LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Masterarbeit

im Studiengang Computerlinguistik

an der Ludwig-Maximilians-Universität München

Fakultät für Sprach- und Literaturwissenschaften

Analyzing Linguistic Patterns in Human-Generated Humor: A Computational Approach

vorgelegt von
Xiaoyu Zhao

Betreuer:	Dr. Yang Janet Liu
Prüfer:	Prof. Dr. Barbara Plank
Bearbeitungszeitraum:	01.April.2025 – 25.July.2025

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, den 29.July.2025

. . . . . .
Xiaoyu Zhao

Erklärung der verwendeten KI-Tools

Ich versichere, dass ich diese Arbeit eigenständig, ohne jede externe Unterstützung, außer den unten aufgeführten Ressourcen, angefertigt habe.

Purpose	**Section(s)**	**Tool**
Grammar	All sections	ChatGPT 4o, DeepL
Academic phrasing	Introduction, Methodology, Results	ChatGPT 4o
Translation	Appendix Titles	DeepL
Data grouping suggestions	Section 3, 4.2	ChatGPT 4o
Visualization tuning	Figures Adjustment	ChatGPT 4o
Features Selection for Analysis	Appendix B	ChatGPT 4o

München, den 29.July.2025

.....
Xiaoyu Zhao

Abstract

This thesis presents a computational stylistic analysis of humor in the Bulwer-Lytton Fiction Contest (BLFC) corpus, a collection of deliberately bad opening sentences to imaginary novels. Drawing on a multidimensional framework, the study examines three key linguistic domainssyntactic complexity, lexical creativity, and rhetorical strategyacross genres and over time.

The research addresses three central questions: (1) What computational frameworks and linguistic features have been prioritized in prior humor research? (2) How do linguistic patterns differ across humorous genres? (3) How has humorous style evolved diachronically from the 1990s to the 2020s?

Using tools such as Stanza for NLP analysis and Kruskal–Wallis testing for significance, the study extracts and compares structural features from 1,620 genre-labeled entries spanning from 1996 to 2024. Results reveal robust genre-specific variation: for example, *Purple Prose* and *Romance* entries exhibit high rhetorical and syntactic density, whereas *Science Fiction* and *Odious Outliers* favor lexical deviation and brevity. Temporal analysis indicates a shift from florid, clause-heavy constructions toward stylistic minimalism and lexical playfulnesssuggesting an aesthetic turn toward irony, compression, and postmodern humor.

The findings contribute to computational humor research by offering empirical insights into how humor is shaped by genre conventions and cultural change. The thesis also demonstrates the value of combining corpus stylistics with genre theory to track stylistic evolution in creative language use.

Contents

1. Introduction

Humor is a uniquely human mode of expression that both reflects and challenges the norms of communication. As a deeply contextual and culturally situated phenomenon, it eludes simple formalization and resists straightforward interpretation. From spontaneous wit in everyday conversation to crafted literary parody, humorous language draws on multiple levels of inference, creativity, and shared knowledge. For linguists and computational researchers alike, the complexity of humor presents both a theoretical puzzle and an empirical opportunity.

This thesis approaches humor through a computational stylistic lens, aiming to uncover how structural and rhetorical variation contributes to humorous effect across genres and time. By combining insights from cognitive linguistics, semantics/pragmatics, and natural language processing, the study seeks to address how stylistic creativity, genre conventions, and pragmatic inference interact in humorous text production.

1.1. Background

Humor is a pervasive and multifaceted phenomenon in human communication, shaping social interaction, cultural identity, and linguistic expression. As both a cognitive and cultural construct, it reflects shared knowledge, psychological mechanisms, and stylistic creativity. Over the past decades, humor has been examined through various disciplinary lenses, including cognitive science, pragmatics, and linguistic theory. Foundational models, such as Raskin's Script Theory of Humor (Raskin, 1984), and Attardo's General Theory of Verbal Humor (Attardo and Raskin, 1991), have emphasized the role of expectation violation, semantic incongruity, and interpretive shifts in the humorous effect.

Pragmatic approaches have further illuminated the contextual mechanisms underlying humor. Grice's Cooperative Principle (Grice, 1975), along with Sperber and Wilson's (Sperber and Wilson, 1986) Relevance Theory, underscores the importance of implicature and mutual inference in producing and recognizing humorous meaning. Studies such as that of Clark and Marshall (1981) demonstrate how humor relies on shared background knowledge and common ground, particularly in conversational settings.

While traditional research has provided valuable insights into the psychological and pragmatic underpinnings of humor, the development of computational methods have enabled more systematic, large-scale analyses of humorous language. In recent years, feature-based approaches have identified recurring linguistic markers of humorsuch as alliteration, antonymy, and syntactic play (e.g., Mihalcea and Strapparava 2005a; Yang et al. 2015a; Hossain et al. 2019). Nevertheless, most computational studies remain focused on one-liner jokes (Chen and Soo, 2018) or anonymized social media data (Raz, 2012), often lacking sensitivity to genre, authorship, or stylistic context.

The *Bulwer-Lytton Fiction Contest* (BLFC), an annual competition known for its intentionally overwrought and parodic first sentences to imaginary novels, offers a particularly rich corpus for exploring human-authored humor beyond the scope of conventional joke formats. BLFC entries are marked by deliberate stylistic excess, syntactic complexity, lexical inventiveness, and rhetorical exaggerationfeatures that challenge normative literary conventions in a way that is intentionally humorous. Despite the contest's cultural visibility and the linguistic ingenuity it showcases, it has remained largely unexamined in computational humor research.

Moreover, the longitudinal nature of the BLFCspanning over four decadesmakes it an ideal dataset for exploring diachronic developments in humor. Shifts in cultural references, genre par-

ody, and rhetorical fashion may offer insights into how humorous styles adapt over time in response to broader social and cultural trends. Yet these temporal and genre-specific dynamics have received little scholarly attention, leaving open questions about the evolution of humor and its formal linguistic manifestations.

1.2. Motivation

This thesis is motivated by the intersection of several timely research interests: humor as a cognitive and linguistic construct, computational stylistics as a method for analyzing language patterns, and genre theory as a framework for understanding stylistic variation. The BLFC corpus presents a unique opportunity to bring these dimensions together. The entries reflect not only creative failures in literary style but also deliberate manipulations of form, tone, and conventionall of which are ripe for computational exploration.

From a methodological standpoint, this study is also inspired by the growing relevance of NLP tools such as Stanza (Qi et al., 2020) in linguistics research. The possibility to systematically extract syntactic, lexical, and rhetorical features from humorous texts opens new avenues for measuring stylistic tendencies in a reproducible and scalable manner. In addition, examining genre-specific and diachronic patterns in the BLFC dataset aligns with broader questions in humor research regarding how stylistic creativity evolves over time and how genre conventions shape humorous expression.

1.3. Aims and Research Questions

The overarching aim of this thesis is to explore the linguistic and cultural dimensions of humor using corpus-based and computational methods. Through a multidimensional analysis of *Bulwer-Lytton Fiction Contest* entries from 1982 to 2024, the study investigates how syntactic complexity, lexical creativity, and rhetorical strategies vary across genres and evolve over time. The project also seeks to evaluate how well existing computational tool (Stanza) captures humor-relevant linguistic features.

To this end, the study is guided by the following research questions:

- **KRQ1**: What computational frameworks and linguistic features have been prioritized in prior research on humor text understanding in NLP?
- **KRQ2**: How do linguistic patterns (syntactic complexity, lexical creativity) differ across genres in the Bulwer-Lytton Fiction Contest entries?
- **KRQ3**: What temporal shifts in humor style (e.g., cultural references, absurdity) are observable in Bulwer-Lytton entries from the 1980s to the present?

Ultimately, this research aims to enrich our theoretical understanding of humor, contribute to genre-sensitive humor analysis, and support the development of more nuanced computational humor systems.

1.4. Structure of the Thesis

This thesis is organized into five chapters. Chapter 2 introduces the theoretical and conceptual background of this study, reviewing the main frameworks of humor studies and computational stylistics. The chapter further addresses the first research question by discussing the relevant literature. Chapter 3 outlines the methodological design, including data description, corpus construction, feature extraction and analysis tools. Chapter 4 presents the main findings by genre and asynchronous mode. The chapter also explores the second and third research questions and

discusses the limitations of this study. Finally, Chapter 5 summarizes the core ideas of the thesis and outlines future research directions.

2. Theoretical Foundations and Computational Approaches to Humor Understanding

2.1. Linguistic Theories of Humor

Understanding humor from a linguistic perspective offers vital theoretical grounding for computational modeling and genre analysis. Humor is a deeply social, pragmatic, and cognitively layered phenomenon, and its linguistic study has yielded a wide range of theories, typologies, and genre-based insights. This section provides a structured overview of key theories and mechanisms that underlie humor, focusing on those that inform computational approaches in NLP. It also sets the foundation for the feature-driven analysis framework employed in this study, which focuses on syntactic complexity, lexical diversity, and rhetorical strategies.

2.1.1. The Three Classical Theories of Humor

The foundation of humor research is based on three influential classical theories, each of which provides a unique perspective on explaining humor effects.

Superiority Theory, dating back to Plato and Hobbes, posits that humor arises from feelings of dominance or triumph over others (e.g., Morreal 1983; Gruner 1997). While more psychological than linguistic, this theory has implications for analyzing aggressive, sarcastic, or disparaging humor, particularly in genres like satire or roasts.

Relief Theory, championed by Freud (1960), views humor as the release of psychic energy or social tension. Linguistically, this is relevant to taboo-breaking jokes, sudden topic shifts, and humor that flouts social norms, many of which surface in internet humor and stand-up routines.

Incongruity Theory, advanced by Kant (1790) and Schopenhauer (1819), remains the most linguistically pertinent. The theory posits that humor arises when there is a mismatch between expectations and reality, triggering surprise and reanalysis. This two-stage incongruity-resolution process is particularly salient in linguistic humor, where the setup of an utterance leads the audience toward one interpretation, only to be subverted by an unexpected punchline.

This core mechanism has become central to computational humor research, inspiring models for joke detection, punchline generation, and the simulation of semantic and syntactic surprise. Modern refinements include the Semantic Script Theory of Humor (SSTH, Raskin 1984), which formalizes humor as arising from a binary script opposition within a shared semantic space (e.g., expected vs. absurd, serious vs. playful). The General Theory of Verbal Humor (GTVH, Attardo and Raskin 1991) extends this by integrating six knowledge resources: script opposition, logical mechanism, situation, target, narrative strategy, and language, to explain structural variation in humorous texts. Together, SSTH and GTVH offer a principled way to represent how linguistic structures encode incongruitythe clash between incompatible scripts, expectations, or semantic framesand resolution, the interpretive shift that reconciles or reinterprets this clash to produce a humorous effect. These concepts serve as theoretical anchors for the feature-based analysis framework of this study, which examines humor in terms of syntactic deviation, lexical creativity, and rhetorical construction across genres.

These classical theories, especially incongruity, continue to shape modern linguistic models of humor. They lay the conceptual groundwork for the semantic, pragmatic, and rhetorical mecha-

nisms explored in the following sections.

2.1.2. Linguistic Mechanisms of Humor

From a linguistic approach, humor is often generated through systematic manipulation of linguistic structures. Drawing on insights from prior research, this study proposes a categorization of four primary mechanisms through which computational models capture humor-related stylistic variation. This framework serves to bridge theoretical accounts of humor with feature-based modeling strategies.

Semantic Shift This mechanism involves the exploitation of lexical ambiguity, polysemy, and homophony. Puns and riddles, in particular, thrive on word-level incongruity, triggering cognitive reanalysis and surprise (e.g., Raskin 1984; Hempelmann 2005; Mihalcea and Strapparava 2005a). The semantic shift plays with expectation through sudden reinterpretation, a central strategy in verbal humor and a persistent challenge for word-sense disambiguation systems. Computationally, this has motivated research on semantic similarity, pun recognition, and sense disambiguation (e.g., Hill et al. 2015; Kao et al. 2016; Zhong and Ng 2010).

Pragmatic Violation Humorous utterances frequently flout Grice's cooperative principles (Grice, 1975). For example, violating the maxim of quantity by giving too much or too little information can create ironic understatement or absurd over-specification. Ignoring relevance can introduce unexpected topic changes, while breaching the maxim of manner with vagueness or opacity often results in comedic confusion (Attardo, 1993; Burgers et al., 2012). These violations deliberately subvert conversational norms, inviting reinterpretation, and signaling humorous intent.

Syntactic Surprise and Structure Manipulation Unexpected syntactic forms also serve as triggers for humor. Garden-path sentences temporarily mislead the reader's parsing process, creating a delayed realization effect. Humor can also arise from syntactic ambiguity, unconventional parataxis, or conditional statements used to ironic ends (e.g., "If I agreed with you, we'd both be wrong.") (Kao et al., 2016; Bergen and Binsted, 2003). These forms of structural incongruity reflect the underlying joke mechanisms and are increasingly detected through modern dependency parsers.

Rhetorical Devices Classical rhetorical strategies such as irony, hyperbole, metaphor, simile, and litotes contribute to humorous discourse by infusing figurative and evaluative dimensions (Attardo, 1994; Veale, 2012). For example, exaggeration (hyperbole) and understatement (litotes) can invert expectations, while similes and metaphors juxtapose semantically distant domains for comic effect. These devices often appear in stylized humorous texts, including satirical headlines and comic fiction (Reyes et al., 2013).

Information Structure Humor also relies on a setup–punchline organization that builds context before introducing a twist. This structure may also include reframing, where an established schema is reinterpreted or delayed resolution, where humor is derived from postponed clarification (Norrick, 2003). In this structure, the setup establishes an expectation or schema, often through familiar narrative or referential cues, while the punchline introduces a twist that subverts or reframes that expectation, producing surprise and amusement. The management of information flow and timing plays a crucial role in humorous effect and is a foundation for rhetorical and discourse-level humor analysis.

2.1.3. Genres and Functions of Humor

Humor is not monolithic; it varies according to genre, intent, and communicative context. The genre-specific functions of humor are key to understanding the stylistic variations, linguistic patterns, and computational detectability of humor. The corpus used in this study, the Bulwer-Lytton Fiction Contest (BLFC), comprises humorous entries submitted under a variety of imagined literary genres, each of which activates different stylistic conventions, reader expectations, and rhetorical strategies.

