04
understand	637	4227	4.0403	3.3183	S	+21.76
expect	516	3591	3.2728	2.8190	S	+16.10
hope	436	4538	2.7654	3.5624	O	-22.37

guess	432	4758	2.7400	3.7351	O	-26.64
forget	368	3518	2.3341	2.7617	O	-15.48
wonder	364	4200	2.3087	3.2971	O	-29.97
suspect	280	1511	1.7759	1.1862	S	+49.71
imagine	261	1510	1.6554	1.1854	S	+39.65
consider	257	1792	1.6301	1.4068	S	+15.88
decide	232	2945	1.4715	2.3119	O	-36.35

"Success" shows more instances of verbs that encode knowledge, memory retrieval, and suspicion (*know*, *remember*, *suspect*), while "Other" works favor verbs associated with conjecture and open-ended uncertainty (*wonder*, *guess*). This could indicate that "successful" detective narratives tend to be more explicitly organized around controlled inference and epistemic management.

4.3.3 Forensic, legal, and police procedural lexicon

The forensic/procedural domain yields one of the clearest stylistic signatures. The top contributors in the "Success" subcorpus strongly emphasize institutional roles and investigative procedure (such as *evidence*, *suspect* and *police*).

Table 3: distribution of forensic and legal lemmas in the subcorpora

item	S_raw	O_raw	S_per10k	O_per10k	direction	S_vs_O
fact	1146	5283	7.2687	4.1473	S	+75.26
follow	820	7595	5.2010	5.9623	O	-12.77
police	678	3455	4.3003	2.7123	S	+58.55
detective	483	2963	3.0635	2.3260	S	+31.70
sergeant	398	496	2.5244	0.3894	S	+548.33
evidence	359	1686	2.2770	1.3236	S	+72.05
examine	347	1527	2.2009	1.1987	S	+83.61
suspect	315	1617	1.9979	1.2694	S	+57.38
inspector	271	1300	1.7189	1.0205	S	+68.44
search	188	1389	1.1924	1.0904	S	+9.35
arrest	181	1033	1.1480	0.8109	S	+41.56
court	167	1115	1.0592	0.8753	S	+21.01

officer	166	1853	1.0529	1.4546	O	-27.62
clue	165	795	1.0465	0.6241	S	+67.68
motive	158	546	1.0021	0.4286	S	+133.79

A salient finding is that not all investigation-related vocabulary patterns in the same way. While Table 3 shows that density of institutional roles and procedural/evidential terms are consistently higher in the "Success" subcorpus, additional checks on more generic action-oriented verbs indicate that these are more characteristic of the "Other" subcorpus: for example, *follow* occurs more often in "Other" (5.96 vs 5.20), whereas *search* is very similar across corpora and slightly more present "Success" (1.19 vs 1.09).

4.3.4 Love and romance language

The love/romance domain shows a nuanced pattern: some high-frequency relationship markers are more common in "Other", while "Success" shows relatively higher rates for terms that are more explicitly linked with the concept of romance (*romance*, *passion*, *intimate*) rather than everyday relational tokens (*love*, *kiss*, *wife*).

Table 4: distribution of love related lemmas in the subcorpora

lemma	S_raw	O_raw	S_per10k	O_per10k	direction	S_vs_O
dear	517	4508	3.2792	3.5389	O	-7.34
love	365	4529	2.3151	3.5554	O	-34.88
wife	277	2957	1.7569	2.3213	O	-24.31
husband	231	2049	1.4652	1.6085	O	-8.91
marry	195	1713	1.2368	1.3447	O	-8.02
desire	191	1426	1.2114	1.1194	S	+8.22
engage	135	1076	0.8563	0.8447	S	+1.37
marriage	98	741	0.6216	0.5817	S	+6.86
intimate	75	315	0.4757	0.2473	S	+92.38
romance	74	240	0.4694	0.1884	S	+149.15
passion	69	326	0.4376	0.2559	S	+71.02
lover	47	507	0.2981	0.3980	O	-25.11
romantic	44	185	0.2791	0.1452	S	+92.22
kiss	36	740	0.2283	0.5809	O	-60.70
affection	32	291	0.2030	0.2284	O	-11.12

This pattern suggests that “Success” is not simply “less romantic” across the board. Instead, Success appears less reliant on frequent, domestic relationship markers, while showing greater relative prominence of vocabulary associated with romance as a thematic motif or intensified register. In “Other”, romance language is more often expressed through everyday relational terms and actions. Importantly, the “Success” profile suggests that romance tends to appear more as an abstract or intensified thematic register than as everyday relational practice. By contrast, the “Other” subcorpus relies relatively more on concrete markers and actions, which are closer to romance as enacted interpersonal life. This distribution is also compatible with how Doyle treats romance in his detective fiction works: in the canonical Sherlock Holmes stories, romantic involvement is not developed as a sustained subplot for the detective, and “love” is more typically present as motive, leverage, or social context within the case rather than as enacted domestic relationship life.

4.4 Discussion

Across the selected domains, the corpus analysis reveals consistent and interpretable differences that connect directly to genre expectations for detective fiction and to historically grounded claims about what “successful” detective stories emphasize.

Overall, the “Success” subcorpus appears less concrete in a descriptive, sensory sense and more oriented toward abstract, procedural, and epistemic framing. Across the analyzed categories successful texts rely less on language that directly evokes lived sensory experience and more on lexicon that foregrounds reasoned interpretation and institutional procedure. “Success” has higher density of inference-oriented items and procedural-investigative vocabulary, all of which are more abstract in the sense that they encode cognitive stance, evidential reasoning, and social institutions rather than immediate physical sensation or domestic affect. In this respect, the “Success” subcorpus seems to concentrate lexical attention on the problem-solving apparatus of detection: what is known, inferred, and officially investigated, while the “Other” subcorpus retains relatively more language associated with physical sensation and more commonly used relational expression.

First, the perception-domain findings suggest that the two subcorpora do not differ in the overall amount of perception language, but in the type of perceptual access they lexicalize. The “Other” subcorpus shows higher rates of *hear*, *listen*, *feel*, and *watch*, verbs that often refer to sensory experience and receiving information though one’s body. By contrast,

“Success” shows higher rates of *notice* and *observe*, which more typically construe perception as selective attention and interpretive uptake, alongside smaller advantages for explicitly evidential verbs such as *perceive*, *detect*, and *witness*. This split is consistent with the structure of detective fiction outlined in Chapters 1 and 2, where what matters is not merely that something is perceived, but that it is perceived as evidence and integrated into an inferential chain. In this light, *notice* signals the registration of salient details, *observe* implies deliberate scrutiny, and verbs such as *perceive*, *detect*, and *witness* construe perception as evidentially consequential rather than purely experiential. Overall, “Success” appears to lexicalize perception less as raw sensation and more as investigative attentional processing, that is, perception oriented toward clue formation and reasoning.

Second, the thinking/inference domain shows the clearest alignment with a rational-problem-solving ideal. While *think* is stable across corpora, “Success” has more occurrences of *know, remember, understand, expect, and suspect.* These items encode knowledge claims, reconstruction of past information, and controlled epistemic stance, precisely the kinds of cognitive operations that structure classical detective plots. On the contrary, “Other” has more cases of *wonder* and *guess*, which seem to signal a more speculative stance. This indicates the use of a different stance profile: “Success” appears more committed to evidential reasoning and suspicion management, while Other relies relatively more on conjecture markers.

However, as mentioned above, not all investigation-related vocabulary patterns in the same way. This contrast reinforces that the “Success” signal is not simply greater action density, but the narration of action through institutionally anchored procedure and evidential work. One plausible explanation is that such institutional anchoring enhances narrative credibility and reader satisfaction: by embedding the detective’s actions within recognizable structures of policing and adjudication, the pathway from clue to solution may feel more legitimate and anchored. At the same time, detective protagonists often operate in productive tension with authority, bending rules or acting independently while still collaborating with institutions for access, information, or closure. The institutional frame can preserve plausibility without reducing the dramatic appeal of the exceptional detective.

Finally, the romance results provide a nuanced perspective on the claim that successful detective fiction suppresses romance. Several high-frequency romance markers are indeed less present in “Success”. At the same time, “Success” has more romance related terms referring to abstract concepts, suggesting that romance vocabulary, when it appears in successful texts, may be used more selectively and in more thematically marked contexts

rather than as pervasive relational background. In this sense, the results are consistent with the Holmes critical tradition: romance is not absent, but it is more often present as a concept (motive, atmosphere, register) than through concrete enactments of relationship life, which aligns with the genre logic that keeps emotional subplots secondary to deduction and procedural clarity.

A brief methodological caveat is warranted. The domain-based patterns discussed in this section are derived from a restricted, manually motivated set of lexical items and from concordance-based inspection of their local contexts. This approach is intentionally interpretable, but it is also necessarily selective and reductive: it does not model higher-level narrative organization directly and it cannot capture syntagmatic relations beyond what is visible in concordance lines. Accordingly, the claims advanced here should be read as descriptive tendencies in lexical signaling rather than as a direct account of narrative structure.

Taken together, these four domain analyses converge on a coherent stylistic portrait of the “Success” subcorpus: stronger procedural-investigative framing, paired with a stance profile that emphasizes knowing, remembering, and suspecting over wondering and guessing, and a lower frequency of concrete romance markers. These findings provide an interpretable bridge between genre theory and the statistical and modelling stages that follow, and motivate the expectation, informing subsequent analyses, that stylistic predictors tied to epistemic control, procedural narration, and attentional processing may correlate with success.