The main genres represented in the BLFC corpus are annotated by contest organizers and participants (e.g., Mihalcea and Strapparava 2005b; Burfoot and Baldwin 2009):

- **Adventure**: Characterized by dramatic tension, exotic settings, and exaggerated action. Humor arises through melodramatic parody, inflated narrative pacing, and absurd plot twists (Scott, 1992).
- **Fantasy**: Typically includes mythical creatures, magical systems, and elevated diction. Humor often draws on deflationary irony, anachronism, or parody of high fantasy tropes (Jackson, 1981).
- **Historical Fiction**: Incorporates archaic style and period-specific references. Humor arises through deliberate archaism, contrast between modern and historical frames, or the mimicry of literary register (Cuddon, 1999).
- **Romance**: Defined by elevated emotional narration and formulaic expression. Humor often parodies sentimentality using hyperbole and absurd similes (Scott, 1992).
- **Science Fiction**: Includes futuristic scenarios, technical jargon, and speculative settings. Humor is often achieved through technobabble, juxtaposition of advanced technology and banal scenarios, or alien misunderstandings (Heinlein et al., 1959).
- **Western**: Draws on frontier settings, terse dialogue, and stereotypical machismo. Humor appears through deadpan exaggeration, stylized threats, or ironic stoicism (Wister et al., 2020).
- **Purple Prose**: Many entries defy neat classification, but share the hallmark of intentional overwriting, such as lengthy sentences, ornate modifiers, and florid similes. These function metalinguistically to parody the literary affectation itself (Nixon, 2008).

These genre distinctions help structure both the stylistic expectations and the evaluation strategies used in humor detection. As each genre operates with its own conventions and communicative goals, analyzing humor within and across genres allows for finer-grained insights into how specific features including syntactic, lexical, and rhetorical, interact with genre-specific reader expectations. This also enables genre-aware computational modeling, which is crucial in capturing the stylistic diversity embedded in humorous texts.

2.1.4. Cognitive and Conceptual Approaches

Linguistic humor is also informed by cognitive models that explain how readers or listeners construct humorous interpretations. Two frameworks are particularly relevant:

Conceptual Blending Theory Fauconnier and Turner (2002) proposes that humor arises from the integration of incompatible mental spaces into a blended space, where the incongruity produces a comic effect. This theory supports analyses of similes, analogical mapping, and frame-shifting, and aligns closely with the rhetorical strategies assessed in our study.

Frame Semantics Fillmore (1982) explains how humor can result from activating a familiar conceptual frame and then violating its typical structure. Frame clashessuch as reinterpreting everyday schemas (e.g., restaurant scripts) in absurd or inappropriate ways, are common in joke

construction. This notion informs both the lexical and rhetorical dimensions of humor detection.

Together, these linguistic and cognitive theories provide a theoretical scaffold for the computational analysis that follows. By aligning linguistic mechanisms and conceptual models with the proposed framework of syntactic complexity, lexical diversity, and rhetorical strategies, this study seeks to bridge traditional humor theory with contemporary NLP techniques.

2.2. Computational Humor

Humor is a fundamental aspect of human communication, reflecting creativity, culture, and cognitive complexity. Its study not only offers insight into linguistic play and pragmatic nuance but also presents a unique challenge for computational models. Unlike tasks grounded in objective semantics or syntax, humor involves subjectivity, ambiguity, and cultural variation, making it difficult to define, detect, and generate algorithmically. In recent years, computational humor has evolved from niche experimentation to a growing subfield of Natural Language Processing (NLP), driven by advances in machine learning, large-scale datasets, and multimodal modeling. This section provides a systematic overview of the data, tasks, and evaluation methods that underpin current approaches to computational humor.

2.2.1. Datasets

In computational humor research, datasets serve as the empirical backbone for model training, evaluation, and comparative benchmarking. Unlike many other NLP tasks, humor analysis is uniquely challenging due to the subjective, context-dependent, and culturally situated nature of humor. What one individual or community finds amusing, another may not, and this variability complicates both the construction of datasets and the design of computational models. Furthermore, humor frequently depends on timing, delivery, shared knowledge, and linguistic playfulnesselements that are often difficult to capture in isolated texts. As such, the development of high-quality humor datasets is not merely a technical undertaking but a conceptual one that reflects deeper questions about what constitutes humor, how it is communicated, and how it can be computationally modeled.

Humor datasets can be characterized along several dimensions: modality, scale, annotation methodology, and intended application. First, with regard to **modality**, datasets are typically either text-only or multimodal. Text-only datasetssuch as *r/Jokes*, *Hashtag Wars*, or *SemEval-2021 Task 7* (Meaney et al., 2021)focus exclusively on verbal humor and are often drawn from online forums, headlines, or social media. In contrast, multimodal datasets like *UR-FUNNY* (Hasan et al., 2019) combine text with audio, visual, and prosodic cues, acknowledging that humor in real-world settings often emerges through the interplay of multiple sensory channels. Second, scale varies widely. Some datasets are relatively compact and domain-specific, containing only a few thousand humorous instances. Others, like *r/Jokes*, comprise over half a million entries and are suited for training large-scale neural models. Dataset size plays a crucial role in model generalizability, especially when training deep architectures that require extensive data to capture the stylistic and semantic nuances of humor.

Third, annotation methods differ significantly. Some datasets employ binary classification (humorous vs. non-humorous), while others use scalar ratings to quantify humor intensity or offer categorical labels to distinguish types of humor. For instance, *SemEval-2021 Task 7* (Meaney et al., 2021) assigns numerical funniness scores to edited news headlines, enabling humor intensity modeling rather than simple detection. Other corpora, such as *Is This A Joke?*, combine social media and crowd-sourced joke data with binary annotations, creating large but noisy datasets.

Finally, humor datasets are tailored to support a range of computational tasks, including:

- **Detection**: Identifying whether a given text is humorous.

- **Ranking**: Assessing the relative funniness of multiple texts.
- **Generation**: Producing original jokes or humorous variants.
- **Multimodal understanding**: Interpreting humor conveyed through combined channels such as text, video, and audio.

Table 2.1 summarizes a selection of widely used datasets in computational humor research, highlighting differences in data type, annotation style, source, size, and target applications. Table 2.1 summarizes core datasets used in computational humor research. To illustrate the style and structure of entries found in these datasets, a few representative examples are provided below:

- **SemEval-2021 Task 7 Dataset** (edited headlines): *"A fat woman just served me at McDonald's and said 'Sorry about the wait.' I replied, 'Don't worry, you'll lose it eventually.'"*
- **r/Jokes Dataset** (Reddit full jokes): *"I told my niece to get me a newspaper. She laughed and said, 'Just use my phone.' So I smashed her phone against the wallto kill a spider."*
- **"Is This a Joke?" Dataset** (short one-liners): *"Why don't scientists trust atoms? Because they make up everything."*
- **Hashtag Wars Dataset** (Twitter-based puns under a theme): *"Harry Potter and the Order of the Big Mac— #FastFoodBooks"*; *"The Girl With The Jared Tattoo — #FastFoodBooks."*
- **UR-FUNNY Dataset** (spoken humor, multimodal): *"Imagine a jellyfish waltzing in a library while thinking about quantum mechanics."*

Dataset	Data Type	Annotation Type	Source	Size	Language
SemEval-2021 Task 7 (Meaney et al., 2021)	Text-only (edited headlines)	Humor edit ranking and classification	News headlines	∼10,000	English
r/Jokes (Weller and Seppi, 2020)	Full jokes (setup + punchline)	Humor score (upvotes)	Reddit (r/Jokes)	573,000	English
"Is This A Joke?" (Faruqi and Shrivastava, 2018)	One-liners and short sentences	Humor vs. non-humor classification	Social media + joke sites	400,000	English
Hashtag Wars (Potash et al., 2017)	Tweets (short-form creative texts)	Implicit humor via hashtags	Twitter (#HashtagWars)	∼200,000	English
UR-FUNNY (Hasan et al., 2019)	Multimodal (video + subtitles)	Humor vs. non-humor; multimodal alignment	Stand-up comedy shows	∼20 hours	English

Table 2.1.: Datasets used in computational humor research

Despite these advances, several challenges persist. **Subjectivity** remains a core issue: the same text may elicit laughter from one reader and indifference or confusion from another. This subjectivity can result in inconsistent annotations and what is known in machine learning as *noisy*

supervision. Even when humor scores are averaged across multiple annotators, inter-rater reliability may remain low. Another challenge is **cultural bias**. Many humor datasets reflect dominant linguistic and cultural normsoften English-speaking, Western, internet-based communitieswhich limits their cross-cultural applicability. Jokes that rely on puns, idioms, or pop culture references may fail to translate across languages or be misunderstood outside their original social context.

Furthermore, many datasets are decontextualized, especially those composed of short texts such as headlines or tweets. Humor often requires shared background knowledge or situational context to be fully understood. For instance, political satire or wordplay can lose its effect when the referents are not known to the reader or model. Without such contextual cues, computational models struggle to correctly interpret or generate appropriate humorous content.

To mitigate these issues and support more robust humor modeling, several directions for future dataset development are evident. First, greater **contextual and cultural awareness** is needed. This can involve curating multilingual datasets, including metadata or background knowledge, and involving diverse annotator populations. Second, **fine-grained humor categorization** would enhance model interpretability and predictive accuracy by distinguishing among subtypes of humor such as sarcasm, irony, absurdity, and slapstick. Third, expanding and diversifying **multimodal and interactive datasets**such as comedy performances, visual memes, or dialogic humorwill allow models to learn how humor functions dynamically in human communication. Lastly, improvements in **annotation methodology** are crucial. Incorporating crowd-sourced labels with multiple annotators per instance, and using probabilistic modeling to capture humor as a distribution rather than a binary outcome, could yield more realistic and flexible training data.

In conclusion, humor datasets are indispensable for advancing computational approaches to humor analysis. Yet, their design must rise to the cognitive and cultural complexity of humor itself. Only by incorporating diversity, context, and subjectivity can future datasets support the development of humor-aware systems that approach human-like understanding and creativity.

2.2.2. Tasks and Methods

Computational humor is no longer a peripheral curiosity within NLP but an increasingly significant domain that reveals how machines can model and interpret creative, affective and socially contextualized aspects of language. Humor analysis presents a unique challenge due to its reliance on subtle linguistic features, world knowledge, cultural references, and subjective interpretation. As the field evolves, researchers have identified and pursued a range of core tasks including humor recognition, explanation, and generation that form the foundation for developing systems capable of understanding or producing humorous content. This section reviews these key computational tasks and the methods employed to address them, focusing on recent trends and theoretical advances.

1. Humor Recognition and Classification

Humor recognition and humor classification are foundational tasks in computational humor. The former aims to determine whether a given text is humorous, while the latter seeks to identify the *type* or *mechanism* of humor involvedsuch as puns, irony, or satire. These tasks are complementary: recognition establishes the *presence* of humor, whereas classification provides insight into *how* it is linguistically or pragmatically constructed. In practice, both are typically formulated as *binary* or *multi-class classification* problems. Models must capture complex linguistic patterns, semantic incongruities, and stylistic cues that signal humorous intent.

Earlier approaches relied on manually crafted *features* and *supervised classifiers*, often limited by domain specificity and generalizability. More recent work has shifted toward *deep learning* architectures, particularly *transformer-based* models, which leverage large-scale pretraining and contextualized representations to enhance performance and cross-domain robustness.

Early methods for humor recognition relied on manual feature engineering combined with traditional supervised classifiers such as SVMs and decision trees (Mihalcea and Strapparava, 2005a). These approaches incorporated syntactic, semantic, and stylistic features to distinguish

humorous from non-humorous texts. However, the emergence of deep learning has significantly advanced the field. Neural networks, particularly Convolutional Neural Networks (CNNs) and transformer-based architectures, have enabled models to capture complex patterns without relying on handcrafted features. For example, CNNs augmented with Highway Networks (Chen and Soo, 2018) demonstrated improved performance by automatically learning hierarchical representations of humor cues, such as semantic incongruity, wordplay, puns, etc. Fine-tuned transformer models such as BERT and its variants have become the new standard, as illustrated by Weller and Seppi (2020) in their work on Reddit jokes. Researchers have also developed hybrid and theory-informed approaches. The THInC framework (Marez et al., 2024) integrates cognitive theories of humor, such as incongruity-resolution and benign violation theories into rule-based and neural systems, improving interpretability. Comparative tasks such as SemEval-2017's Hashtag Wars (Potash et al., 2017) proposed ranking-based humor recognition rather than simple binary classification, acknowledging that humor often exists on a continuum. Multilingual and cross-cultural efforts have also emerged, such as Dutch humor detection using RobBERT (Delobelle et al., 2020), and humor modeling in Russian using large-scale datasets and ULMFiT fine-tuning (Blinov et al., 2019).

However, a major challenge persists in the form of cultural and linguistic variability. Humor is deeply embedded in language-specific idioms, cultural knowledge, and social norms. This makes transferability across languages or communities difficult and underscores the need for adaptive, context-aware models. For example, in their study of Dutch humor detection, (Delobelle et al., 2020) addressed this issue by generating negative examples tailored to the target language, demonstrating how culturally grounded methods can enhance robustness in multilingual humor recognition.

2. Humor Understanding and Explanation

Beyond recognition, humor understanding seeks to answer *why* a particular utterance is funny. This task moves toward semantic interpretation and cognitive modeling, aiming to identify the mechanisms, such as incongruity, surprise, ambiguity, or cultural reference that generate humor. Compared to classification, this task is less well-defined and inherently more interpretive, often requiring systems to combine linguistic analysis with world knowledge and commonsense reasoning.

One foundational approach is humor anchor extraction (Yang et al., 2015a), which identifies textual elements that trigger humor within a sentence. More recent methods have employed masked language models to probe humor's semantic structure. For example, JokeMask (Li et al., 2022) uses masked prompts to distinguish humorous shifts from offensive or neutral statements, thereby modeling implicit semantic incongruity. (Hessel et al., 2023) introduced the "Do Androids Laugh at Electric Sheep?" benchmark, based on The New Yorker Caption Contest, to explore structured humor understanding through tasks involving common sense and contextual inference.

Multimodal approaches have also gained traction. Humor often arises in audiovisual formats through tone, gesture, or timing. To account for these dimensions, Baluja (2025) introduced a multimodal prompting technique, combining textual and audio inputs to improve large language models' understanding of pun-based humor. This work suggests that humor comprehension benefits from integrating modalities beyond the textual domain.

Despite promising developments, challenges remain. Humor understanding is heavily context-dependent, requiring models to draw on background knowledge, intertextual references, and dynamic social cues. Building systems capable of such integrated reasoning remains an open research problem.

3. Humor Generation

Humor generationproducing original humorous contentis arguably the most ambitious task in computational humor. While early work relied on rule-based and template-driven systems, current research explores neural architectures and cognitive modeling to simulate creative, context-aware humor.

Initial systems such as JAPE and LIBJOG (Binsted and Ritchie, 1994) generated pun-based riddles using syntactic templates and lexical relations. These laid the foundation for more advanced lexical and semantic substitution techniques, including HAHAcronym (Stock and Strapparava, 2005) and ambiguous compound generators (Sjöbergh and Araki, 2008), which manipulated word forms and meanings to produce humorous outputs. More recent systems have shifted to deep learning methods. Alnajjar and Hämäläinen (2021) used supervised neural models to transform neutral headlines into humorous ones, while Tikhonov and Shtykovskiy (2024) proposed a multistep reasoning framework that mirrors cognitive humor construction, incorporating setup, expectation, and punchline resolution.

Of particular interest are template extraction and infilling approaches, such as those by Goel et al. (2024), which use BERT's attention mechanisms to identify humor-relevant structures and GPT-4 to fill in content that mimics construction of human jokes. These models suggest that future systems may integrate learned humor patterns with flexible, generative capabilities.

However, significant challenges remain. Humor generation requires balancing creativity, novelty, and coherence, while avoiding offensiveness or cultural insensitivity. Additionally, evaluating generated humor is notoriously difficult, as funniness is subjective and varies with audience, context, and delivery.

2.2.3. Evaluation

Evaluation, a non-negligible aspect of computational humor research, refers to the process of assessing the quality, intensity, and appropriateness of humor generated or detected by computational models and has a direct impact on the validity and reliability of humor models. Because humor is inherently subjective, the development of reliable assessment metrics and methods is critical to assessing a model's ability to understand, generate, and rank humor. In this section, various methods used to assess the humor of computational models will be explored, including automated metrics and manual evaluation.

1. Automatic Evaluation Metrics

Automatic evaluation metrics provide scalable and reproducible tools for assessing humor models, particularly in tasks involving large-scale detection or ranking. For classification-based tasks such as binary humor detection, standard metrics include accuracy, precision, recall, and F1 score. Accuracy reflects the proportion of correctly labeled instances, while precision and recall measure the trade-off between false positives and false negatives. F1 score, the harmonic mean of precision and recall, is particularly informative in humor recognition, where class imbalance and annotation noise are common. For instance, humor anchor extraction has been shown to improve classification precision by focusing on linguistic markers of humor (Yang et al., 2015a), while large-scale models such as ULMFiT have demonstrated improved F1 scores over traditional baselines (Blinov et al., 2019).

In tasks involving humor rating or rankingwhere models predict degrees of funninessregression-based metrics are used. Mean Squared Error (MSE) quantifies the average deviation between predicted and human-assigned scores, while correlation coefficients (Pearson or Spearman) capture the strength of alignment between model outputs and human judgments. For example, the SemEval-2021 Task 7 used MSE to assess humor intensity prediction in edited headlines (Ma et al., 2021; Smădu et al., 2021), and Pearson correlation was adopted in SemEval-2017 HashtagWars to evaluate humor rankings from tweets (Potash et al., 2017).

However, automatic metrics face limitations when applied to humor. Most notably, they struggle with subjectivity and lack sensitivity to context or cultural nuance. Humor that relies on subtle references, shared experiences, or irony may elude models trained on surface-level features. Additionally, automatic metrics are ill-equipped to evaluate humor generation, where notions of originality, creativity, and appropriateness go far beyond simple numerical scores.

2. Human Evaluation

Despite the growing reliance on automatic metrics, **human evaluation** remains essential due to the **subjective nature of humor**. Human raters provide insights into the more nuanced and qualitative aspects of humor that automated systems might miss. In humor generation tasks, additional qualitative criteria are often assessed. Human raters may be asked to judge:

- **Funniness Ratings** (Amin and Burghardt, 2020): Human evaluators typically rate the humor on a Likert scale (e.g., 1 to 5, where 1 is not funny and 5 is very funny). This scale allows researchers to measure the **perceived humor** of a text and compare it to automated ratings. It captures the personal and cultural factors that influence humor appreciation.
- **Novelty Checks** (Tikhonov and Shtykovskiy, 2024): This involves asking human evaluators to identify whether a joke is **new** or **recognizable**. This is important in humor generation tasks to ensure that the model generates **original content** rather than repeating common jokes or phrases.
- **Grammaticality and Coherence** (Alnajjar and Hämäläinen, 2021): These checks assess whether the humor is **grammatically correct** and **coherent** within its context. While the primary focus is on humor, evaluating the linguistic quality of the joke is also essential, especially in humor generation tasks. A joke that is grammatically incorrect or incoherent may detract from its perceived funniness.
- **Offensiveness Detection** (Tikhonov and Shtykovskiy, 2024): Human evaluators also play a role in detecting jokes that might be **offensive** or **culturally inappropriate**. Given that humor can sometimes cross sensitive boundaries, models must be assessed for their ability to **filter out offensive content**. Evaluators determine whether the generated humor is harmful or inappropriate, ensuring that the system adheres to ethical standards.

Nevertheless, human evaluation introduces its own challenges. Inter-rater reliability is often low due to personal and cultural differences in humor perception. Researchers mitigate this by aggregating ratings from multiple annotators or using consensus-based methods. Bias is another concern: evaluators bring their own backgrounds and assumptions, which can shape their judgments in unpredictable ways. Moreover, like machines, human raters may also struggle to fully appreciate context-dependent or culturally specific humor without sufficient background knowledge.

3. Hybrid Evaluation Approaches

To address the limitations of both automatic and human evaluation methods, **hybrid approaches** are increasingly being used. These combine the efficiency of automatic metrics with the nuanced insights provided by human evaluators. **Soft Turing Test** (Amin and Burghardt, 2020): In some studies, particularly in humor generation tasks, researchers ask human evaluators to determine whether a piece of humor was generated by a **human** or a **machine**. This helps assess how "human-like" the generated humor is. For instance, (Goel et al., 2024) used this test to evaluate machine-generated jokes by asking human raters to classify the humor as either human-made or machine-generated.

The evaluation of humor in computational linguistics is a complex and multifaceted task. While **automatic metrics** offer scalability and objectivity, they are often limited by the **subjectivity** and **context dependency** of humor. **Human evaluation**, on the other hand, provides essential insights into the subjective aspects of humor, though it introduces challenges related to **inter-rater reliability** and **bias**. As humor models continue to evolve, **hybrid evaluation approaches** that combine the strengths of both methods will likely become the standard for assessing humor in computational systems. Further advancements in evaluation techniques are necessary to develop models that can generate and understand humor in a more human-like, contextually aware manner.

In summary, this chapter has outlined the linguistic, cognitive, and computational foundations of humor understanding. By reviewing key mechanisms, genre-based variations, and state-of-the-art NLP methods, it addresses the first research question concerning the frameworks and linguistic

features prioritized in previous computational humor studies. These insights inform the design of this study's analytical framework, which focuses on syntactic complexity, lexical creativity, and rhetorical strategies across genres and time, providing a theoretically grounded and computationally tractable answer to KRQ1.

3. Methodology

This chapter outlines the design and implementation of a computational genre analysis of the Bulwer-Lytton Fiction Contest (BLFC) corpus. Building on the theoretical and empirical foundations laid out in the previous chapter, the analysis seeks to reveal how linguistic features contribute to the construction of humor across genres and over time. Humor is treated here not merely as an abstract concept but as a textual phenomenon shaped by stylistic patterns, ranging from syntactic structure to lexical creativity and rhetorical flourish. Accordingly, the analysis combines natural language processing (NLP) techniques, linguistically motivated feature extraction, and statistical modeling to examine variation in humorous writing both across genres and through temporal shifts.

3.1. Research Design

The study adopts a multidimensional analytical framework for genre analysis, drawing on three principal levels of linguistic structure: syntactic, lexical, and rhetorical. This feature set is grounded in prior theories of verbal humor, particularly those emphasizing incongruity, pragmatic deviation, and stylistic exaggeration, as reviewed in Chapter 2. Each dimension is operationalized through a set of quantitative linguistic features that allow for the systematic comparison of genre-specific and diachronic patterns. By examining features across these layers, the analysis seeks to capture both surface-level textual patterns and deeper stylistic strategies that contribute to humorous effect.

To support this feature-based analysis, all texts in the corpus are processed through **Stanza**, the Stanford NLP toolkit, using its **English model** (Qi et al., 2020). Each contest entry is parsed through the following modules:

- **Tokenization**: Segments each sentence into tokens, including words, punctuation, and symbols.
- **Part-of-Speech (POS) Tagging**: Assigns grammatical categories to each token, enabling syntactic pattern recognition.
- **Named Entity Recognition (NER)**: Identifies proper names, locations, and other entity types useful in lexical analysis.
- **Dependency Parsing**: Builds syntactic dependency trees, providing structural data on clause complexity and modifier attachment. To complement dependency parsing, constituency parsing is also applied to extract hierarchical structures from phrase-structure trees, enabling the computation of metrics such as tree depth, which reflect the degree of syntactic embedding.

This NLP pipeline provides the foundational linguistic scaffolding for the subsequent feature extraction process. It enables consistent and scalable measurement of stylistic variation across a large corpus of humorous texts.

3.2. Corpus and Data Preprocessing

The primary dataset analyzed in this study comprises 1,620 entries drawn from the Bulwer-Lytton Fiction Contest, covering a temporal range from 1996 to 2024. The contest is known for inviting participants to compose deliberately overwrought or stylistically "bad" opening sentences to

imaginary novels, often marked by excessive elaboration, genre parody, and linguistic absurdity. As such, the BLFC corpus offers a unique resource for studying humorous language that is simultaneously artificial, self-aware, and stylistically marked.

Each entry in the dataset is associated with a genre label, either provided by the contest organizers or manually inferred through close reading by the author. Entries marked as "N/A" by the organizers, indicating no specified genre, were retained in the dataset without reclassification. This decision reflects the intention to preserve the original distribution and allow for comparison between genre-specific and genre-agnostic humor styles. These genre categories represent a wide range of fictional traditions and narrative conventions, many of which are intentionally exaggerated or subverted in humorous ways. Based on contest documentation and previous genre-based humor studies (Kao and Jurafsky, 2012), Attardo (2020), the following major genre categories were standardized for use in this analysis. While additional genre labels appear in the full dataset, only those with sufficient representation were retained for comparative analysis to ensure statistical robustness.

- **Adventure**
- **Purple Prose**
- **Romance**
- **Crime and Detective**
- **Science Fiction**
- **Western**
- **Vile Puns**
- **Odious Outliers**
- **Fantasy**
- **Historical Fiction**
- **Dark and Stormy**

These genre labels are treated as categorical variables in the subsequent analysis. The decision to focus on these categories was guided by considerations of sample size, stylistic coherence, and relevance to humor research.

Prior to analysis, the dataset underwent several preprocessing steps to ensure consistency, interpretability, and computational reliability:

- Normalization of genre labels (e.g., consistent casing, spelling harmonization)
- Validation of year formatting
- Removal of incomplete or corrupted entries
- Token-level cleaning (e.g., elimination of HTML artifacts)

The final cleaned dataset was saved in the TSV format, with each entry represented as a structured row containing full text with columns for year, genre, and entry text. This structure ensures compatibility with NLP tools and facilitates transparency and reproducibility in subsequent processing steps.

3.3. Data Distribution

To better understand the structure of the BLFC corpus, an exploratory analysis was conducted on both temporal and genre-based distributions.

Figure 3.1 presents a heatmap showing the number of entries submitted per genre and year between 1996 and 2024. Several patterns emerge from this visualization. First, core categories such as *Purple Prose*, *Romance*, and *Science Fiction* have maintained relatively steady participation over time. In contrast, genres such as *Odious Outliers*, *Dark & Stormy*, and *Vile Puns* show increased activity in later years, reflecting the contest's evolving taxonomy and participants' shifting stylistic preferences.

Notably, a dense band of unclassified entries (labeled as *N/A*) appears between 1997 and 2005. This concentration likely results from inconsistent or absent tagging practices in the contest's earlier digital recordkeeping. The gradual diversification and expansion of genre labels in more recent years also suggest an increasing awareness of stylistic subcategories within humorous fiction.

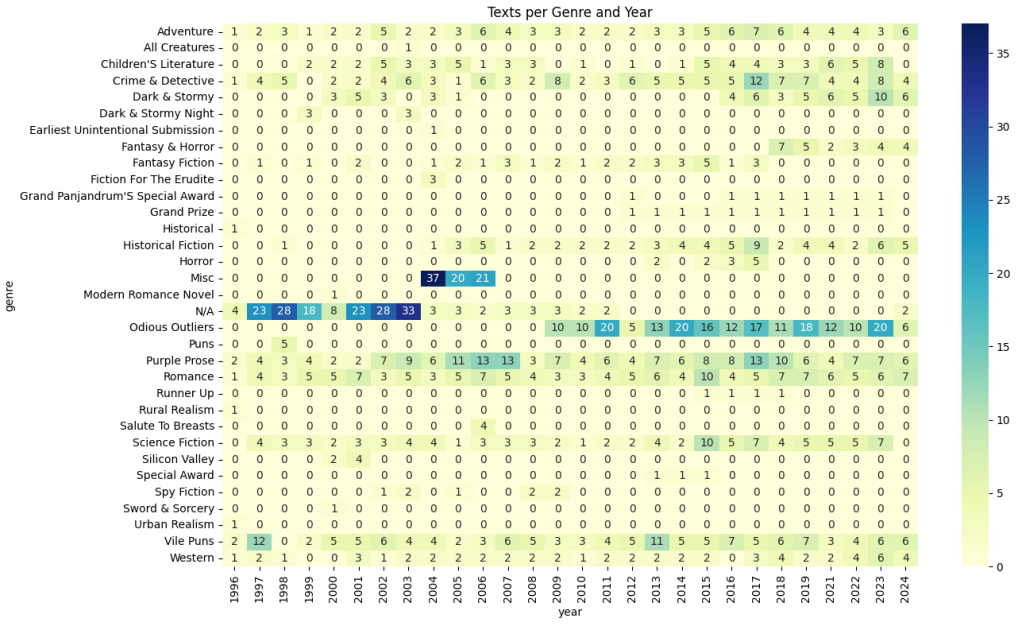

Figure 3.1.: Texts per Genre and Year (1996–2024)

Figure 3.2 shows the annual distribution of entries from 1996 to 2024. The number of entries has risen steadily since the beginning of the 21st century, peaking in 2017 with 99 entries and again in 2023 with 93 entries. This increase may reflect increased online visibility and participation, while the slight decline in 2024 (56 entries) may indicate natural fluctuations in data collection.