Chapter 5: Statistical analysis of stylistic predictors of success

5.1 Aim and rationale

This chapter concerns the statistical component of the thesis, contributing to addressing RQ1 by testing whether the five selected linguistic variables—Concreteness, Arousal, first-person pronoun density, third-person pronoun density, and subordination density—are statistically associated with Project Gutenberg downloads. The corpus construction and feature-extraction pipeline were presented in Chapter 3, together with the motivation for treating downloads as a pragmatic measure of present-day attention rather than as a direct indicator of historical popularity or intrinsic literary quality. All analyses described in this Chapter use the Statistical Analysis configuration of the corpus. This included the use of the JSON file containing all features extracted from the 281 books and the .csv file containing metadata.

Because Arthur Conan Doyle is both a canonical reference point for the genre and recurrent among the highest-download texts in the corpus, the statistical results are interpreted throughout considering the Holmes-focused discussion in Chapter 2. The aim, however, is not to model “Doyle versus others,” but to test whether the same document-level stylistic dimensions that characterize detective narration in general can explain variation in download intensity across authors within a shared period and genre.

A central methodological challenge is that raw download counts are not directly comparable across books that have been available on Project Gutenberg for different amounts of time. A text uploaded earlier has had substantially more opportunity to accumulate downloads, introducing a strong exposure effect unrelated to the linguistic properties of the text. To address this issue, the analysis models downloads as a rate (downloads per unit time) by introducing an exposure offset based on the number of days between each book’s Project Gutenberg issue date and the dataset acquisition date (catalog snapshot). In other words, rather than asking whether a book has more downloads in total, the statistical model tests whether it has a higher expected download intensity given the time it has been available online.

Given that the dependent variable is a count (downloads) and that download distributions are typically highly skewed and overdispersed (Hilbe, 2011), the analysis relies on a Negative Binomial regression model with a logarithmic link function. Although downloads are integers, this model treats “success” as a continuous variable: it estimates how the

predictors shift the expected value of downloads (and, with the offset, the expected downloads/day) through multiplicative effects. Results are reported as Incidence Rate Ratios (IRR), which quantify the proportional change in download rate associated with a one-standard-deviation increase in a predictor while holding the other predictors constant (Kasyoki Muoka et al., 2016).

The analysis covers 281 detective fiction texts drawn from Project Gutenberg, with an outcome variable given by the download count observed at the snapshot date and an exposure term defined as the number of days between a text's release date and the snapshot date.

Download counts are non-negative integers and are strongly right-skewed and heavy-tailed. In the present dataset, raw downloads range from 155 to 26,832 (median = 282, IQR = 113; mean = 758.68, SD = 2,123.27). A small number of texts account for a disproportionate share of total downloads, which motivates the use of a Negative Binomial model and the robustness checks reported below (trimming and winsorization) to assess sensitivity to extreme high-download cases.

The number of days since release varies across texts (min = 39, max = 11,403, median = 4,261), reflecting the fact that older texts have had more time to accumulate downloads. Modelling therefore uses an exposure offset so that effects are interpreted on a rate scale rather than as raw totals.

All analyses were conducted in R (version 4.4.2). The main packages used were MASS (negative binomial regression), dplyr, tidyr, tibble, readr, and stringr (data import, cleaning, and reshaping), lubridate (date handling for exposure calculations), and jsonlite (parsing feature data stored in JSON format).

5.2 Data preparation and variables

The analysis uses the 281 detective fiction books in the Statistical Analysis corpus described in Chapter 3.

For each book, the dataset contains:

- Outcome: Project Gutenberg download count at snapshot date.
- Exposure: computed as the number of days between the book's issue date and the dataset acquisition date.
- Predictors: document-level measures of the five chosen features (see Chapter 3 for operational definitions and extraction).

To make effect sizes comparable across predictors measured on different scales, each predictor was converted to a z-score. A z-score expresses an observation in terms of how far it lies from the sample mean, measured in units of the sample standard deviation (Abdi, 2007):

$$z_i = \frac{x_i - M_x}{SD_x}$$

where x_i is the unstandardized value of a predictor for book i, M_x is the sample mean of that predictor across the corpus, and SD_x is its sample standard deviation. After this transformation, each standardized predictor has mean 0 and standard deviation 1. Therefore, $z_i = 0$ indicates a book at the average value of that predictor, $z_i = 1$ indicates a value one standard deviation above the average, and $z_i = -1$ indicates a value one standard deviation below the average.

A “1 standard deviation increase” therefore means increasing the predictor by an amount equal to its sample standard deviation on the original scale (e.g., if the SD of subordination density is 0.012, then a 1 SD increase corresponds to +0.012) (Livingston, 2004). This does not correspond to a fixed unit across variables; rather, it is a corpus-relative shift that captures a “typical” magnitude of variation for that feature. Using z-scores makes coefficients directly comparable across different features: each coefficient describes how the expected download rate changes when the predictor increases by one standard deviation, while the other predictors remain constant.

5.3 Model choice: negative binomial regression with exposure offset

Download counts are non-negative integers and, in the present dataset, they display a strongly right-skewed, heavy-tailed distribution. Raw downloads range from 155 to 26,832, with a median of 282 and a mean of 758.68, indicating that a small number of highly downloaded books substantially increases the average relative to the typical observation. This skewness is also evident in Figure 5, which shows the histogram of $\log(Downloads + 1)$: although the logarithmic transformation compresses extreme values and facilitates visualization, a pronounced right tail remains, confirming substantial heterogeneity in readership intensity across titles.

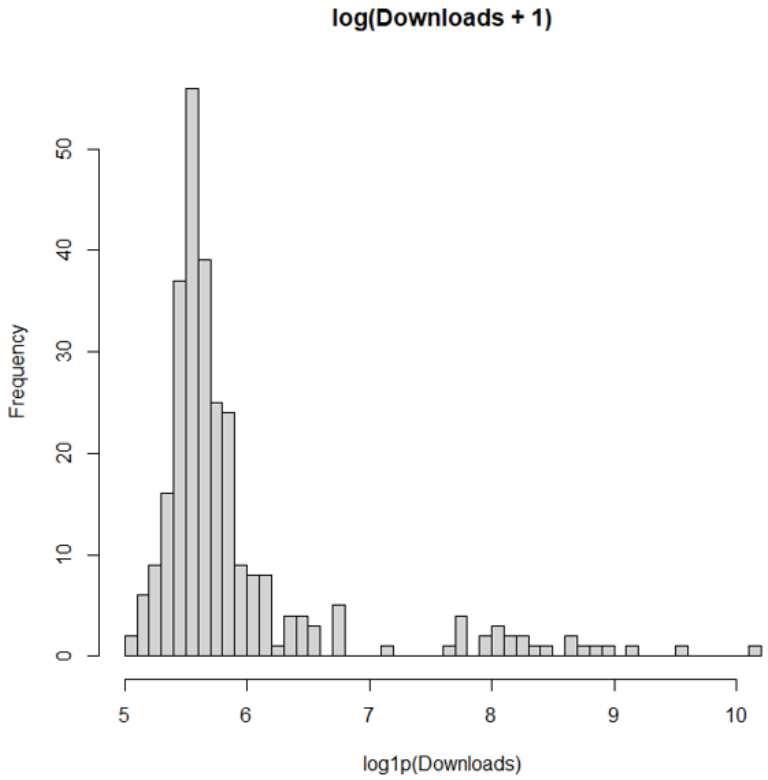

Figure 5: Frequency of compressed downloads

A further dataset-specific diagnostic concerns dispersion. A Poisson regression model assumes equidispersion, i.e. equality between the mean of distribution and its variance (Consul & Famoye, 1992). However, in this dataset the sample variance of downloads is 4,508,287, far exceeding the sample mean of 758.68. The resulting variance-to-mean ratio is 5,942.28, providing clear evidence of strong overdispersion relative to the Poisson assumption. For this reason, Poisson regression is too restrictive for this dataset, as it would underestimate variability and potentially yield overly optimistic standard errors and significance tests.

Accordingly, downloads were modelled using a Negative Binomial generalized linear model with a logarithmic link function. The Negative Binomial model avoids the Poisson equidispersion constraint by introducing an additional dispersion parameter, allowing the variance to exceed the mean and better accommodating heavy-tailed count outcomes. This makes the model a suitable choice for modelling overdispersed count data in settings where measures exhibit substantial heterogeneity across observations (Cameron & Trivedi, 2013).

Finally, as mentioned in 5.1, an important methodological issue is that download totals are not directly comparable across books that have been available on Project Gutenberg for different lengths of time. A text issued earlier has had more opportunity to accumulate downloads, independently of its linguistic profile. To account for this exposure effect, the

model includes an offset term computed as the number of days between a book's Project Gutenberg issue date and the dataset acquisition date. With this specification, coefficients are interpreted in terms of download intensity (downloads per day) rather than cumulative totals. In other words, the model estimates whether books with different stylistic profiles are associated with systematically higher or lower expected download rates, holding constant the time each book has been available online.

Because the chosen model uses a log link, coefficients are reported as Incidence Rate Ratios (SAS Institute Inc., 2004). For the standardized predictors, an IRR represents the change in the expected download rate associated with a one-standard-deviation increase in a predictor, holding other predictors constant (UCLA Institute for Digital Research & Education, n.d.). For example, an IRR of 1.50 corresponds to a 50% increase in expected downloads/day, whereas an IRR of 0.70 corresponds to a 30% decrease.

5.4 Results

Table 5 reports coefficient estimates on the log-rate scale (β), their exponentiated form as incidence rate ratios (IRR), 95% confidence intervals, and p-values for the five standardized predictors in the Negative Binomial model with an exposure offset. Figure 6 provides a visual summary of the same effects by plotting IRRs with 95% confidence intervals: the dashed vertical line at IRR = 1 marks the null effect (no change in expected downloads/day). Because predictors are z-scored, each IRR expresses the multiplicative change in expected download intensity (downloads/day) associated with a one-standard-deviation increase in the predictor, holding other predictors constant. Values of IRR greater than 1 indicate higher expected download intensity, whereas values below 1 indicate lower expected download intensity; confidence intervals that do not cross 1 indicate effects that are statistically distinguishable from zero at conventional thresholds.

Table 5: Negative Binomial model with exposure offset coefficient estimates (β), incidence rate ratios (IRR), 95% confidence intervals, and p-values for standardized predictors.

Predictor (z-scored)	β (log-rate)	IRR = exp(β)	95% CI for IRR	% change in down-loads/day	p-value
Concrete-ness	-0.843	0.431	[0.336, 0.556]	-56.9%	6.82e-16

Arousal	0.557	1.745	[1.369, 2.219]	+74.5%	7.50e-07
1st-person pronouns	0.094	1.098	[0.658, 1.834]	+9.8%	0.653
3rd-person pronouns	-0.008	0.992	[0.576, 1.699]	-0.8%	0.968
Subordina-tion density	-0.693	0.500	[0.378, 0.666]	-50.0%	2.35e-11

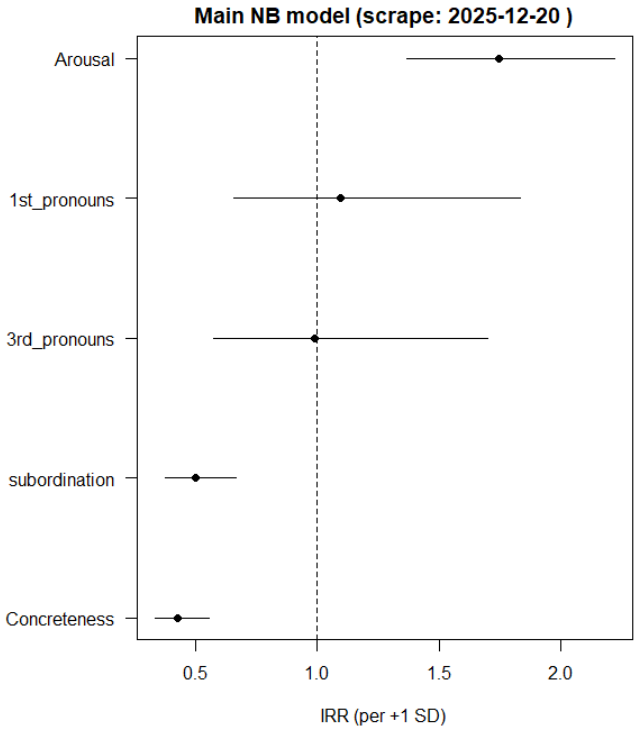

Figure 6: Forest plot of Incidence Rate Ratios from the main Negative Binomial model with exposure offset.

Three predictors show clear and statistically robust associations with download intensity. Concreteness is negatively related to downloads/day (β = −0.843; IRR = 0.431; 95% CI [0.336, 0.556]; p < .001), indicating that a one-standard-deviation increase in concreteness is associated with an expected download rate that is multiplied by 0.431: approximately a 56.9% decrease. In contrast, Arousal is positively associated with download intensity (β = 0.557; IRR = 1.745; 95% CI [1.369, 2.219]; p < .001): a one-standard-deviation increase in

arousal corresponds to an expected 74.5% increase in downloads/day. Finally, subordination density shows a strong negative association (β = −0.693; IRR = 0.500; 95% CI [0.378, 0.666]; p < .001), implying that a one-standard-deviation increase in syntactic subordination is associated with roughly a 50.0% decrease in expected download rate.

By contrast, the two pronoun-density measures do not show evidence of an independent association with downloads/day once concreteness, arousal, and subordination are controlled for. The estimated effects for first-person pronoun density (IRR = 1.098; 95% CI [0.658, 1.834]; p = .653) and third-person pronoun density (IRR = 0.992; 95% CI [0.576, 1.699]; p = .968) both have confidence intervals that include 1, indicating that the data are compatible with either modest increases or decreases in download intensity. In substantive terms, within this multivariate specification pronoun density does not appear to contribute a stable, separable signal of “success” as operationalized by download rate.

5.4.1 Robustness checks

To evaluate whether the main findings are driven by a small number of extremely high-rate titles, two robustness strategies were applied and the same Negative Binomial model with exposure offset was re-estimated (Table 6). First, books were ranked by download intensity (rate = Downloads/days_available) and the two most extreme cases were removed to reduce the influence of extreme observations on parameter estimates (Wilcox, 2023). Second, as an alternative to excluding observations, a winsorization procedure was used to cap downloads at the level corresponding to the 99th percentile of the rate distribution, scaled by each book’s exposure time, thereby limiting the influence of outliers while retaining all observations (Wilcox, 2023). Rather than deleting high-rate titles, this process keeps all books in the sample and replaces extreme download counts above the p99 threshold with the corresponding cap value.

Table 6: Robustness checks

Predictor (z-scored)	Main IRR [95% CI]	p	Trim_top2 IRR [95% CI]	p	Winsor_p99 IRR [95% CI]	p
Con-crete-ness	0.431 [0.336, 0.556]	< .001	0.576 [0.471, 0.705]	< .001	0.523 [0.420, 0.651]	< .001
Arousal	1.745 [1.369, 2.219]	< .001	1.634 [1.346, 1.979]	< .001	1.671 [1.355, 2.056]	< .001
Subordi-nation density	0.500 [0.378, 0.666]	< .001	0.604 [0.482, 0.759]	< .001	0.547 [0.430, 0.697]	< .001
1st-per-son pro-nouns	1.098 [0.658, 1.834]	.653	1.003 [0.679, 1.494]	.986	1.037 [0.672, 1.610]	.850
3rd-per-son pro-nouns	0.992 [0.576, 1.699]	.968	0.802 [0.535, 1.205]	.240	0.860 [0.548, 1.353]	.442

Across both robustness checks, the core pattern observed in the main model remains stable. Concreteness continues to show a negative association with download intensity (IRR < 1, p < .001 in both specifications), as does subordination density (IRR < 1, p < .001), while Arousal remains positively associated with downloads/day (IRR > 1, p < .001). By contrast, first- and third-person pronoun density do not reach statistical significance in either robustness specification and their confidence intervals overlap the null value (IRR = 1). Overall, these results strengthen confidence that the main associations (negative concreteness and subordination; positive arousal) are not solely artifacts of a few extreme observations in the heavy-tailed download distribution. Model fit was evaluated using likelihood-based measures. The full NB model significantly improved fit relative to the null model (LR $\chi^2(5) = 129.41$, $p < .001$), reducing deviance from 490.23 to 360.82 (≈26.4% deviance reduction). Model adequacy was assessed using likelihood-based criteria. The fitted Negative Binomial model had AIC = 4517.8; while AIC is primarily intended for

comparing alternative specifications rather than providing an absolute goodness-of-fit measure, it can be used to contrast the model with plausible alternatives.

5.5 Discussion

The negative binomial model estimates associations, not causal effects: it tests whether books that differ in the chosen document-level stylistic measures also differ, on average, in download intensity at the snapshot date.

Confronting results with the background literature reviewed in Chapter 1, the direction of effects is partly confirmatory and partly corrective. The positive association between lexical arousal and download intensity is consistent with the audience-engagement mechanism discussed in the review: Berger and Milkman (Berger & Milkman, 2012) claim that high-arousal emotions increase social transmission while low-arousal affect reduces it, suggesting that "activation" is a robust dimension of engagement even when valence is controlled. Although their data are journalistic rather than fictional, the present results support the broader claim that emotionally activating language can correlate with measurable forms of audience attention and diffusion.

Second, the negative association with concreteness is also compatible with patterns reported in prior computational work. As previously highlighted, da Silva et al. (2024) identify lexical tendencies separating more successful from less successful books and note that less successful works contain more concrete words, with frequent references to body parts, while more successful texts show broader lexical variety. This result provides an external point of convergence for interpreting the present finding: higher aggregate concreteness in the Statistical Analysis corpus may reflect a stylistic profile closer to the lexical tendencies observed in "less successful" texts. Taken together, these results could indicate that, in this corpus, an "overly concrete" register may correlate with lower download intensity. These findings are consistent with the results obtained in Chapter 4.

In contrast, pronoun density did not show a stable independent relationship with download intensity in the multivariate specification. In Chapter 1, pronoun-based perspective was treated as a theoretically motivated dimension, since Holmesian narration is often described as being mediated through Watson: a narratorial configuration that structures suspense via limited focalization, regulates reader alignment through stance, and supports the "fair-play" management of information (Emmott & Alexander, 2024; Jann, 1990; Krasner, 1997). Both first- and third-person pronoun measures yielded IRRs close to null value and wide confidence intervals crossing 1, suggesting that once arousal, concreteness, and

subordination are controlled for, variation in pronoun frequency alone does not account for systematic differences in downloads/day. A plausible explanation is that pronoun density is a more fine-grained marker of narratorial stance than broader affective or structural dimensions, and its contribution may be conditional rather than additive at the document level. In other words, pronouns may carry predictive value primarily in combination with other cues, or when measured more selectively (for instance, distinguishing dialogue from narration). Accordingly, the absence of a clear pronoun effect here should be interpreted as "no detectable independent signal in this multivariate specification," rather than as evidence that perspective is irrelevant to detective fiction engagement.

Whereas arousal and concreteness align with previous findings discussed in Chapter 1 and Chapter 4, the direction of the subordination effect diverges from earlier work. The negative result for subordination density diverges from a prominent precedent. Ashok et al. (Ashok et al., 2013) report that more successful novels can be less readable on standard metrics and interpret this as evidence that syntactic/stylistic complexity may correlate positively with success in fiction, potentially because it supports more complex plots or characterization.