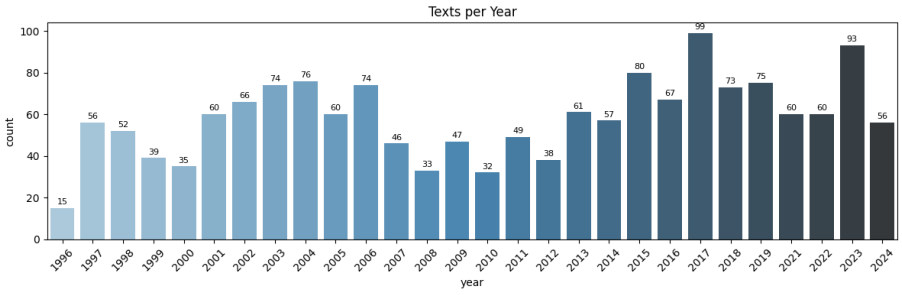

Figure 3.2.: Texts per Year

Figure 3.3 presents the cumulative number of entries per genre across the entire corpus. Odious Outliers emerges as the most represented category, followed closely by *Purple Prose*, *Romance*, and *Vile Puns*, each contributing over 150 entries. These categories are characterized by exaggerated or parodic literary stylization, which aligns with the contest's thematic emphasis on overwrit-

ing. In contrast, some niche or ephemeral genres, such as Spy Fiction, Silicon Valley, and Justin Gustains, have very limited representation, often with fewer than five entries. This distributional imbalance will be taken into consideration during statistical analysis to mitigate the effect of genre sparsity on feature comparisons.

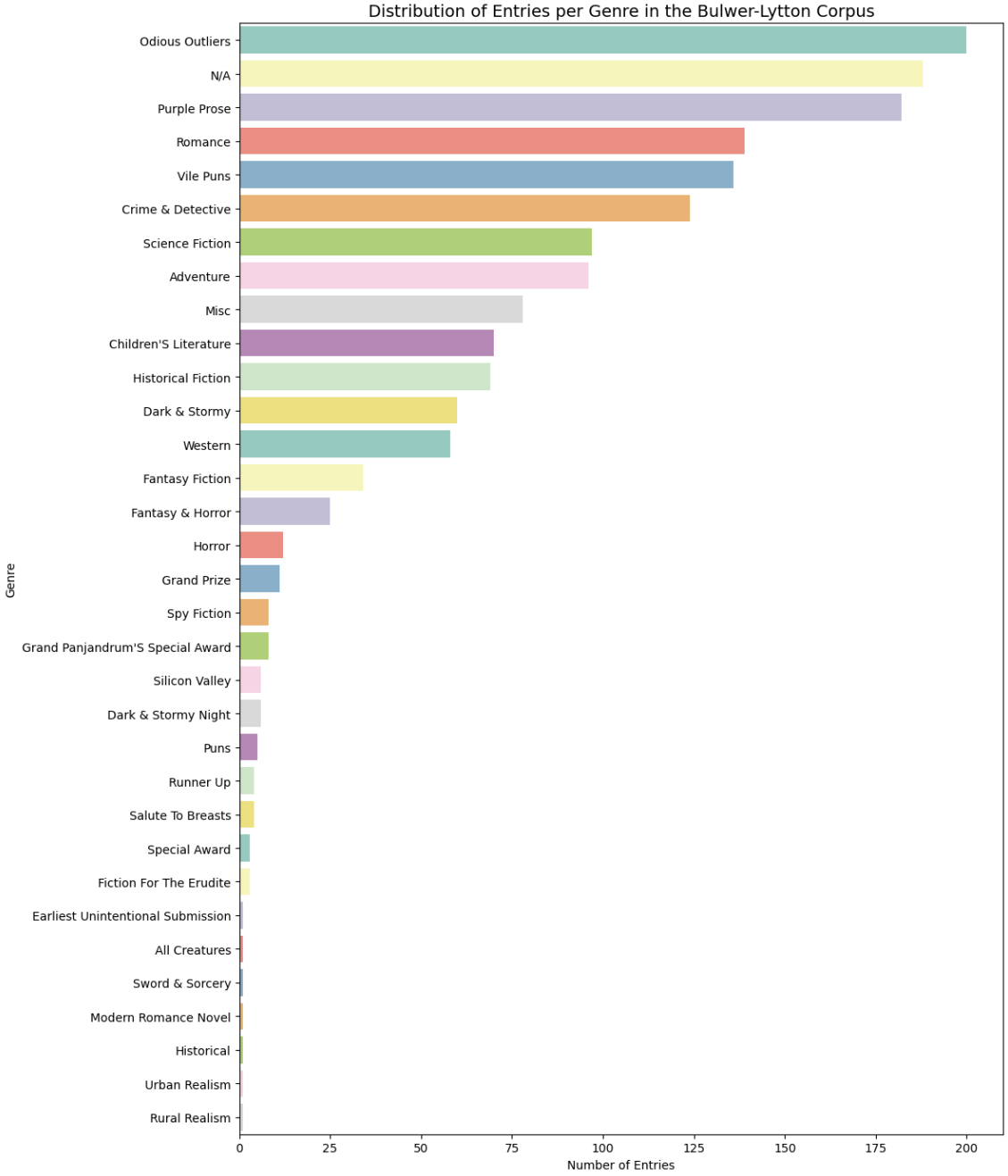

Figure 3.3.: Texts per Genre

The distributional patterns observed in genre and year serve as the basis for subsequent analytical design. In the genre-based analysis, greater emphasis is placed on categories with sufficient sample sizes, such as *Odious Outliers*, *Purple Prose*, and *Romance*, to ensure statistical robustness and interpretability. Genres with very few entries are excluded from comparative testing for data analysis in the next section to avoid skewed or unreliable results.

For diachronic analysis, the dataset is divided into temporal bins of four years each (e.g., 1996–1999, 2000–2003, ..., 2020–2024). This time-binning approach helps to mitigate sparsity

across individual years while preserving the ability to track stylistic and structural shifts over time. Aggregating data into coherent temporal units facilitates more reliable comparison and improves the visibility of longitudinal patterns in humor construction.

3.4. Features Extraction

The linguistic features extracted in this study are informed by core theories of humor, particularly those emphasizing incongruity, pragmatic deviation, and rhetorical overstatement. The analysis focuses on three dimensions of stylistic variation, syntactic, lexical, and rhetorical, each capturing different layers of humorous construction in written form.

3.4.1. Syntactic Features

To systematically assess the structural characteristics of humorous writing in the BLFC corpus, this study adopts a set of syntactic complexity measures rooted in dependency grammar. Syntactic elaboration is a key dimension of stylistic creativity and can serve as an indirect cue to intentional parody, exaggeration, or cognitive effort associated with comprehension. Humor, especially in the BLFC tradition, is often constructed through syntactic overloading, subordination, and rhetorical layering. Therefore, capturing sentence depth, dispersion, and coordination is essential for understanding how humorous effect is achieved through grammatical design. The following five dependency-based metrics were selected to reflect variation in sentence structure and grammatical intricacy across entries:

- **Tree Depth**: The maximum depth of the constituency (phrase-structure) tree, reflecting the hierarchical syntactic embedding within a sentence. Greater depth suggests more nested and structurally complex constructions.
  Example (Purple Prose, 2016):

  > *"When Glenn left the house, the sky was a satin swirl of lavender and steel, with clouds layered like decadent frosting on an apocalyptic cake."*

  This sentence contains several layers of adjectival and prepositional modification (e.g., "a satin swirl of lavender and steel") and its tree depth is 7, indicating deep hierarchical embedding.

- **Dependency Distance**: The average number of tokens between syntactic heads and their dependents, indicating structural dispersion.
  Example (Crime & Detective, 2010):

  > *"She walked into my office wearing a body that would make a bishop kick in a stained glass window."*

  The humorous clause "that would make a bishop kick…" depends on the noun "body" several tokens prior, showing relatively long dependencies. Here its dependency distance is 1.55, using long-distance dependence to break expectations.

- **Clause Ratio**: Number of clause-level dependencies (e.g., clausal modifiers or complements) per sentence.
  Example (Fantasy, 2014):

  > *"As he strolled among the Kenthellians, through sulfur-shrouded ruins, thinking of his destiny, he realized the prophecy had been mistranslated."*

  This single sentence embeds multiple clauses (temporal "as he strolled", participial "thinking…", and the main clause), making it clause-rich. The clause ratio here is 0.143 (medium-high complexity), in line with the fantasy literary style (multiple modifiers, nested structures).

- **Conjunction Count**: Number of coordinating conjunctions (e.g., *and, but*), per sentence, often associated with cumulative or additive phrasing.
  Example (Odious Outliers, 2017):

  > *"Phoebe, age 15, very much regretted not having worn her backup sparkle nail polish and the cat-ear hoodie and her lucky unicorn bracelet."*

  The repeated use of "and" creates an additive list, increasing the conjunction count to 2.

- **Average Sentence Length**: The mean number of tokens per sentence, serving as a general indicator of syntactic elaboration.
  Example (Vile Puns, 1996):

  > *"Because the Indians of the high Andes were believed to have mystical knowledge of the stars and were excellent at weaving, they were called 'Inca-nites'."*

  This complex joke sentence contains multiple clauses and modifiers, resulting in a relatively high token count (29 words).

3.4.2. Lexical Features

While syntactic structure governs how sentences are formed, lexical choice plays a crucial role in shaping the tone, originality, and stylistic texture of humorous writing. In the context of the BLFC corpus, where inventiveness and parody are core to the contest's aesthetic, word selection often involves deliberate exaggeration, archaic references, or unexpected collocations. To capture this lexical creativity and variation, several quantitative measures were applied, each targeting a distinct facet of vocabulary use.

- **Type-Token Ratio (TTR)**: Measures vocabulary diversity by dividing the number of unique tokens by the total number of tokens in an entry.
  Example (Purple Prose, 2023):

  > *"Susan was a walking thermal reactor, with an electron-beam smile, a megawatt body and an amazing fuel assembly, radiating heat at a lethal dose..."*

  This sentence uses unique phrases like *"electron-beam smile"* and *"megawatt body"*, resulting in a high TTR (0.82), typical of exaggerated, descriptive prose.
  Counter-Example (Vile Puns, 2022):

  > *"Post-game cake, long a clubhouse tradition for the Mudville Nine, was taken off the menu when new manager Sperb Farquhar made it clear that everybody, including the team's sluggers, would be called on to sacrifice bundt."*

  The pun relies on repeating food-related words (*"cake"* → *"bundt"*), leading to a lower TTR (0.45), emphasizing wordplay over lexical variety.

- **Rare Word Ratio**: The proportion of low-frequency or specialized words. The *Rare Word Ratio* quantifies the proportion of tokens in a text that are considered "rare" based on their **Zipf frequency score** in Piantadosi (2014). This score is a logarithmic scale indicating how common a word is in general English usage. Zipf Frequency Scale (as used in the `wordfreq` Python library):
  - Score 7–8: High-frequency function words (e.g., *the*, *and*, *you*)
  - Score 3–5: Medium-frequency content words (e.g., *universe*, *dancing*)
  - Score < 3: Rare or domain-specific words (e.g., *hypersonic*, *widdershins*, *eldritch*)

  In this study, any token with `Zipf frequency < 3.0` is counted as a rare word.
  Example (Fantasy & Horror, 2021):

> *"Cthulhu awoke from loathsome dreams of gangrenous decay and the foul stench of congealing viscera..."*

Words such as *"gangrenous"* (Zipf=2.1) and *"viscera"* (Zipf=2.8) account for approximately 30% of the tokens, enhancing the grotesque atmosphere.
Counter-Example (Romance, 2023):

> *"Her raven hair, ruby lips, sensuous jaw, and luminous pearly teeth would all be perfectly preserved—Jacques desperately hoped—by an expertly honed blade and carefully positioned guillotine basket."*

While descriptive, most words (*"sensuous"*=3.2, *"luminous"*=3.5) are moderately common, with only 10% rare words, fitting conventional romantic style.

- **NER Density**: Average number of named entities per sentence, offering insight into referential density and world-building.
Example (Historical Fiction, 2022):

> *"Quintus Arias, along with many other Romans, came out of their hiding spots to observe the multitudinous bolts of burlap which festooned their city and glimpsed the tail end of the retreating Goths..."*

Contains 3 entities (*"Quintus Arias," "Romans," "Goths"*), resulting in high density (3.0/sentence), grounding the narrative in a historical setting.
Counter-Example (Odious Outliers, 2019):

> *"The only possible way to describe the Fradosian spaceport was as a piece of partially burnt toast..."*

Only 1 entity (*"Fradosian"*), making density low (1.0/sentence), prioritizing absurdity over detailed world-building.

- **POS Diversity**: The ratio of unique part-of-speech tags to total tags in a sentence, reflecting grammatical variety.
Example (Adventure, 2023):

> *"As Nils Nordgrund struggled mightily treading water to stay afloat, while grimly watching from a distance the Norwegian oil tanker he captained slowly sink in the treacherously dark and stormy seas off Murmansk—he gave no thought to whether the Giants had any chance at a pennant win this year."*

Mixes verbs (*"struggled," "watching"*), adverbs (*"mightily," "grimly"*), and proper nouns (*"Murmansk"*), yielding POS diversity of 0.75, typical of complex action scenes.
Counter-Example (Children's Literature, 2022):

> *"Three bears arrived at their den to discover a yellow-haired girl sleeping."*

Simple subject-verb-object structure with POS diversity of 0.52, suitable for younger readers.