In the Statistical Analysis corpus, however, greater subordination is associated with lower download intensity. A plausible interpretation is that this relationship is genre- and medium-sensitive. In fact, while classical detective fiction is often framed as privileging clarity, pacing, and descriptive economy, Ashok, et al. (Ashok et al., 2013) drew on multi-genre novels from Project Gutenberg and operationalize success primarily via download-count thresholds within genres, so their readability-related findings reflect a broad setting rather than genre-specific detective fiction. As previously stated, Van Dine explicitly cautioned against overly long descriptive passages and narrative digressions that slow down the story or distract from the puzzle structure. This also aligns with the Holmes–Watson narrative frame discussed in Chapter 2, where the procedural pacing of clue presentation favors accessible packaging of information.

Read in this light, heavier clausal embedding may impede processing fluency and dilute the forward drive of puzzle plotting, thereby reducing engagement. More speculatively, syntactic and lexical complexity may be less optimal in detective narratives, where part of the reader's satisfaction depends on being able to track the same evidential cues available to the detective; if information is packaged in overly embedded structures, it may become harder for readers to retain and integrate clues, weakening the payoff when the solution is revealed. This could help explain why subordination behaves differently here than in broader mixed-genre corpora where complexity may correlate with other dimensions of literary success.

In this respect, the negative association of heavy embedding is also compatible with the critical characterization of Holmes narration as procedurally oriented and readability-aware, where clue tracking benefits from relatively transparent packaging of information (see Chapter 2).

5.6 Conclusion

Using negative binomial regression with an exposure offset to model download intensity, this chapter has tested whether five document-level stylistic variables are associated with engagement. Results indicate that arousal is positively associated with download rate, while concreteness and subordination density are negatively associated with it; pronoun density measures do not show stable independent effects. Robustness checks suggest that the core pattern is not driven solely by a small number of extreme high-download cases. Taken together, these findings support the broader claim that a compact set of interpretable, text-derived features contains measurable signal related to the operationalized notion of "success."

At the same time, because the analysis is observational and the outcome is derived from Project Gutenberg downloads, this proxy should be interpreted cautiously: download counts may also reflect extra-textual influences that are not captured by the present covariates, even when time-since-release is controlled via the exposure offset.

Crucially, these results also provide a principled foundation for the next chapter's NLP-oriented work. Building on the inferential evidence obtained here, the following chapter shifts from association estimation to out-of-sample predictive evaluation, assessing whether models trained on the same features can classify individual books as more or less successful, and how predictive performance varies across modelling choices.

Chapter 6: Training NLP models to predict success

6.1 Purpose and role in the pipeline

The purpose of this part of the analysis is to evaluate whether success labels for early twentieth-century detective fiction can be predicted using only document-level, text-derived stylistic features. This predictive step is explicitly designed to avoid collapsing “success” into author identity: since a highly visible canonical author such as Arthur Conan Doyle is present in the corpus and appears in the upper tail of downloads, the Predictive Modelling Corpus applies author-balancing constraints (see Chapter 3) so that models cannot achieve good performance by learning a single author signature. The goal is therefore to test whether the five interpretable stylistic dimensions capture a signal that generalizes across authors within the same genre and period.

The feature extraction process produced a single JSON file containing, for each book, a structured collection of numeric indicators and associated metadata. Because standard machine learning models require rectangular numeric input, this JSON representation was converted into numeric feature matrices suitable for model training and evaluation. Although many features are extracted (Chapter 3), the predictive models in this chapter are trained only on the five theory-driven variables defined in Chapter 1; the conversion procedure is therefore specified with respect to this fixed input space.

The JSON output was converted into numeric feature matrices through a deterministic matrix construction procedure described in Section 6.4.

Three models were tested for this analysis: Logistic Regression, Random Forest, and a Neural Network. Model performance was evaluated using two threshold-independent metrics derived from the predicted probabilities of the positive (“successful”) class: ROC AUC and PR AUC. ROC AUC (area under the Receiver Operating Characteristic curve) quantified the model’s ability to rank a randomly chosen successful book above a randomly chosen unsuccessful one; it summarized the trade-off between true positive rate and false positive rate across all possible decision thresholds, with 0.5 corresponding to chance-level ranking and 1.0 to perfect separation. PR AUC (area under the Precision–Recall curve, also known as average precision in scikit-learn) summarized the trade-off between precision and recall across thresholds and was emphasized because the positive class had few members; unlike accuracy, PR AUC directly reflected how concentrated true successes were among the highest-scoring predictions. Under a random ranking (chance performance), the

expected PR AUC equaled the positive-class prevalence in the evaluation set, that is, the fraction of books labeled as successful. In this case, since 10% of the books were successful, a random model would have PR AUC ≈ 0.10; PR AUC values above this level therefore would indicate that the model ranked successful books disproportionately near the top of its predictions.

In addition to these threshold-free measures, performance was also summarized using the F1 score, a threshold-dependent metric computed after converting predicted probabilities into binary decisions at a chosen operating point (Sokolova & Lapalme, 2009). F1 is defined as the harmonic mean of precision and recall, $F1 = 2\frac{PR}{P+R}$, and therefore rewards models that achieve a balance between identifying many true successes (high recall) and keeping predicted successes relatively pure (high precision). Because F1 depends on a specific decision threshold, it complements ROC AUC and PR AUC by characterizing model behavior at an actionable classification point, which is especially relevant under class imbalance.

This stage begins with the per-book feature JSON produced by the extraction pipeline and ends with a model-ready representation consisting of a numeric matrix X and a binary label vector y. This representation is the sole input to the classifiers described in Section 6.6 and therefore defines what information the models are permitted to use. Sections 6.2-6.5 document the conversion procedure because choices such as feature-space definition, ordering, and missing-value policy directly affect comparability across models and the reliability of the evaluation reported in Section 6.7.

6.2 Design requirements for the numeric representation

Before any classifier can be trained, the nested JSON output must be transformed into a single, deterministic feature space shared by all books. This means converting the extracted measures into a two-dimensional feature matrix (a table with one row per text and one column per feature), with a fixed column order and identical definitions across the Predictive Modelling corpus. This step is not only a technical requirement for standard machine-learning workflows (which expect matrix-like numeric input), but also a methodological safeguard: it prevents feature misalignment across texts, reduces leakage risks from identifiers, and ensures that all reported results can be reproduced from the same feature specification.

To ensure validity and stability in the context of model training, the conversion process should satisfy a set of methodological constraints concerning the structure and interpretation of the resulting data. First, all samples had to be represented by feature vectors of identical dimensionality. This requirement implies that a single, global definition of the feature space had to be established in advance, specifying both which features are included and the exact order in which they appear. Without this constraint, it would be impossible to represent the corpus as a single numeric matrix suitable for statistical modelling.

Closely related to this requirement was the necessity of consistent feature ordering across all samples. Each column of the resulting matrix had to correspond to the same linguistic or psycholinguistic metric for every document. This is particularly important given the nature of the JSON format, in which data are stored as key-value mappings rather than ordered structures. Although some JSON parsers preserve insertion order, relying on this behavior can lead to subtle inconsistencies. Therefore, feature ordering had to be imposed explicitly and deterministically during the conversion process.

Finally, the feature representation had to be strictly numeric, since statistical learning models operate in a numerical vector space and cannot directly process symbolic or categorical information. For this reason, all non-numeric fields present in the original JSON structure, most notably document identifiers such as the book field, were excluded from the feature matrix. Any value that could not be interpreted as a valid finite number was discarded rather than propagated into the model input, thereby preventing invalid or undefined numerical behavior during training.

6.3 Feature flattening process

The feature extraction output was stored as nested JSON objects, which were convenient for inspection but not directly compatible with statistical learning models requiring tabular inputs. To bridge this gap, each book's nested feature structure was transformed into a flat set of numeric variables. The flattening step therefore produced a single-level mapping from feature names to numeric values for each book, which could then be assembled into a rectangular matrix.

The exact flattening and numeric validation logic used in the experiments is specified in Section 6.4.2.

6.4 Conversion from JSON to model-ready matrices

This stage implements the methodological bridge between the feature extraction pipeline (Chapter 3) and the supervised models (Section 6.6). Its purpose is to produce a

deterministic mapping from the extracted JSON records to a numeric design matrix X and a binary label vector y, such that all evaluated models differ only in learning algorithm, not in data representation or alignment.

6.4.1 Inputs, outputs, and alignment assumptions

The input to this conversion is the JSON produced by feature extraction, restricted to the Predictive Modelling corpus configuration described in Section 3.7 and introduced in Chapter 3. In the modelling configuration used here, the resulting dataset contains 246 book records, and each record is paired with a binary success label (successful vs. other) defined using the corpus-internal download threshold.

The procedure produces three aligned objects:

1. X: a numeric feature matrix (one row per book),
2. y: the aligned binary label vector,
3. feature_names: an ordered list defining the column semantics of X.

Two alignment constraints are enforced throughout: (a) row order is consistent across X, y, and identifiers (so the *i*-th row always corresponds to the same book), and (b) column order is fixed and deterministic so that each column denotes the same linguistic metric across all cross-validation folds and the final held-out evaluation.

6.4.2 Flattening and numeric validation

Each book-level record contains nested JSON objects grouping related features (Section 3.9). These are flattened into a single-level mapping from feature name → numeric value by recursively traversing the structure and concatenating keys into dot-delimited paths (e.g., *Phrase.subordinate_density*). This preserves the interpretability of the original grouping while producing a uniform tabular representation.

Only finite numeric values are retained as candidate inputs. Non-numeric fields, most importantly the book identifier and any metadata, are treated strictly as identifiers for traceability and are never passed to the model, preventing accidental leakage from document IDs into prediction.

6.4.3 Feature set definition for the predictive models

Although the extraction pipeline produces a broad set of stylistic variables (Chapter 3), the predictive experiments in this chapter are intentionally based on the restricted, theory-driven

feature set introduced in Chapter 1. Concretely, the model input space is defined as the following five document-level variables: concreteness, arousal, first-person pronoun density, third-person pronoun density, and subordination density.

Defining the feature space in this explicit way ensures that the conversion step directly implements the methodological claim of this chapter: the models are evaluated on whether these five interpretable stylistic and psycholinguistic dimensions carry predictive signal for the operationalized success label, independent of topic modelling or high-dimensional lexical representations.

6.4.4 Missing values and imputation policy

Since some documents may lack a given extracted key, missing values must be handled explicitly to maintain a rectangular input matrix. In this study, any missing value among the five selected features is imputed as 0.0 during matrix construction. This choice was suited to the nature of the extracted metrics, which primarily represent normalized frequencies or densities for which a value of zero has a clear semantic interpretation. Moreover, this strategy guaranteed that all feature vectors maintain the same dimensionality, thereby preserving the structural integrity of the numeric matrix used for classification.

Because missing values can in principle reflect extraction failure rather than true absence, zero-imputation was interpreted as valid primarily for frequency- or density-based variables; if systematic extraction failures were present, they could bias estimates, and this risk motivated the exclusion of features found to behave as preprocessing proxies.

This policy is applied uniformly across all models, ensuring that performance differences reported in Section 6.7 are attributable to model behavior rather than inconsistent handling of missingness.

6.4.5 Matrix construction and traceability

Given the fixed ordered feature list, the design matrix X is constructed by iterating through books in corpus order and retrieving each of the five feature values in the predefined column order. The label vector y is constructed in the same order using the success labels defined in the corpus metadata and operationalized as described in Chapter 3.

For auditing and interpretive follow-up, book identifiers are retained in a parallel aligned list (or optional CSV) but kept strictly separate from X. This preserves reproducibility and error

analysis capability while ensuring that the supervised models in Section 6.6 operate only on the intended stylistic predictors.

6.5 Output artifacts and reproducibility

The conversion procedure conceptually yields three artifacts: X, y, and the ordered feature_names list that defines the meaning of each column. In the experiments reported in this thesis, these objects are constructed and used within the experimental scripts (rather than stored as separate files), but the representation is fully deterministic given the same JSON input and the same five-feature definition.

Separating numeric inputs (X) from metadata (feature names and identifiers) ensures stable column semantics across repeated runs and cross-validation folds, which is necessary for a fair comparison of the probabilistic classifiers evaluated in Section 6.6 and summarized in Section 6.7.

6.6 Models

To assess whether the proposed linguistic features can distinguish successful detective fiction books from other works, three probabilistic classifiers were trained and compared under an identical feature representation and an identical evaluation protocol. All models output a probability for the positive class (“successful”), which allows both threshold-free evaluation (ROC-AUC and PR-AUC) and threshold-based classification at a fixed operating point. Class imbalance was addressed in every model by adjusting the weight values, so that the minority class receives higher effective weight during training.

Logistic regression was implemented in scikit-learn as a probabilistic linear classifier trained to predict the odds of the positive class as a weighted sum of the five features (Hosmer Jr et al., 2013). The model was fit within a pipeline that standardizes inputs using StandardScaler followed by LogisticRegression with class_weight="balanced" to address the strong class imbalance. A liblinear solver with L2 regularization was employed for penalties (scikit-learn default penalty), max_iter=5000, and otherwise default settings (C=1.0, fit_intercept=True, tol=1e-4).

The second model is a random forest classifier (Breiman, 2001) intended to capture feature interactions while remaining relatively robust in such low-dimensional settings. The forest is trained with n_estimators=1000 trees, max_depth=5, min_samples_leaf=5, and max_features="sqrt". Class imbalance is handled via class_weight="balanced_subsample", which reweights classes within each bootstrap sample. Parallelization is enabled via

n_jobs=-1. To ensure that variability estimates in repeated cross-validation reflect both data partitioning and model stochasticity, the forest's random_state is varied across repeated runs; for the final held-out evaluation, the model is refit on the full development partition with a fixed seed consistent with the held-out split seed.

Scikit-learn defaults are retained for the random forest, including bootstrap aggregation (bootstrap=True), the Gini impurity criterion (criterion="gini"), minimum samples to split an internal node (min_samples_split=2), and no pruning constraints beyond the specified max_depth and min_samples_leaf. The predicted probability corresponds to the mean of the per-tree class probabilities across the ensemble.

The third model is a compact feed-forward neural network trained on the same five standardized inputs. The architecture is a multilayer perceptron with a single hidden layer of four units, followed by dropout regularization, and a sigmoid output unit. Specifically, the network maps the five-dimensional input into a dense layer with HIDDEN_UNITS=4 and ReLU activation, applies dropout with rate DROPOUT=0.2, and produces a scalar probability through a single-unit sigmoid layer. The model is trained with the Adam optimizer and binary cross-entropy loss. Because the dataset is small and imbalanced, class weights are computed on each training split using a balanced inverse-frequency rule and passed to the training procedure. During development cross-validation, early stopping is employed to reduce overfitting: training proceeds for up to 200 epochs with mini-batches of 16 and early stopping monitors validation PR-AUC with patience 15, restoring the best weights observed on the validation partition. For the final held-out evaluation, the network is retrained on the entire development partition using the same optimization settings and a fixed number of epochs.

Early stopping is used only inside the development cross-validation loop, where an explicit inner validation split is available for model selection; in that setting, the callback monitors validation PR-AUC and restores the best observed weights. For the final held-out evaluation, the network is retrained once on the full development partition (after standardization fitted on DEV) and run for a fixed budget of 200 epochs with the same optimizer and batch size; no held-out data are used for monitoring or stopping.

For reproducibility, random seeds are controlled at each run by setting the Python hash seed and the pseudo-random generators of Python, NumPy, and TensorFlow.

6.6.1 Class imbalance and metrics

The dataset is split once into a development partition (80%) and a held-out test partition (20%) using stratified sampling with a fixed random seed. All model selection is performed exclusively on the development partition via repeated stratified cross-validation with 5 folds and 10 repeats, repeated across five independent seeds (1–5), yielding 250 outer test evaluations. Within each outer training fold, an additional stratified split reserves 20% of the fold for validation, which is used solely to select a decision threshold on a fixed grid of 201 evenly spaced values in [0,1]. The selected threshold maximizes validation F1, and a single global threshold per model is then fixed as the median of the 250 selected thresholds; this global value is used for the final single evaluation on the held-out test set. The threshold that maximizes validation F1 is selected; ties are resolved by preferring higher precision and, if still tied, the lower threshold to ensure determinism.

Because the successful class represents approximately 10% of the Predictive Modelling corpus, metrics that depend heavily on the majority class can be misleading (He & Garcia, 2009). For this reason, performance is summarized using both:

Threshold-free metrics

- ROC-AUC, which evaluates ranking quality across all possible thresholds.
- PR-AUC, which is often more informative under strong class imbalance because it emphasizes the purity of predicted positives (Davis & Goadrich, 2006).

Threshold-based metrics (at a fixed operating point)

- Precision, Recall, and F1 for the positive class
- Accuracy
- Balanced Accuracy, which was included to reduce the impact of class imbalance on interpretation.

For comparison, a baseline classifier that always predicts “other” would achieve an accuracy close to the negative-class prevalence (≈ 0.90) while having Recall = 0 and F1 = 0 for the successful class. Balanced accuracy for such a model would be 0.50. Similarly, the expected PR-AUC of a random ranking is approximately equal to the positive prevalence (≈ 0.10). These baseline expectations motivate reporting PR-AUC and balanced accuracy in addition to accuracy.

6.6.2 Protocol and threshold selection

All model selection and robustness estimates are computed on the development partition (DEV) using repeated stratified cross-validation (Kohavi, 1995). Specifically, DEV is evaluated with stratified 5-fold cross-validation (CV_SPLITS = 5), repeated ten times per shuffle (CV_REPEATS = 10) across five independent random seeds (SEEDS = [1, 2, 3, 4, 5]), yielding 250 outer test-fold evaluations overall. Within each outer training fold, threshold selection is nested via an additional stratified split: the outer training data are divided into inner-train and inner-validation subsets with an 80/20 split (VAL_SIZE = 0.20), using a deterministic per-run seed derived from the outer seed and fold index. Models are trained exclusively on inner-train; predicted probabilities on inner-validation are used to select an operating threshold.

Thresholds are selected using a deterministic policy that maximizes the positive-class F1 score. Candidate thresholds are evaluated on a uniform grid of 201 values in [0, 1]. The threshold achieving the highest validation F1 is chosen; ties are broken by preferring higher precision, and then by selecting the lower. The selected threshold is then applied to the corresponding outer test fold to compute threshold-dependent metrics (precision, recall, F1, accuracy, balanced accuracy). In parallel, threshold-free metrics (ROC-AUC and PR-AUC) are computed directly from the probability outputs on the same outer test fold.

To obtain a single operating point per model for the final evaluation, the thresholds selected across the 250 DEV runs are aggregated and the global threshold is fixed as their median. The final thresholds are 0.6825 for Logistic Regression, 0.4250 for Random Forest, and 0.5750 for the Neural Network. Each model is then retrained on the full DEV split and evaluated exactly once on the held-out test partition using the corresponding fixed threshold, ensuring that neither model configuration nor threshold definition is influenced by held-out outcomes.