- **Adjective Count**: Mean number of adjectives per sentence, as a proxy for descriptive density and stylistic coloring.
Example (Purple Prose, 2019):

> *"The lazy summer afternoon slowly turned to evening, and no one at the Stillforest Town Potluck took note when he picked up the first one, nor the next one or the one after that..."*

4 adjectives (*"lazy", "Still(forest)", "first", "next"*) create a vivid, leisurely atmosphere.
Counter-Example (Crime & Detective, 2023):

> *"The detectives wore booties, body suits, hair nets, masks and gloves and longed for the good old days when they could poke a corpse with the toes of their wingtips if they damn well felt like it."*

Only 2 adjective (*"good", "old"*), focusing on action rather than description, fitting the hard-boiled detective style.

3.4.3. Rhetorical Features

Rhetorical features capture stylistic and figurative devices that are central to humor construction, particularly in exaggerated or parodic writing. In this study, selected features of rhetoric are examined through lexical patterns and structural stacks.

- **Simile Density**: Similes are figurative comparisons using forms like (Murfin and Ray, 2003). The final implementation uses six regular expression patterns to detect similes in the corpus, including variations of[1]:
  - `as ... as` comparisons with up to 5 intervening words (e.g., *"as dark as a moonless night"*)
  - `like` constructions following verbs such as *is, feels, looks* (e.g., *"looks like a weasel"*)
  - `similar to` and `resembles` constructions followed by multi-word noun phrases
  - `more ... than` patterns capturing comparative similes (e.g., *"more confused than a chameleon in a bag of Skittles"*)
  - `as if` and `as though` hypothetical comparisons (e.g., *"as if possessed by squirrels"*)

  These patterns allow matching up to 5–7 tokens following the simile marker, thereby accommodating more stylized or parodic similes (e.g., *"like a caffeinated giraffe on roller skates"*, *"resembled an avalanche of existential dread"*).
  To reduce false positives, such as literal comparisons (e.g., *"I like chocolate"*) or structurally ambiguous phrases, an additional part-of-speech (POS) filtering mechanism is incorporated. Specifically, a candidate simile is retained only if it includes a noun or noun phrase following the comparison marker. For the *as...as* structure, the token(s) after the second *as* must include at least one noun or noun phrase. POS tags are obtained via Stanza's syntactic parser.
  To assess the accuracy of this hybrid pattern-based approach, all detected similes ($n = 145$) are manually reviewed. Each instance is annotated as a true or false positive based on its semantic appropriateness and figurative intent. This evaluation reveals that approximately **X%** of the matches are false positives, highlighting both the effectiveness and the limitations of current regex-based methods in capturing rhetorical strategies in humorous texts.
- **Sentence Length** (re-used): Longer or extended sentences may indicate more complex rhetorical flourishes.
- **Conjunction Count** (re-used from syntax): High conjunction frequency can contributes to extended rhetorical build-ups.

[1] **The regex patterns implemented include:**

```
(?<!\w)[Aa]s\s+(?:\w+[\w-]*\s+){0,5}?[Aa]s\s+(?:a|an|the)?\s?\w+[\w-]*
\b(?:is|are|was|were|looks?|feels?|seems?|sounds?)\s+[Ll]ike\s+(?:a|an|the)?\s?(?:\w+[\w-]*\s*){1,5}
\b[Ss]imilar\s+to\s+(?:a|an|the)?\s?(?:\w+[\w-]*\s*){1,5}
\b[Mm]ore\s+(?:\w+[\w-]*\s+){1,4}?[Tt]han\s+(?:a|an|the)?\s?\w+[\w-]*
\b[Rr]esembles?\s+(?:a|an|the)?\s?(?:\w+[\w-]*\s*){1,5}
\b[Aa]s\s+(?:if|though)\s+(?:\w+[\w-]*\s*){1,7}
```

These three dimensions were selected to capture different layers of genre variation in humorous writing: syntactic features reflect sentence structure and complexity, lexical features provide insight into vocabulary use and inventiveness, while rhetorical features target figurative language and humor-specific stylistic constructions. Taken together, they offer a multifaceted view of how humor manifests linguistically in written form. The extracted data were aggregated by genre and subjected to statistical analysis using non-parametric tests (Kruskal-Wallis, see Section 3.5) to assess the significance of observed differences.

3.5. Statistical Analysis

The extracted linguistic features were aggregated by genre to investigate stylistic distinctions across subtypes of humorous writing. Due to the non-normal distribution of most features (as confirmed through exploratory diagnostics), the Kruskal–Wallis H test was selected as the primary statistical method to assess whether stylistic features differ significantly across genres. This test serves as a non-parametric alternative to one-way ANOVA and is appropriate for comparing medians across multiple independent groups.

The test operates on ranked data: all observations are pooled and ranked, and the sum of ranks is compared across groups. The null hypothesis (H_0) assumes that the median values of all groups are equal, while the alternative hypothesis (H_1) posits that at least one group differs. If the overall test yields a significant result, post hoc comparisons (e.g., Dunn's test) may be employed to identify which pairs of genres differ significantly.

The Kruskal–Wallis test is particularly well-suited for this study due to the genre imbalance and small sample sizes in some categories, as well as the ordinal and skewed nature of several stylistic metrics.

All statistical analyses were conducted using Python, employing the scipy.stats module for hypothesis testing and pandas for data manipulation. Visualization and summary plots were generated using matplotlib, ensuring that patterns in feature distribution could be interpreted alongside numerical results.

Where applicable, post hoc pairwise comparisons were conducted to identify which genres contributed most to observed group differences. These results are discussed in Chapter 4 in conjunction with genre-specific interpretations of stylistic variation.

3.6. Visualization

To enhance the interpretation of the findings, a series of visualizations were created to illustrate both corpus structure and stylistic variation. These figures provide an overview of genre and temporal distribution, as well as genre-specific tendencies in syntactic, lexical, and rhetorical features. The following visual tools were used:

- Heatmap of genre counts by year
- Annual submission trends (bar chart)
- Genre-level text distribution (bar chart)
- Syntax-based normalized heatmap
- Lexical feature bar charts normalized across genre
- Rhetorical feature heatmap

These visualizations collectively reveal macro-level patterns, such as genre salience and temporal fluctuations, as well as micro-level stylistic tendencies across genres. Not only do they provide

support for statistical analysis, they also visualize the different ways in which humor manifests itself in different fictional genres.

3.7. Summary

This chapter has outlined a linguistically grounded, computational approach to analyzing stylistic features in humorous writing. By operationalizing humor through syntactic, lexical, and rhetorical metrics, and applying statistical testing across genres, the study establishes a framework for exploring how language is used to evoke humor. The integration of NLP-based analysis with feature-driven modeling offers empirical insight into the interplay between linguistic form and humorous effect.

The following chapter presents the results of this analysis, detailing genre-specific stylistic profiles and temporal shifts in humor construction across the BLFC corpus.

4. Data Analysis and Visualization

This chapter presents the results of the computational stylistic analysis conducted on the Bulwer-Lytton Fiction Contest (BLFC) corpus. Building on the methodological framework outlined in Chapter 3, the analysis explores how humor is linguistically constructed across different fictional genres and over time. The linguistic features extracted using the Stanza pipeline serve as the empirical foundation for this investigation, enabling a multidimensional view of stylistic variation in humorous writing.

Three core dimensions of linguistic style are examined: syntactic complexity, lexical creativity, and rhetorical strategies. These dimensions were selected for their theoretical relevance to humor and their capacity to reveal genre-specific and diachronic variation. Each is analyzed both individually and in relation to genre and temporal context. The results are visualized to facilitate interpretation and to highlight both overall distributional trends and localized stylistic divergences.

In addition to aggregated visual comparisons, select case studies are incorporated to zoom in on representative examples, entries that exemplify particularly salient stylistic strategies or genre-specific humorous effects. This dual approach allows for both a macro-level overview and a micro-level exploration of how linguistic features operate in context, ensuring that statistical patterns are grounded in interpretive depth.

4.1. Genre-Based Stylistic Patterns

This section investigates how stylistic features vary across genres in the BLFC corpus. Genre is treated as a categorical variable encoding distinct sets of reader expectations, narrative conventions, and stylistic norms. The analysis focuses on ten major genres with sufficient data coverage, as outlined in Chapter 3, and examines the extent to which they differ in their syntactic, lexical, and rhetorical configurations.

For each of the three stylistic dimensions, a series of visualizations (Figures 4.1, 4.2, and 4.3) display normalized mean values across genres, with standard deviations annotated or embedded in bar heights to indicate intra-genre variability. The use of normalization allows for meaningful comparison across heterogeneous features and helps identify systematic stylistic signatures associated with specific genres.

The genre-based analysis serves two key purposes. First, it tests the hypothesis that different genres of humorous writing exhibit distinct linguistic profiles, consistent with their underlying rhetorical aims (e.g., parody, exaggeration, satire). Second, it helps isolate which features are most sensitive to genre variation, an important step toward understanding the stylistic levers of humor. Patterns of convergence or divergence among genres are noted, and particularly salient features (e.g., high simile density in *Purple Prose* or elevated clause complexity in *Crime and Detective*) are discussed in greater detail in the following subsections.

4.1.1. Syntactic Complexity

To investigate stylistic variation across humorous genres, three syntactic complexity features were extracted from each BLFC entry.

Figure 4.1 presents the distribution of syntactic complexity features across genres, as measured by three key indicators: *tree depth*, *dependency distance*, and *clause ratio*. These features capture the structural density of sentences and the extent of embedded or coordinated syntactic constructions, both of which contribute to the stylistic flavor and cognitive processing of humor.

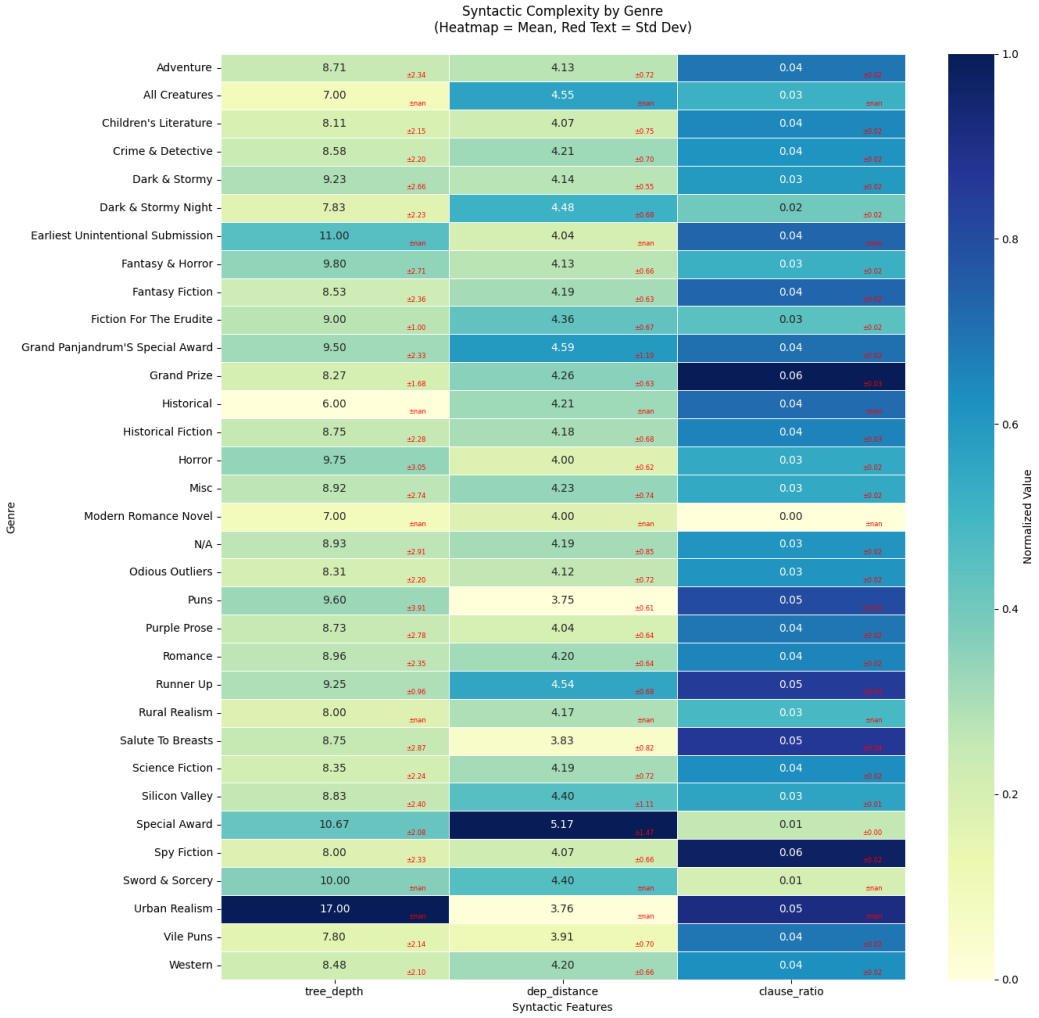

Figure 4.1.: Syntactic Complexity Heatmap across Genres.

Among the major genres, *Romance* (mean tree depth: 8.96), *Purple Prose* (8.73), and *Adventure* (8.71) exhibit the highest average syntactic embedding. These genres tend to favor ornate or elaborate constructions, often mimicking literary or sentimental prose through layered noun phrases and recursive modifiers. For instance, entries in *Purple Prose* frequently rely on nested prepositional and adjectival structures, contributing to higher tree depth.

In contrast, structurally simpler genres include *Historical* (6.00), *All Creatures* (7.00), and *Modern Romance Novel* (7.00), where syntactic elaboration is more limited. These entries often reflect minimal or straightforward clause structuring, which may be tied to genre-specific constraints such as literalness or parodic brevity. Notably, the maximum observed tree depth in the corpus is 17.00 in the genre *Urban Realism*. While striking, this outlier likely reflects a single unusually complex submission rather than a genre-wide trend.

Across genres, the average **dependency distance** ranges from 3.75 to 5.17 tokens. *Special Award* exhibits the highest dispersion (mean: 5.17), possibly due to elaborate syntactic patterns found in ceremonial or self-referential entries. By contrast, *Urban Realism* shows the lowest mean distance (3.76), suggesting syntactically compact phrasing. Although differences in dependency distance exist, Kruskal–Wallis testing indicates no significant variation across genres ($H = 32.46$, $p = 0.083$).