Scikit-learn was used for Logistic Regression and Random Forest, and TensorFlow/Keras for the Neural Network.

6.7 Results

6.7.1 Cross evaluation results

Table 7 reports mean ± standard deviation over all repeated DEV-CV runs. Standard deviations are substantial for positive-class metrics, reflecting both scarcity of positives and the sensitivity of F1 to threshold choice in imbalanced settings.

Table 7: Development-set performance (mean ± std) with thresholds chosen on inner validation

Model	Prec	Recall	F1	Accu-racy	Bal. Acc	PR-AUC	ROC-AUC	Chosen-thresh-old
Logistic Regress	0.3179± 0.1745	0.6037± 0.3061	0.3953± 0.1883	0.8299± 0.0782	0.7289± 0.1385	0.4182± 0.1782	0.8008± 0.1224	0.6825
Random Forest	0.3302± 0.1795	0.5630± 0.3026	0.3882± 0.1797	0.8410± 0.0675	0.7168± 0.1316	0.4459± 0.1765	0.7852± 0.1375	0.4250
Neural Network	0.2279± 0.1984	0.4660± 0.3410	0.2791± 0.2048	0.7771± 0.1590	0.6382± 0.1484	0.3513± 0.1932	0.7309± 0.1705	0.5750

Across repeated stratified cross-validation on the development partition, logistic regression and random forest exhibit closely comparable performance, with differences that are small relative to the observed cross-run variability. Logistic regression attains a slightly higher mean F1 (0.3953 ± 0.1883) and balanced accuracy (0.7289 ± 0.1385), whereas random forest yields a marginally higher mean PR-AUC (0.4459 ± 0.1765). Both models achieve ROC-AUC values substantially above chance (0.8008 ± 0.1224 for logistic regression; 0.7852 ± 0.1375 for random forest), indicating that the five-feature representation carries discriminative signal beyond the strong class imbalance (random PR-AUC ≈ 0.0976). The neural network underperforms on average (mean F1 0.2791 ± 0.2048; balanced accuracy 0.6382 ± 0.1484) and shows pronounced instability, a pattern consistent with flexible models trained on limited, low-dimensional inputs under heavy imbalance.

The contrast between threshold-free metrics (ROC-AUC, PR-AUC) and threshold-dependent metrics (F1, balanced accuracy) is typical here: AUC-based measures evaluate the ordering of scores across all possible thresholds, whereas F1 evaluates performance at a single selected cutoff and is therefore more sensitive to calibration and small score shifts. Accordingly, the selected thresholds (values shown in Table 7) should be interpreted as model-specific operating points determined by their score distributions, rather than as directly comparable cutoffs.

6.7.2 Held-out test results

Table 8: Held-out test results with chosen threshold

Model	Threshold	Matrix	Precision	Recall	F1 (pos)	Acc.	Bal. Acc	PR-AUC	ROC-AUC
Logistic Regr.	0.6825	[[41, 4], [0, 5]]	0.5556	1.0000	0.7143	0.9200	0.9556	0.7893	0.9689
Random Forest	0.4250	[[44, 1], [3, 2]]	0.6667	0.4000	0.5000	0.9200	0.6889	0.6824	0.8933
Neural Network	0.5750	[[42, 3], [2, 3]]	0.5000	0.6000	0.5455	0.9000	0.7667	0.7112	0.9511

For the held-out evaluation, each model was retrained on the full development partition and tested once on the held-out split using a threshold chosen exclusively within DEV. On this split, logistic regression achieves the strongest threshold-dependent outcome (Precision = 0.556, Recall = 1.000, F1 = 0.714; balanced accuracy = 0.956), corresponding to no false negatives at the selected operating point (TN = 41, FP = 4, FN = 0, TP = 5). Random forest attains the same overall accuracy (0.920) but adopts a more conservative operating profile (Precision = 0.667, Recall = 0.400), while the neural network yields intermediate performance (Precision = 0.500, Recall = 0.600).

Threshold-free metrics are also high on this split (e.g., ROC-AUC = 0.969 for logistic regression), consistent with the DEV-CV finding that the models capture a ranking signal. However, because the held-out set contains only five positive instances, these single-split estimates are inherently high-variance: changes in one or two cases can substantially shift recall, F1, and AUC. For that reason, the held-out results are best read as a confirmatory check under a fully untouched split, while the repeated DEV-CV distributions remain the primary basis for model comparison and inference.

Overall, logistic regression emerges as the most robust choice under the present experimental conditions: it matches or exceeds random forest on DEV-CV in balanced accuracy and F1, and it also attains the best held-out threshold-dependent outcome on the selected split. Random forest offers a competitive alternative with similar DEV-CV performance and, depending on the desired operating objective, may be preferable when higher precision and lower recall is acceptable. The neural network does not appear

advantageous in this setting, likely due to the limited dataset size and the small number of informative features, which restricts the benefits of additional model capacity.

6.8 Discussion

This study examined whether a compact set of linguistically interpretable document-level variables contains a predictive signal for a corpus-defined proxy of success in early twentieth-century detective fiction. The analysis is intentionally conservative in its representational assumptions: rather than attempting to model topic or plot content, it focuses on stylistic and psycholinguistic features that plausibly reflect narrative stance, syntactic packaging, and lexical choice. Within this framing, the results support the conclusion that these features carry measurable information about the operational success label, while also illustrating the limits imposed by small sample size and strong class imbalance.

Across repeated cross-validation on the development partition, both logistic regression and random forest achieve discrimination well above chance, as reflected by ROC-AUC values around 0.8 and PR-AUC substantially above the prevalence baseline described in 6.6.1. Importantly, these threshold-free results indicate that the models are not merely exploiting the majority class but are learning a ranking that tends to prioritize successful novels. At the same time, threshold-dependent metrics show considerable variability across runs, which is expected when the minority class is rare and F1 is evaluated at a single operating point chosen from a finite validation set. In practical terms, this means that performance should be interpreted as a distribution rather than a single number, and that conclusions about small differences between model families should be tempered by the observed cross-run dispersion.

The comparison between logistic regression and random forest suggests that the structure of this five-feature representation may include mild non-linearities. Random forest yields slightly better PR-AUC on average in DEV-CV, whereas logistic regression achieves marginally higher mean F1 and balanced accuracy and remains competitive in overall discrimination. Importantly, logistic regression is also the most transparent model in this study: its decision rule is a single linear boundary, and model behavior can be summarized by a small set of coefficients whose sign and magnitude directly indicate how each standardized feature shifts the predicted log-odds of success (and can be expressed as odds ratios). This pattern is consistent with a setting in which much of the usable signal can be captured by relatively simple decision boundaries, with limited benefit from higher-

capacity function approximation. From an interpretability standpoint, the linear model also offers a more direct route for linking predictive behavior to hypotheses about narratorial stance and affective/lexico-syntactic cues, whereas the forest primarily provides predictive flexibility without comparably simple global parameters.

The neural network does not provide consistent gains and is substantially less stable. This outcome is unsurprising given the combination of a low-dimensional input space and a small number of positive examples. Even with class weighting and early stopping, a flexible model can overfit, yielding large swings in precision and recall across runs. Under richer representations (e.g., sentence embeddings, discourse-level trajectories, or lexical distributional features) neural architectures may become more competitive, but with only five aggregate features the model capacity is not the limiting factor; rather, the bottleneck is the information content and the scarcity of positive instances available for robust estimation.

On the held-out test set, logistic regression achieves the strongest threshold-dependent results under the fixed operating point derived from DEV, while random forest exhibits a more conservative profile with fewer false positives but more false negatives. Although these outcomes are informative in illustrating how different model families realize different precision–recall trade-offs, they should not be over-weighted as decisive evidence. The held-out partition contains only five positives, therefore small changes in one or two instances can produce large shifts in recall, F1, and even AUC values. For these reasons, the held-out evaluation is best understood as a confirmatory check that performance remains above chance under a strictly untouched split, while DEV-CV remains the primary basis for comparative model assessment. More generally, the sensitivity of held-out outcomes to the stratified split reinforces the importance of reporting repeated cross-validation distributions in imbalanced literary datasets of this size.

Point estimates on this split are expected to exhibit substantial variance; therefore, conclusions rely primarily on the repeated DEV cross-validation distributions, using the held-out evaluation as an additional sanity check rather than as the sole basis for model ranking. Still, the results are compatible with the hypothesis that narrative stance (as proxied by pronoun usage), syntactic embedding (subordination density), and lexical-semantic properties (concreteness and arousal) are weak but detectable correlates of the operationalized success proxy. This should not be interpreted as a claim that “success” is primarily determined by linguistic style; rather, it indicates that stylistic decisions measurable at the document level vary with the download-based label within the Predictive Modelling corpus, potentially reflecting reader accessibility, narrative immediacy, or affective

engagement. Because Project Gutenberg downloads are influenced by many non-textual factors—availability, author notoriety, metadata quality, and modern rediscovery—any causal interpretation would be unwarranted. The most plausible conclusion is associative: these features encode a modest signal that can be exploited by standard classifiers in a strictly evaluated setting.

Future work should aim to reduce variance and clarify what aspects of language drive the observed signal. Two directions are particularly important: increasing the number of positive instances by enlarging the corpus or by considering alternative operationalizations of success, and strengthening evaluation against confounds (for example, using author-stratified splits to limit author identity leakage, or controlling for publication date and series effects). In parallel, expanding the feature set while preserving interpretability, such as adding dialogue proportion, sentiment dynamics, or measures of narrative pacing, could test whether the present signal is primarily stylistic or whether it reflects broader discourse organization that correlates with reader engagement.

6.9 Conclusion

Using a strict no-leakage protocol with repeated stratified cross-validation on a development set and a single held-out confirmatory evaluation, this chapter has compared logistic regression, random forest, and a feed-forward neural network for predicting success in detective fiction from five interpretable document-level features: concreteness, arousal, first-person pronoun density, third-person pronoun density, and subordination density. Results indicate that the linear and tree-based baselines provide the most stable performance in this imbalanced, low-dimensional setting, while the neural model does not yield consistent improvements. Overall, findings suggest that these text-derived stylistic features encode measurable information associated with the operational success label, while also highlighting the need for larger datasets and/or richer representations for more robust prediction.

Chapter 7: General discussion and conclusions

7.1 Thesis summary

This thesis set out to investigate whether stylistic features that can be operationalized as measurable linguistic variables are systematically associated with literary "success" in late nineteenth- and early twentieth-century detective fiction, and whether these features can be used to train a model with the goal to distinguish "successful" books. Rather than treating literary value as reducible to a single indicator (i.e., a number), the study treated "success" as an explicitly pragmatic, corpus-internal category intended to capture present day public-domain attention. "Success" was operationalized using Project Gutenberg download counts, with the top 10% of the corpus by downloads classified as successful, and the remaining works treated as a comparison set. The overarching goal was associative and comparative: to test whether a small, interpretable set of stylistic signals aligns with engagement as approximated by downloads, under controlled genre and period constraints, and to examine whether any such signals generalize across authors.

The empirical work combined three components that moved from corpus interpretation to statistical testing and then to modelling. Initially, a corpus-analysis stage focused on the stylistic profile of "success", with particular attention to concreteness through motivated lexical domains connected to genre theory and to historical prescriptions about detective. Second, a statistical stage tested whether five theory-driven document-level variables were associated with download intensity, while explicitly controlling exposure time. Finally, a supervised modelling stage evaluated whether those same five variables contained sufficient signal to classify individual texts as successful versus other under a conservative evaluation protocol designed to limit trivial author-signature learning.

Across these stages, the results converge on a coherent picture of stylistic correlations of download-based success within this detective-fiction corpus.

7.2 The stylistic profile of successful texts

Chapters 4-6 develop a single interpretive thread that starts from Doyle-era detective narration and then tests it quantitatively at the document level. In the corpus comparison (Chapter 4), the "Success" subset is characterized less by the presence/absence of investigation-related language than by distributional emphasis: in the perception domain (Section 4.3.1), "Success" favors items such as *notice* and *observe* while "Other" shows relatively higher rates for more sensory-oriented verbs such as *feel* (and related items); in

the cognition domain (Section 4.3.2), both subcorpora contain cognition language, but with differences in epistemic weighting; and in the forensic/institutional domain (Section 4.3.3), “Success” shows a stronger procedural/institutional framing. Romance (Section 4.3.4) further nuances this picture: romance is not absent in “Success”, but it is less overtly foregrounded and tends to appear through more abstract or selective marking. Statistical modelling then tested whether these corpus-level tendencies aligned with document-level variables: the negative binomial regression (Chapter 5) shows a robust negative association for concreteness and subordination density, and a positive association for arousal, while pronoun densities do not retain stable independent effects when modelled jointly. Finally, the modelling evaluation (Chapter 6) indicates that these five features contain a modest but detectable signal for the operationalized “Success” label under repeated cross-validation, supporting RQ2 without implying a one-to-one mapping between lexical domains and aggregate features.

The corpus analysis provided an interpretable qualitative-quantitative bridge between genre expectations and the modelling results. In the perception domain, the successful subset did not simply show “more perception” in a raw sensory sense; rather, it showed relatively higher rates of verbs that construe perception as selective attention and interpretation, consistent with detection as the transformation of observation into evidential reasoning. In the cognition domain, the successful subset showed relatively higher rates of verbs encoding epistemic control and investigative stance, whereas the comparison corpus relied relatively more on conjectural or open-ended stance markers. In the forensic/procedural domain, successful texts more strongly foregrounded institutional roles and evidential procedure, and this effect was not reducible to general action intensity, since more generic action verb counts were not uniformly higher in the successful subset. Finally, romance language showed a nuanced pattern: successful texts were lower on frequent domestic and relational markers, yet relatively higher on terms that frame romance as a thematic register, suggesting that when romance appears in detective fiction it may function more as motive, atmosphere, or marked concept than as a sustained enacted domestic subplot. Taken together, the lexical-domain results characterize the “successful” subset as more procedurally anchored and more explicitly organized around evidential attention and controlled inference, with a more limited role of everyday romance enactment.

The statistical modelling stage sharpened this picture by testing whether the five core variables retained systematic associations with downloads when modelled jointly. Using negative binomial regression with an exposure offset to account for time-available effects,

three predictors showed robust associations with download intensity. Arousal was positively associated with download rate: books with more emotionally activating language tended, on average, to have higher expected downloads per day. Concreteness was negatively associated with download rate, suggesting that texts with higher aggregate concreteness scores tended to have lower download intensity. This direction is consistent with the corpus-level domain patterns discussed in Chapter 4 insofar as the “Success” subset tends to foreground more interpretive and institutionally framed investigative language rather than primarily sensory depiction, an orientation that aligns with a more abstract lexical profile. Subordination density was also negatively associated with download intensity, indicating that heavier clausal embedding correlated with lower engagement as measured here. In contrast, first- and third-person pronoun densities did not show stable independent effects in the multivariate model: their confidence intervals overlapped the null, and their p-values did not indicate reliable associations once the other predictors were controlled. Robustness checks preserved the same core pattern: positive arousal, negative concreteness, and negative subordination remained stable, reinforcing the view that the results were not driven solely by a few extreme high-download titles.

Interpreting these findings against the theoretical background proposed in Chapter 1 yields several implications. The arousal effect aligns with the audience-engagement argument that emotionally activating language supports diffusion and attention. The negative concreteness association resonates with prior computational work (as mentioned in Chapter 1) suggesting that less successful texts can show more concrete lexical tendencies, implying that within this genre-and-period slice, a strongly concrete register may correlate with lower present-day attention. The negative subordination association diverges from broader mixed-genre claims that complexity can correlate with success. However, it becomes plausible when interpreted as genre-sensitive: detective fiction’s puzzle structure and “fair-play” clue tracking may benefit from relatively transparent information packaging, and heavy embedding may reduce processing fluency or obscure evidential cues, weakening the reader’s ability to integrate clues and experience payoff.

7.3 Supervised modelling results

The predictive modelling stage asked whether the five-feature representation (concreteness, arousal, first-person pronouns, third-person pronouns, subordination) contains a generalizable signal for success classification under a conservative protocol. The modelling setup emphasized interpretability and avoided richer content features on purpose: the intent

was not to predict success by topic, but to test whether stylistic and psycholinguistic dimensions encode discriminative information. Evaluation also emphasized metrics appropriate to strong class imbalance (ROC-AUC and PR-AUC, alongside threshold-dependent measures such as F1 and balanced accuracy) and relied primarily on repeated stratified cross-validation distributions on the development partition, using a single held-out split only as a confirmatory check.

Across repeated development-set evaluations, logistic regression and random forest performed comparably and substantially above chance in ranking terms (ROC-AUC around 0.8, with PR-AUC well above the prevalence baseline). Threshold-dependent metrics showed substantial variability across runs, which is expected given the rarity of positives and the sensitivity of F1 to operating-point choice. The neural network underperformed and was less stable, consistent with the combination of low-dimensional inputs and a small number of positive examples: additional capacity did not translate into reliable gains under these constraints. On the held-out split, logistic regression achieved the strongest threshold-dependent outcome under the fixed threshold derived from the development procedure, while random forest adopted a more conservative precision–recall trade-off. Because the held-out set contained very few positives, these single-split numbers are inherently high-variance; the more important conclusion is that, across repeated evaluation, standard classifiers can exploit a modest but detectable signal in the five stylistic features.

The results presented above support a single, convergent takeaway: across complementary methods, a limited set of interpretable stylistic variables shows a stable, although modest, association with the operationalized notion of “success”. The next section steps back from the details of individual analyses to articulate what can be generalized from these findings at the level of research, design and interpretability. It summarizes the thesis contributions in terms of the framework, methodological choices, and transferable insights enabled by this multi-level approach.

7.4 Thesis contributions

The thesis contributes an integrated, interpretable framework for studying success-related variation in historical detective fiction that does not rely on opaque high-dimensional representations it shows, first, that the “successful” subset exhibits a coherent lexical profile tied to evidential attention, epistemic control, and procedural framing, features that are strongly compatible with genre theory and with Holmes scholarship on detection as observation plus inference. At the same time, the corpus-based domain analysis suggests

that romance is not simply absent in “Success” but tends to be less overtly foregrounded and more selectively lexicalized, with a relative shift toward more abstract romance-related concepts rather than concrete relational enactment.

Secondly, it demonstrates through statistical modelling that arousal, concreteness, and subordination density show robust correlation with download intensity when exposure time is controlled, while pronoun densities do not produce stable independent effects in the main multivariate specification. Third, it confirms through out-of-sample evaluation that these stylistic features contain measurable discriminative information: they support above-chance prediction across repeated cross-validation, even under class imbalance and with author-balancing constraints designed to reduce trivial identity learning.

7.5 Limitations and cautions

Several limitations constrain interpretation and point to future work. The most relevant being that the dataset is small, especially for supervised prediction, and the positive class is rare. The main statistical analyses are conducted on 281 texts, with “success” defined as the top 10% by downloads, spanning 28 titles. For supervised classification, additional author-balancing constraints (cap of ten texts per author, plus overlap constraints between classes) further reduced the Predictive Modelling corpus to 246 texts. Under the same top-decile definition, this implies only few successful instances in the entire predictive dataset, and when the data are split into DEV (80%) and held-out TEST (20%), the held-out evaluation contains only five positive cases. This scarcity creates strong class imbalance and makes threshold-dependent metrics (precision/recall/F1) high-variance: changing the classification of one or two successful books can materially shift the results, especially on the held-out split.