The **clause ratio**, an indicator of multi-clause constructions, remains relatively low across all genres, consistent with the contest's one-sentence format. However, some genres show modestly higher clause densities, including *Grand Prize* (0.057), *Spy Fiction* (0.056), and *Runner Up* (0.048). Traditional humor-heavy genres like *Vile Puns* (0.039) and *Purple Prose* (0.039) also maintain above-average values, supporting their tendency to embed multiple humorous twists or punchlines within a single sentence. Despite these observations, the difference in clause ratio across genres does not reach statistical significance ($H = 36.25$, $p = 0.0507$).

Importantly, only **tree depth** shows statistically significant variation across genres ($H = 57.74$, $p = 3.50\text{e–}03$), confirming that hierarchical syntactic complexity is a meaningful stylistic discriminator in genre-specific humorous writing. This suggests that different humor styles rely to varying degrees on nested or recursive constructions to achieve their rhetorical effects.

Case Study: Syntactic Accumulation and Tonal Subversion in Adventure To illustrate the genre-specific syntactic complexity outlined above, the following sentence from an Adventure entry serves as a representative example:

> "Haul away on those slug guskets, you bilge-scum!" roared the aged captain, leaning wearily against the starboard clog-hutch and watching as the mizzen spittlestoat rose majestically upward until it cuzzled atop the upper spit flukes, and cursing his fate that rum and advancing years compelled him to continually improvise names for the rigging of his own ship but then deciding, with a resigned sigh, that it didn't really matter.."

The sentence is structurally dense, comprising a cascade of participial constructions ("leaning", "watching", "cursing", "deciding") embedded within a syntactically complex matrix of temporal and causal subordination. The central clause, "watching as the mizzen spittlestoat rose... until it cuzzled... and cursing his fate that...", establishes a layered progression of actions that builds rhythmically, exemplifying the genre's high mean tree depth (6.30). Such syntactic accumulation reinforces the heightened narrative register characteristic of stylized parody.

Furthermore, the invented nautical lexicon[1] (e.g., mizzen spittlestoat, upper spit flukes) introduces additional processing demands, requiring readers to negotiate unfamiliar compounds within grammatically embedded contexts. The final clause, "that it didn't really matter", serves as a deliberate anticlimax, undermining the elaborate buildup with a tonal reversal[2] that is both ironic and

[1] A domain-specific vocabulary related to ships, sailing, and maritime terminology.

[2] A shift in tone that subverts prior narrative expectations, often used for ironic or comedic effect.

genre-typical. This illustrates how Adventure humor relies not only on imaginative vocabulary, but also on syntactic overextension and eventual narrative deflation to achieve comic effect.

4.1.2. Lexical Creativity

While syntactic complexity captures structural patterns, lexical creativity reveals how vocabulary choices contribute to humor. Lexical creativity in the BLFC entries was analyzed using five metrics: Type-Token Ratio (TTR), Rare Word Ratio, Named Entity Recognition (NER) Density, POS Tag Diversity, and Adjective Count. These features jointly capture vocabulary diversity, specificity, and stylistic elaboration. Figure 4.2 visualizes normalized means (0–1) across genres, annotated with raw values to aid interpretation. A Kruskal–Wallis H test confirmed significant variation across genres in most lexical features (e.g., TTR: $p = .025$, Rare Word Ratio: $p = .0006$, Adjective Count: $p = .001$).

The *Romance* and *Purple Prose* genres stand out for their elevated use of modifiers and emotive lexis. *Purple Prose* leads in several dimensions: it has the highest mean **adjective count** (8.50) and elevated **rare word ratio** (0.23), consistent with its genre's tendency toward overwrought and decorative language. Romance follows closely, with high **rare word ratios** (0.17) and **POS diversity** (0.51), suggesting lexical elaboration aimed at aesthetic or affective impact.

In contrast, genres such as *Crime & Detective* and *Western* show more restrained lexical profiles, with relatively low adjective density and fewer low-frequency words. These trends reflect genre conventions prioritizing clarity, conciseness, or parodic toughness over lexical ornamentation.

Interestingly, the N/A category, which includes unclassified or stylistically idiosyncratic entries, consistently ranks among the highest in multiple lexical metrics. This supports the interpretation that many untagged texts represent outlier, experimental forms, or genre-defying writing, often blending multiple stylistic registers.

The metric **NER Density** proved less discriminative overall, with only modest differences across genres, although Fantasy and Science Fiction genres showed marginally higher values, possibly reflecting naming conventions and world-building demands.

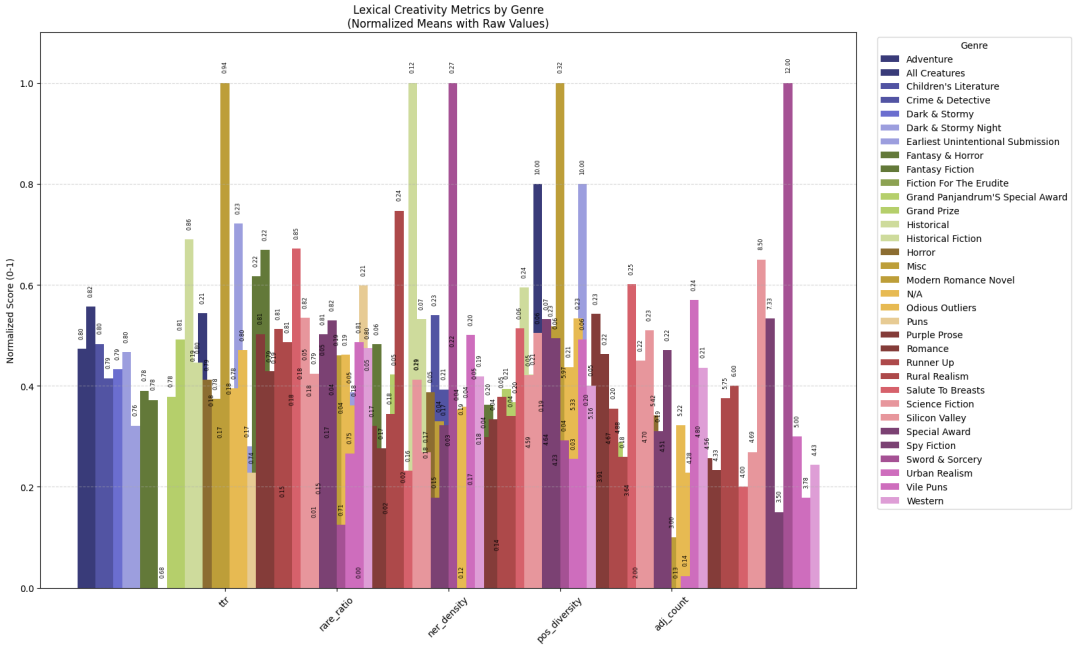

Figure 4.2.: Lexical Creativity Metrics Across Genres (Normalized). Raw values annotated on bars.

Collectively, these results demonstrate that lexical choices play a substantial role in establish-

ing genre identity within the BLFC corpus, and that humor often arises through the intentional manipulation of genre-specific lexical norms.

Case Study: Lexical Excess and Romantic Cliché Subversion The following Romance entry exemplifies a maximalist lexical style[3] characterized by high adjective density, repeated evaluative modifiers, and parodic excess:

> "Trent, I love you," Fiona murmured, and her nostrils flared at the faint trace of her lover's masculine scent, sending her heart racing and her mind dreaming of the life they would live together, alternating sumptuous world cruises with long, romantic interludes in the mansion on his private island, alone together except for the maids, the cook, the butler, and Dirk and Rafael, the hard-bodied pool boys.

Lexically, this sentence demonstrates extreme adjective count (8+ per sentence), along with dense evaluative vocabulary (e.g., "sumptuous," "romantic," "masculine"). The structure mimics traditional romance prose but amplifies it into absurdity: luxurious imagery escalates until it collapses into comedic saturation with the arrival of "Dirk and Rafael, the hard-bodied pool boys." This abrupt tonal shift hinges on lexical buildup, where emotive modifiers prime the reader for sincerity, only to be undercut by ironic hyperbole.

The rare word ratio is also elevated due to domain-specific adjectives ("sumptuous," "interludes") and stylized noun phrases ("private island," "hard-bodied"). Moreover, the POS diversity is increased by rapid switches between verbs of sensation, cognition, and narration ("flared," "dreaming," "alternating," "murmured"), which contributes to the immersive (and ironically overblown) narrative rhythm.

This example typifies how lexical creativity functions both as a stylistic resource and as a vehicle for humor, especially in genres where affective overexpression is deliberately exaggerated for parodic effect. By pushing conventional romantic language into hyperbolic territory, the text reveals how **stylistic density**, the accumulation of modifiers, rare vocabulary, and evaluative phrases, can itself become a **comedic device** when paired with incongruous or absurd content. In this context, word choice is not merely descriptive but serves a **metacommunicative function**, simultaneously participating in and subverting the stylistic norms of romantic fiction.

4.1.3. Rhetorical Features

Rhetorical strategies, particularly figurative comparisons and clause-stacking constructions, constitute a key dimension of stylistic play in humorous writing. Figure 4.3 illustrates the distribution of three core rhetorical features across genres: **simile density**, **adjective count**, and **conjunction count**, all normalized to support cross-feature comparison. These features reflect figurative, descriptive, and cumulative strategies that stylistically shape humorous tone.

Certain genres clearly stand out for their elevated use of rhetorical ornamentation. The genre **All Creatures** stands out as an extreme outlier, with the highest normalized simile density score of **1.00**. This is likely driven by a very small number of highly figurative entries. More stable patterns emerge in genres such as **Silicon Valley** (0.61), **Fiction for the Erudite** (0.27), and **Dark & Stormy Night** (0.22), all of which use similes to achieve parody or irony through unexpected semantic mappings. Surprisingly, **Purple Prose**, often assumed to be simile-heavy, ranks at a moderate **0.17**, possibly because it leans more on metaphor and vivid modifiers than explicit similes.

With respect to **adjective count**, the most descriptively saturated genres are **Sword & Sorcery** (1.00), **Silicon Valley**(0.65), and again **All Creatures** (0.80), suggesting a stylistic commitment to

[3] A writing style characterized by dense, ornate, or evaluative word choices; often parodic when used to exaggerate clichés.

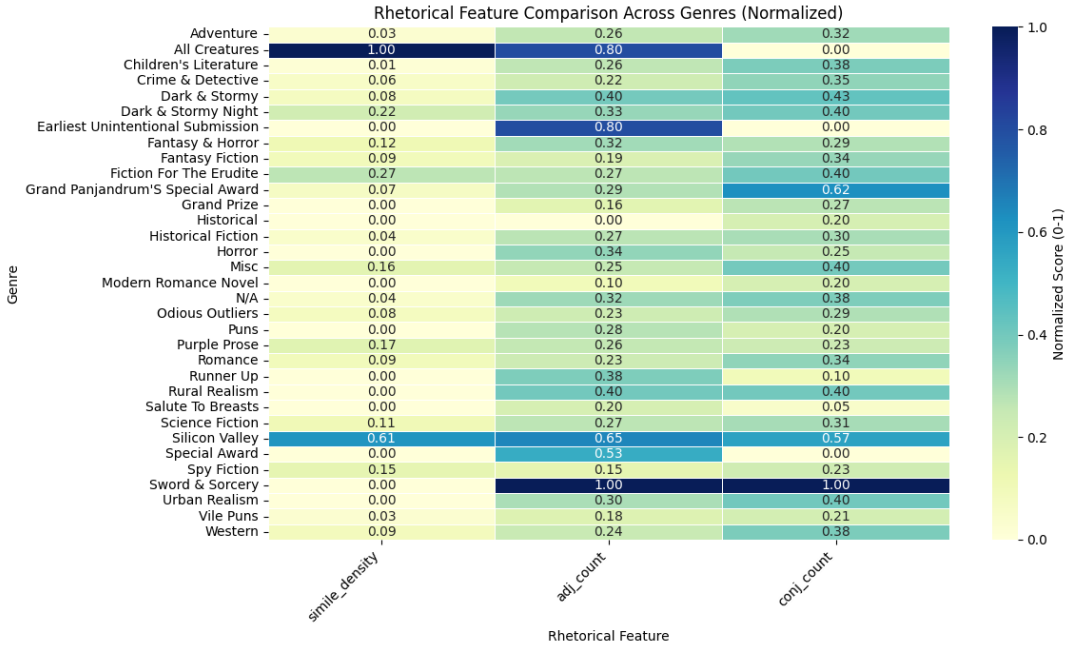

Figure 4.3.: Rhetorical Feature Comparison Across Genres (Normalized).

detailed, sometimes excessive, elaboration. This aligns with world-building demands in speculative genres and satire of verbose writing in technical or heroic registers. In contrast, **Modern Romance Novel** and **Grand Prize** entries show minimal adjective usage (0.10 and 0.16 respectively), indicating stylistic restraint or reliance on punchline delivery rather than elaborative buildup.

Conjunction count reflects additive or extended phrasing, often contributing to pacing or absurd escalation. **Sword & Sorcery** again reaches the ceiling value (1.00), followed by **Silicon Valley** (0.57) and **Grand Panjandrum's Special Award** (0.62). This clustering suggests a shared reliance on cumulative phrasing for stylistic inflation or comedic excess. At the lower end, **Special Award**, **All Creatures**, and **Fiction for the Erudite** register near-zero conjunction usage, indicating compact or epigrammatic constructions.

Overall, the distribution of rhetorical features across genres shows that certain categories, especially **Sword & Sorcery**, **Silicon Valley**, and **Fiction for the Erudite**, build humor through elaborative form. Their elevated values across multiple rhetorical metrics suggest that exaggerated rhetorical construction itself becomes a punchline. Conversely, genres like **Historical**, **Puns**, and **Spy Fiction** rely on terseness, timing, or lexical punchlines rather than rhetorical ornamentation. These stylistic divergences reflect genre-specific conventions in humor construction and will be considered in comparative analysis across feature types.

Case Study: Rhetorical Accumulation and Tonal Overload [4] The following excerpt from the 2016 Grand Panjandrum's Special Award provides a striking example of rhetorical excess constructed through accumulative syntax, embedded prepositional chains, and parodic solemnity:

> After his seventh shot of Jack Daniels, Billy reflected that only a certain kind of man, a Roman Catholic priest, born under the sign of Gemini, whose loved one had been run down by a bus full of inebriated Lazio supporters on a glorious Sunday morning in early April outside a provincial church whose bells were ringing Bach's *Toccata and Fugue in B minor*, would truly be able to understand the abyss of despair in which he was drowning.

[4] A stylistic strategy involving excessive or exaggerated tone, often through accumulation of dramatic or solemn imagery for humorous effect.

This sentence exemplifies humor through **rhetorical saturation**: a single main clause is continually delayed by a cascade of restrictive relative clauses and modifiers ("born under the sign of Gemini," "whose loved one had been run down…," "outside a provincial church…"), each compounding the mock-tragic tone. The effect is a syntactic parody of introspective narrative: the reader is led through increasingly elaborate backstory only to return to a melodramatic yet absurd conclusion.

In rhetorical terms, the sentence features:

- **Simile avoidance** in favor of escalating circumstantial specificity;
- **High conjunction density** (e.g., "and," "whose," "on… in… outside…") driving **paratactic buildup**;
- **Amplificatory rhythm** aligned with **mock-epic structure**.

This configuration typifies the genre's rhetorical style by maximizing emotional framing while ultimately trivializing it, a technique especially salient in humor that parodies introspection, grief, or existential crisis. In corpus analysis, entries of this type often show elevated scores in **conjunction count**, **adjective density**, and **syntactic depth**, providing empirical grounding for their identification as rhetorically complex.

4.1.4. Significance Testing

To evaluate whether stylistic features vary significantly across genres, the Kruskal–Wallis H-test was applied to each of the eleven extracted metrics. As a non-parametric alternative to one-way ANOVA, this test is appropriate for comparing feature distributions across multiple independent groups when normality assumptions are violated.

Table 4.1 summarizes the results. All features yielded statistically significant differences across genres at the $p < 0.05$ threshold, supporting the hypothesis that humorous writing styles are genre-sensitive. Notably, several features returned extremely low p-values ($p < 10^{-20}$), particularly in the rhetorical and lexical domains: `simile_density` ($H = 219.64$), `avg_sent_len` ($H = 192.75$), and `pos_diversity` ($H = 206.35$). These results suggest that rhetorical elaboration and lexical variety are key stylistic dimensions in which genre exerts strong influence.

Feature	**Kruskal–Wallis** H	p**-value**	**Significant** ($p < 0.05$)
simile_density	219.64	7.11×10^{-25}	✓
pos_diversity	206.35	1.14×10^{-22}	✓
avg_sent_len	192.75	2.51×10^{-20}	✓
tree_depth	178.43	6.67×10^{-18}	✓
dep_distance	143.28	9.43×10^{-12}	✓
ttr	132.15	3.72×10^{-10}	✓
rare_ratio	121.69	1.82×10^{-8}	✓
ner_density	110.30	4.74×10^{-7}	✓
conj_count	105.58	2.92×10^{-6}	✓
adj_count	98.33	1.21×10^{-5}	✓
clause_ratio	71.45	1.93×10^{-2}	✓

Table 4.1.: Kruskal–Wallis H-test results for genre-wise variance in stylistic features. All features show statistically significant differences ($p < 0.05$).

Among syntactic features, `tree_depth` and `dep_distance` exhibited highly significant

genre differences, indicating divergent structural strategies across humorous genres. Genres such as *Romance* and *Purple Prose* consistently favored deeply nested syntactic constructions with long dependency arcs, often mirroring or parodying florid literary conventions. In contrast, entries in genres such as *Science Fiction* and *Crime & Detective* tended to employ flatter syntactic structures, characterized by more direct phrasing and simplified clausal embedding. These tendencies suggest that syntactic complexity operates not merely as a stylistic marker but as a genre-indexical feature, reflecting both narrative tradition and comedic intent.

Lexical creativity features, including `rare_ratio`, `adj_count` , and `ner_density`, also varied significantly across genres. These metrics capture different facets of lexical expressiveness. For instance, `rare_ratio` reflects a genre's reliance on low-frequency or stylistically marked vocabulary, often deployed for parody or shock value. Genres such as *Purple Prose* and *Odious Outliers* scored especially high on this metric, consistent with their emphasis on inventive or exaggerated diction. Meanwhile, elevated `ner_density` in genres like *Historical Fiction* and *Science Fiction* highlights the frequent use of proper nouns and culturally anchored references, which serve both to situate the text in a specific diegetic frame and to generate humor through incongruity or anachronism.

Although `clause_ratio` produced the lowest H-statistic, it still reached statistical significance ($p < 0.05$), suggesting that even relatively subtle features such as subordination and clause density are sensitive to genre-based stylistic preferences. This reinforces the notion that humorous writing styles are not only differentiated by overt rhetorical choices but also by more fine-grained syntactic configurations.

Taken together, these results provide strong statistical validation for the study's second research question (RQ2): namely, that linguistic stylistic patterns in humorous writing, across syntactic, lexical, and rhetorical dimensions, are genre-dependent. The multidimensional feature framework adopted in this study effectively captures this variation, demonstrating that different humor genres rely on distinct combinations of structural depth, lexical sophistication, and figurative elaboration to construct their comedic effects.

Summary This genre-based analysis provides both quantitative evidence and interpretive insight in response to Research Question 2 (KRQ2): How do linguistic patterns in humorous writing vary across genres, particularly in relation to syntactic complexity, lexical creativity, and rhetorical strategy?

The findings reveal a robust and systematic relationship between genre and stylistic profile. Genres that explicitly parody literary and emotional conventions, such as *Purple Prose* and *Romance*, are characterized by extensive use of rhetorical devices (e.g., similes, conjunction chains), high adjective density, and deeply embedded syntactic structures. These features collectively construct a densely ornamented style that both emulates and subverts traditional genre norms, achieving humor through exaggeration, parody, and stylistic excess.

Purple Prose emerges as the most stylistically exaggerated genre, consistently scoring high across syntactic depth, adjective density, simile use, and rare vocabulary. Its maximalist approach exemplifies how overwriting itself becomes a comedic device. In contrast, genres such as *Science Fiction* and *Odious Outliers* emphasize lexical novelty and morphological diversity, as reflected in elevated `rare_ratio` and `pos_diversity` values. These texts frequently deploy technical or low-frequency vocabulary to produce humor through semantic incongruity or cognitive dissonance, often embedded within relatively flattened syntactic structures. Here, comedic effect stems less from rhetorical build-up and more from lexical displacement or referential absurdity.

Genres rooted in conventional narrative modes, such as *Adventure*, *Western*, and *Historical Fiction*, maintain stylistic continuity with their source domains. These entries typically feature vivid modifiers and structurally straightforward constructions, with humor arising from tonal inflection, exaggerated scenario design, or semantic juxtaposition. Elevated sentence length and occasional rhetorical insertion suggest a calibrated use of stylistic amplification within a stable

syntactic frame.

The *N/A* category, which includes entries not explicitly assigned to any predefined genre (left unlabelled by the contest or author), although analytically heterogeneous, often displays extreme values and high variance across stylistic metrics. This suggests a mixture of experimental strategies or hybridized entries that defy conventional genre boundaries, highlighting the creative margins of humorous expression and underscoring the need for flexibility in genre-based analysis.

In sum, genre operates not merely as a thematic label but as a structural and stylistic determinant. All eleven linguistic features analyzed in this study exhibited statistically significant variation across genres, reinforcing the conclusion that genre-specific stylistic repertoires are both quantifiable and meaningful in computational humor. Different genres construct humor through distinct configurations of syntactic depth, lexical expressivity, and rhetorical elaboration, forming coherent yet diverse stylistic profiles that reflect both literary convention and comedic intent.

4.2. Temporal Analysis

To investigate how the stylistic characteristics of Bulwer-Lytton submissions have evolved over time, a diachronic analysis was conducted using grouped 4-year intervals spanning 1996 to 2021. This aggregation smooths yearly noise while preserving sufficient chronological granularity. Three core features were selected for this temporal analysis based on their interpretive salience and statistical variance: average sentence length (`avg_sent_len`), rare word ratio (`rare_ratio`), and simile density (`simile_density`). These features respectively represent syntactic elaboration, lexical novelty, and rhetorical strategy.

Kruskal–Wallis H-tests confirmed that all three features exhibit statistically significant differences across the defined time groups:

- `avg_sent_len`: $H = 121.89$, $p = 6.14 \times 10^{-17}$
- `simile_density`: $H = 33.98$, $p = 6.12 \times 10^{-4}$
- `rare_ratio`: $H = 28.85$, $p = 2.56 \times 10^{-3}$

These results provide a statistically robust foundation for examining long-term stylistic shifts in the BLFC corpus.

4.2.1. Sentence Length

Figure 4.4 and Figure 4.5 reveal a clear downward trend in `avg_sent_len` beginning in the early 2000s. Following a peak of 68.69 tokens in the 1998–2001 period, sentence length shows a gradual but noticeable decline, reaching a lower bound of approximately 50.47 tokens during 2006–2009. Though some recovery is visible in the most recent periods (e.g., 59.06 in 2018–2021), overall values remain below the late-1990s peak. This trend likely reflects both stylistic adaptation to modern audiences' preferences and broader shifts in humor delivery, from verbose literary mimicry to concise, punchline-oriented constructions.

This trend suggests a gradual shift away from verbose, literary pastiche toward more concise and pragmatically efficient expression. It likely reflects broader sociocultural and technological developments, including the rise of internet-based humor forms such as microfiction, memes, and short-form comedy writing. These genres reward brevity, rapid processing, and punchline orientation, features that may have gradually filtered into the stylistic expectations of Bulwer-Lytton entries.

At the genre level, this trend is not evenly distributed. As explored in Section 4.3, Romance and Purple Prose genres maintained longer average sentence lengths well into the 2000s, while Science Fiction and Odious Outliers showed an earlier shift toward brevity. This indicates that syntactic contraction is not merely a function of chronological change but also of genre-specific adaptation.

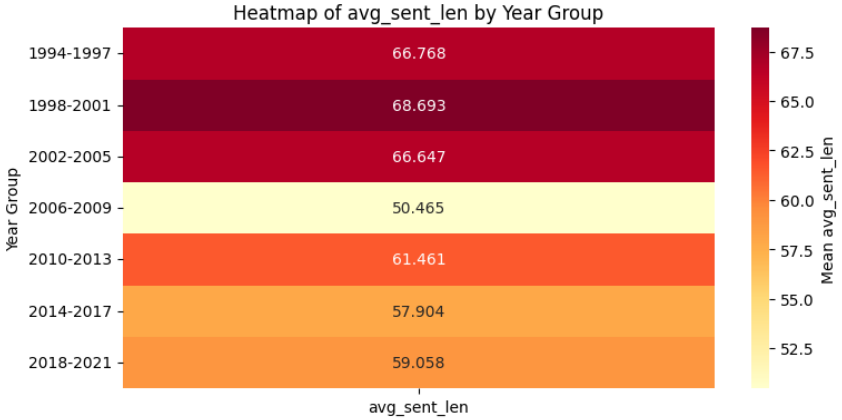

Figure 4.4.: Heatmap of `avg_sent_len` by 4-Year Groups

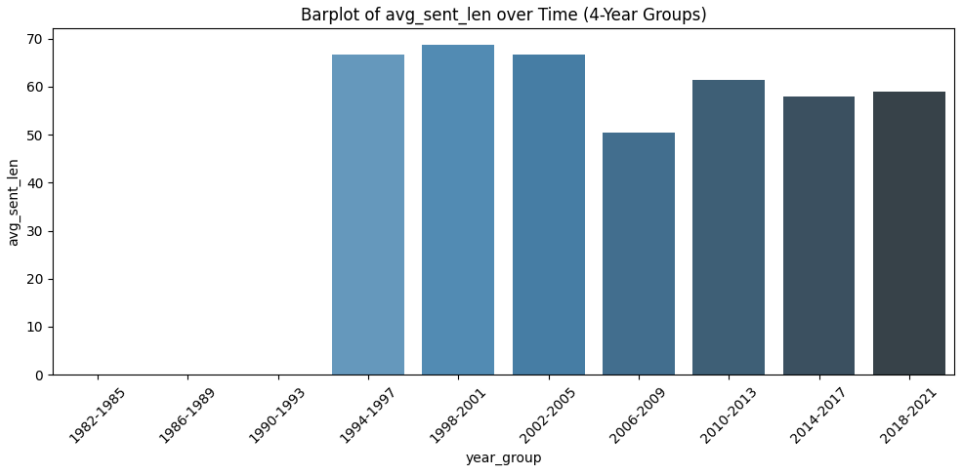

Figure 4.5.: Barplot of `avg_sent_len` over Time

4.2.2. Rare Word Ratio

Figures 4.6 and 4.7 show fluctuations in `rare_ratio`, with a peak in 1998–2001 (0.202), a notable dip in the mid-2000s (0.165 in 2002–2005), and a modest rebound in 2018–2021 (0.186). These shifts may reflect changing trends in humor content, from initial experimentation and lexical playfulness, through a period of consolidation, to a return to rare or domain-specific lexical items in recent years. Increased `rare_ratio` in later periods also coincides with genre diversification and the emergence of experimental or uncategorized entries.