Regarding the corpus analysis reported in Chapter 4, it is important to note that it’s conducted over individual lemmas and some items occur at relatively low frequencies, several observed contrasts are small in absolute terms; accordingly, the domain-based patterns reported in this thesis should be interpreted as descriptive tendencies rather than as definitive evidence of large stylistic differences.

Furthermore, Project Gutenberg downloads are a proxy for digital engagement, not a direct measure of historical popularity, sales, or intrinsic literary merit. Downloads are shaped by extra-textual factors such as author notoriety, digitization and metadata quality, educational reuse, and present-day rediscovery dynamics; therefore, the statistical associations identified here must not be read causally. The corpus also inherits selection and survival

bias from Project Gutenberg: what is available and widely downloaded is already filtered by institutional and historical factors.

Another key limitation concerns the fact that the five-feature set is intentionally compact and document-level: it cannot capture discourse structure, plot dynamics, dialogue proportion, pacing, or emotional trajectories over narrative time, all of which may be important to reader experience. Finally, narratorial stance is likely more complex than global pronoun density can represent; distinguishing narration from dialogue, capturing focalization shifts, and modelling stance markers beyond pronouns may be necessary to detect perspective effects more reliably.

7.6 Future directions

Future work could expand both data and representation while preserving interpretability. Enlarging the corpus or adopting alternative operationalizations of success (or multi-dimensional success measures) would increase the number of positive instances and reduce variance. Evaluation protocols could be strengthened further to address confounds, for example by using stricter author-stratified splits, controlling for series effects, and incorporating publication-date covariates more directly. On the feature side, adding interpretable measures that reflect narrative organization—dialogue proportion, pacing proxies, discourse connectives, suspense markers, sentiment/arousal trajectories across narrative time, and more fine-grained stance indicators—could clarify whether the signal observed here is primarily lexical-affective, syntactic, or discourse-structural. In particular, adding the other features extracted in this study could provide further insight on the topic. More broadly, integrating stylistic signals with carefully chosen non-textual metadata could help separate linguistic correlates of engagement from repository and cultural-visibility effects.

7.7 Closing remarks

Within the constraints of a historical, public-domain corpus and a deliberately conservative feature design, this thesis provides converging evidence that stylistic properties of language, especially emotional activation, degrees of concreteness, and syntactic packaging are meaningfully associated with a download-based proxy of engagement in early twentieth-century detective fiction. The effect sizes and predictive performance indicate a modest signal rather than a deterministic rule, but the results demonstrate that even a small set of interpretable stylistic variables can illuminate how genre-specific narrative practices relate

to present-day attention, while also setting clear methodological boundaries for what such analyses can and cannot claim.

References

Abdi, H. (2007). Z-scores. In *Encyclopedia of Measurement and Statistics*, 3, 1055–1058.

Allan, J. M., & Pittard, C. (2019). *The Cambridge companion to Sherlock Holmes*. Cambridge University Press.

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Ashok, V. G., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1753–1764. https://aclanthology.org/D13-1181.pdf

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Bode, K. (2012). *Reading by numbers: Recalibrating the literary field.* Anthem Press.

Boghrati, R., Berger, J., & Packard, G. (2023). Style, content, and the success of ideas. *Journal of Consumer Psychology*, *33*(4), 688–700. https://doi.org/10.1002/jcpy.1346

Boucher, A., & Perkins, R. (2020). The case of Sherlock Holmes and linguistic analysis. *English Literature in Transition, 1880–1920*, *63*(1), 77–98. https://muse.jhu.edu/article/743930

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge University Press.

Cawelti, J. G. (2014). *Adventure, mystery, and romance*. University of Chicago Press.

Ciotti, F. (2021). Distant reading in literary studies: A methodology in quest of theory. *Testo e Senso*, (23), 195–213.

Citron, F. M., Gray, M. A., Critchley, H. D., Weekes, B. S., & Ferstl, E. C. (2014). Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, *56*, 79–89. https://doi.org/10.1016/j.neuropsychologia.2014.01.002

Conan Doyle Estate. (n.d.). *Physician*. Retrieved January 13, 2026, from https://arthurconandoyle.co.uk/physician

Consul, P. C., & Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, *21*(1), 89–109. https://doi.org/10.1080/03610929208830766

da Silva, G. D., Silva, F. N., de Arruda, H. F., e Souza, B. C., Costa, L. F., & Amancio, D. R. (2024). Using full-text content to characterize and identify best seller books: A study of early 20th-century literature. *PloS One,* *19*(4), e0302070. https://doi.org/10.1371/journal.pone.0302070

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 233–240. https://doi.org/10.1145/1143844.1143874

Doyle, A. C. (1892). *A Study in Scarlet: A Detective Story*. Ward, Lock, Bowden, and Company.

Doyle, A. C. (1993). *The case-book of Sherlock Holmes*. Wordsworth Editions.

Doyle, A. C. (2006). *The coming of the fairies*. University of Nebraska Press.

Emmott, C. & Alexander, M. (2024). "You see, but you do not observe": Sensory manipulation and sense-making in the Sherlock Holmes detective stories. In Pillière, L., Sorlin, S. (eds) *Style and Sense(s)*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-54884-0_6

Forsyth, R. S. (2000). Pops and flops: Some properties of famous English poems. *Empirical Studies of the Arts*, *18*(1), 49–67. https://doi.org/10.2190/E7Q8-6062-K6H4-XFRW

Hall, J. W. (2012). *Hit lit: Cracking the code of the twentieth century's biggest bestsellers.* Random House.

Harvey, J. (1953). The content characteristics of best-selling novels. *Public Opinion Quarterly*, *17*(1), 91–114. https://doi.org/10.1086/266441

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Jann, R. (1990). Sherlock Holmes codes the social body. *ELH*, *57*(3), 685–708.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kasyoki Muoka, A., Owino Ngesa, O., & Gichuhi Waititu, A. (2016). *Statistical models for count data*. https://ir.ttu.ac.ke/handle/123456789/45

Knight, S. T. (2003). *Crime fiction 1800-2000: Detection, death, diversity*. Palgrave Macmillan. https://orca.cardiff.ac.uk/id/eprint/3716

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1145). Morgan Kaufmann. https://dl.acm.org/doi/10.5555/1643031.1643047

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1318–1326. https://aclanthology.org/P11-1132.pdf

Krasner, J. (1997). Watson falls asleep: Narrative frustration and Sherlock Holmes. *English Literature in Transition, 1880-1920*, *40*(4), 424–436.

Leydesdorff, L., & Bornmann, L. (2011). *Percentile ranks and the integrated impact indicator (I3)* (arXiv:1112.6281). arXiv. https://doi.org/10.48550/arXiv.1112.6281

Liggins, E., & Vuohelainen, M. (2019). Introduction: Reassessing the Strand magazine, 1891–1918. *Victorian Periodicals Review*, *52*(2), 221–234.

Livingston, E. H. (2004). The mean and standard deviation: What does it all mean? *Journal of Surgical Research*, *119*(2), 117–123.

McGann, J. (1998). Textual scholarship, textual theory, and the uses of electronic tools: A brief report on current undertakings. *Victorian Studies*, *41*(4), 609–619.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174–184. https://aclanthology.org/P18-1017/

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *2*(1), 54–68.

O'Gorman, F. (2007). Conan Doyle, Sherlock Holmes, and the Victorian media. *Linguæ &- Rivista Di Lingue e Culture Moderne*, *5*(1), 53–60.

Project Gutenberg. (n.d.). *Project Gutenberg: Free eBooks*. Retrieved February 10, 2026, from https://www.gutenberg.org/

R Core Team. (2024). *R: A language and environment for statistical computing (Version 4.4.2.)*. R Foundation for Statistical Computing.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*(1), 31. https://doi.org/10.1140/epjds/s13688-016-0093-1

SAS Institute Inc. (2004). *Modelling rates and estimating rates and rate ratios in PROC GENMOD (SAS Knowledge Base Article 24188).* Retrieved February 10, 2026, from https://support.sas.com/kb/24/188.html

Scaggs, J. (2005). *Crime fiction*. Routledge.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Thomas, R. R. (1999). *Detective fiction and the rise of forensic science* (Vol. 26). Cambridge University Press.

Todorov, T. (2019). The typology of detective fiction (1966). In *Crime and Media* (pp. 291–301). Routledge.

UCLA Institute for Digital Research & Education. (n.d.). *Poisson regression | R data analysis examples.*

Underwood, T. (2019). *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Van Dine, S. S. (2015). *Twenty rules for writing detective stories.* Booklassic.

Vezzani, F., & Di Nunzio, G. M. (2019). (Not so) Elementary, my dear Watson! A different perspective on medical terminology. *Umanistica Digitale*, *6*, 59–75.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432. https://doi.org/10.1002/asi.22708

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., & Barabási, A.-L. (2019). Success in books: Predicting book sales before publication. *EPJ Data Science*, *8*(1), 31. https://doi.org/10.1140/epjds/s13688-019-0208-6

Wilcox, R. R. (2023). *A guide to robust statistical methods*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41713-9

Wilson, P. K. (2019). Arthur Conan Doyle | Biography & facts. In *Encyclopædia Britannica.* https://www.britannica.com/biography/Arthur-Conan-Doyle

Yucesoy, B., Wang, X., Huang, J., & Barabási, A.-L. (2018). Success in books: A big data approach to bestsellers. *EPJ Data Science*, *7*(1), 7. https://doi.org/10.1140/epjds/s13688-018-0135-y