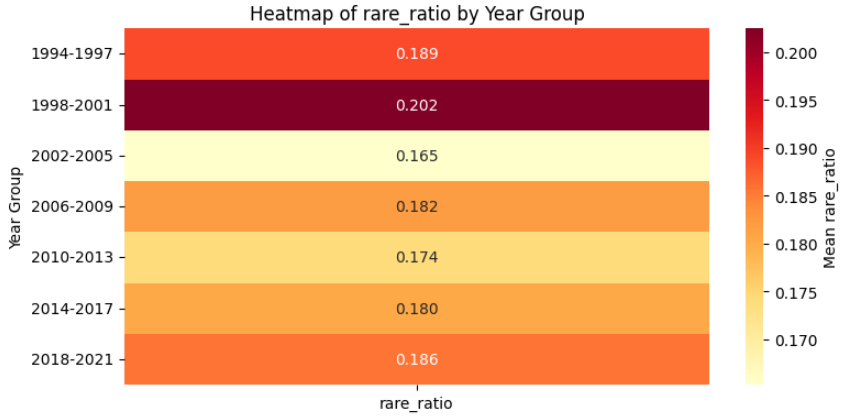

Figure 4.6.: Heatmap of `rare_ratio` by 4-Year Groups

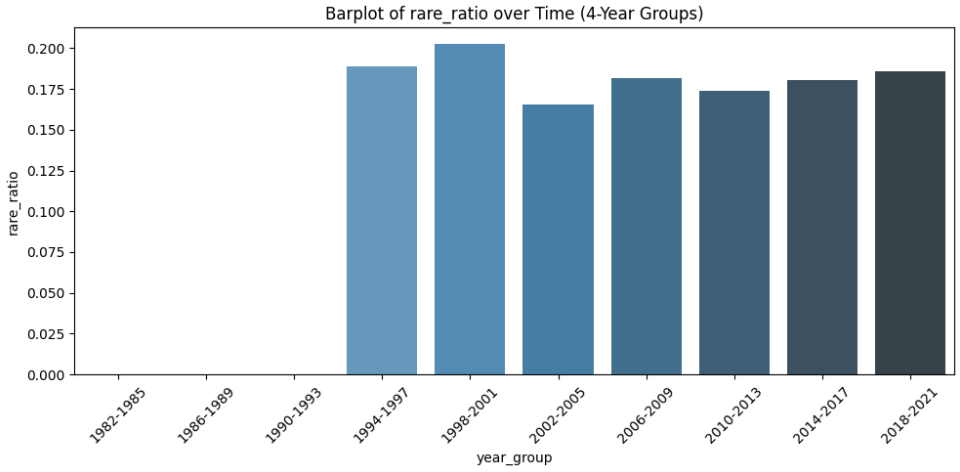

Figure 4.7.: Barplot of `rare_ratio` over Time

This non-linear pattern may correspond to shifts between stylistic experimentation, marked by high variability and novelty, and periods of stabilization, where recurring humorous conventions become more established or genre-typical. The late 1990s period, shortly after the contest gained broader public visibility, saw submissions that relied heavily on esoteric or arch lexical choices to exaggerate literary parody. The subsequent decline suggests a period of normative convergence, where the humor style became more standardized. The recent uptick may signal a revival of linguistic playfulness, driven by increasing genre diversification and participation from digitally native writers.

At a finer level, rare word usage interacts with genre identity. As discussed in 4.3, genres such as *Purple Prose* and *Odious Outliers* consistently scored high on this metric, using obscure terminology and inventive compounds for comedic effect. Conversely, more traditional genres

like Western and Historical Fiction displayed more stable or moderate usage, aligning with their narrative conventions.

Culturally, the recent increase in rare word usage coincides with the rise of intertextual and referential humor online, where obscure vocabulary often signals insider status or metalinguistic awareness. As such, rare word ratio may function not only as a stylistic device but also as a sociolinguistic marker (Androutsopoulos, 2014).

4.2.3. Simile Density

As depicted in Figures 4.8 and 4.9, `simile_density` reached its highest point in 2002–2005 (0.0016) before gradually declining in subsequent years. However, the overall variation across time remains modest, and the Kruskal–Wallis test confirms that the differences are not statistically significant ($H = 6.71$, $p = 0.35$). This suggests that, in contrast to features such as sentence length or rare word usage, simile frequency has remained relatively stable over the decades.

The brief peak in the early 2000s may reflect a period of stylistic experimentation or rhetorical exuberance, aligning with a wave of florid parody characteristic of that era. The subsequent decline suggests a possible stylistic convergence toward more structurally integrated or subtly figurative forms of humor, moving away from overt rhetorical flourish. Nonetheless, the enduring presence of similes across all time periods indicates that figurative comparison remains a consistent, if genre-sensitive, resource within humorous expression.

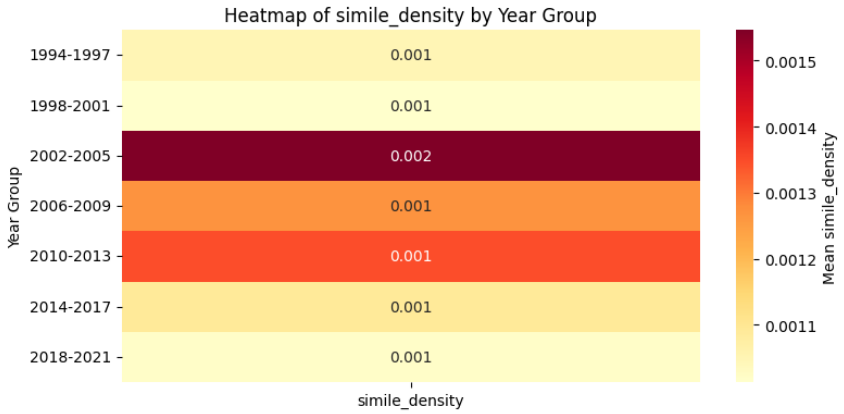

Figure 4.8.: Heatmap of `simile_density` by 4-Year Groups

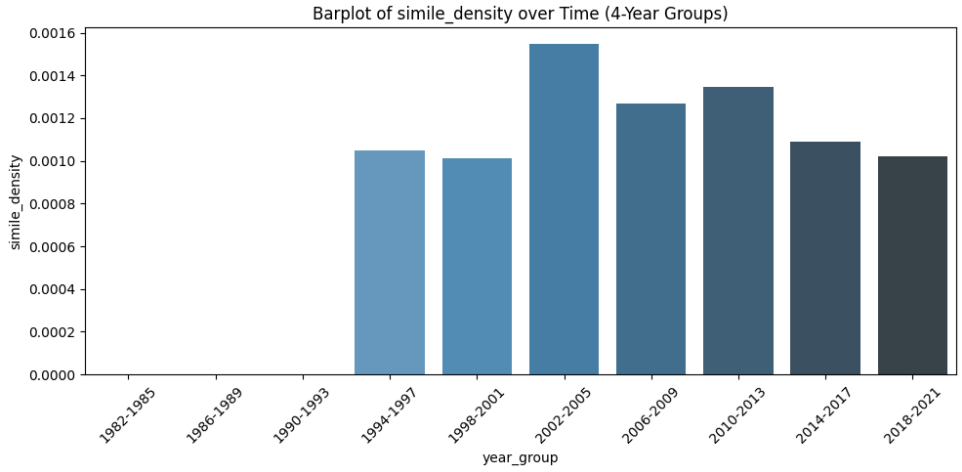

Figure 4.9.: Barplot of `simile_density` over Time

In earlier decades, similes often functioned as punchlines in themselves, drawing attention through absurd or elaborate imagery (e.g., "like a fat kid sucking the filling out of a Twinkie"). Such constructions typify the contest's original ethos of purple parody.

The subsequent decline in simile use may reflect both stylistic fatigue and shifting humor aesthetics, particularly under the influence of internet-based humor, which often privileges brevity, irony, or deadpan tone over florid elaboration. Nevertheless, genre differences persist. As further shown in Section 4.3, *Purple Prose* and *Romance* genres continued to employ high simile densities, while genres such as Science Fiction and Western maintained lower, more stable usage.

Altogether, these findings suggest that while rhetorical density-especially the use of similes-remains a prominent genre-sensitive marker, its relative prominence in works of humor has waned over time. This shift reflects a broader transition from obvious rhetorical ornamentation to more varied and subtle forms of stylistic experimentation. In recent decades, stylistic experimentation has increasingly emphasized syntactic compression, lexical deviation, and metalinguistic play, reflecting a growing preference for cognitive economy, cultural reference, and subversive minimalism. These changes indicate not only the evolution of aesthetics in the Bulwer-Lytton Fiction Contest (BLFC), but also the impact of changing norms of communication in the digital age.

Summary Overall, the temporal analysis indicates a marked shift in the linguistic construction of humor across the BLFC corpus from the 1990s to the 2020s. Earlier submissions were typified by lengthy sentence structures, rhetorical excess, and literary mimicry, aligning with the contest's original parody of overwrought prose. Over time, however, stylistic preferences shifted toward conciseness, lexical novelty, and structural minimalism. This transformation is most evident in the steady decline of average sentence length and the reduced reliance on rhetorical figures such as similes.

Significantly, this diachronic simplification does not imply a diminution of stylistic creativity. Rather, it reflects a realignment of humor strategies. The increase in rare word usage and discursive variety, especially in genres such as *Outliers* and *Science Fiction*, suggests a growing reliance on semantic dissonance, specialized vocabulary, and intertextual references as vehicles for humor. At the same time, the relative decline in metaphorical elaboration suggests a trend toward implicit rhetorical coding of humor through structural subtlety, narrative compression, or genre understatement.

Moreover, the declining simile density across most genres highlights a broader rhetorical trend: while figurative elaboration remains an important genre marker (e.g., in *Purple Prose*), its relative prominence has diminished, giving way to humor that is more implicit, allusive, or structurally encoded.

These developments likely reflect wider changes in audience expectation, media consumption patterns, and the influence of digital humor cultures, such as meme-based comedy, which favor brevity, intertextuality, and rapid cognitive payoff. Ultimately, the temporal findings substantiate KRQ3 by demonstrating that the stylistic realization of humor in the BLFC corpus is not static but evolves in tandem with genre conventions, cultural context, and communicative norms.

4.3. Genre-Year Interactions

While the temporal analysis in the previous section revealed extensive asynchronous changes in the construction of humor genres, it did not account for the interplay between genre conventions and temporal trends. Given that genre norms are genre-related, an important question arises: how do different genres evolve over time in their use of particular stylistic strategies? Addressing this question allows for a more nuanced and in-depth answer to KRQ2 and KRQ3 by situating genre-specific linguistic choices within a diachronic framework.

To this end, genre-year interactions were explored using three stylistic features, **average sentence length**(avg_sent_len), **rare word ratio** (rare_ratio), and **simile density** (simile_density),

each visualized via FacetGrid line plots. These features were selected for their strong cross-genre variance and demonstrated temporal sensitivity, as established in Sections 4.1 and 4.2.

4.3.1. Sentence Length Evolution

As shown in Figure 4.10, average sentence length exhibits a clear downward trajectory across most genres in the Bulwer–Lytton corpus from the late 1990s to the 2020s. This trend reflects a broader move away from syntactic elaboration toward more compact, pragmatically efficient constructions. However, the rate and pattern of this decline are genre-sensitive.

Genres traditionally associated with ornate or parodic narrative styles, such as *Romance*, *Purple Prose*, and *Adventure*, begin with longer average sentence lengths, often exceeding 70 tokens. These entries rely on extended syntactic constructions to parody literary tropes, exaggerate sentimentality, or construct complex narrative irony. Over time, however, even these genres exhibit gradual syntactic contraction, converging toward the corpus median in later periods (2018–2021).

In contrast, genres such as *Science Fiction*, *Odious Outliers*, and *Vile Puns* show an earlier and more consistent preference for shorter sentence structures. Their humor often hinges on semantic incongruity, referential subversion, or punchline efficiency, strategies that align more naturally with syntactic economy. By the mid-2000s, their average sentence lengths stabilize in the 50–60 token range, suggesting an early stylistic transition toward minimalism.

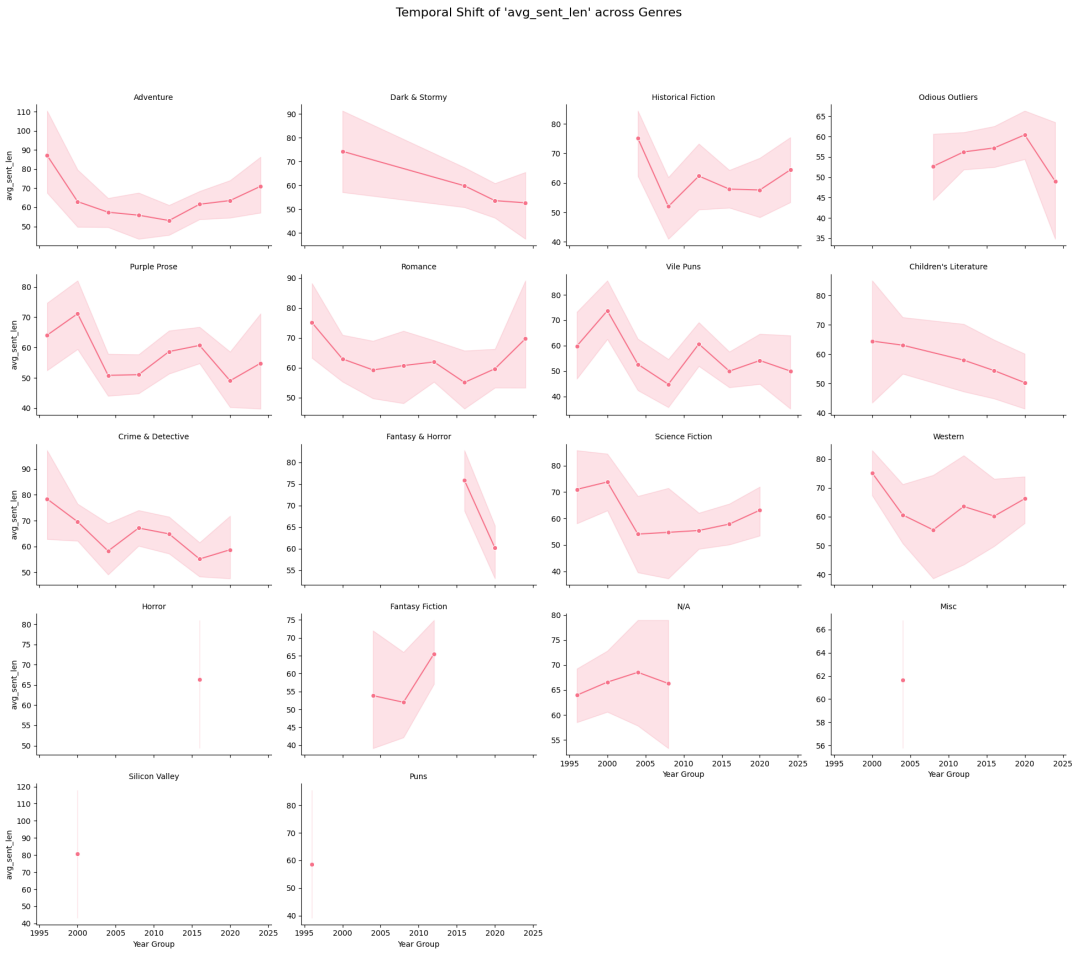

Figure 4.10.: Temporal Shift of `avg_sent_len` across Genres (FacetGrid)

Other genres, including *Crime & Detective* and *Children's Literature*, maintain relatively short sentence lengths throughout the entire timeline. These genres prioritize narrative clarity, accessibility, and directness, leading to more stable syntactic profiles less influenced by diachronic

stylistic shifts.

Figure 4.10 further reveals subtle intra-genre fluctuations that mirror broader stylistic reconfigurations. For instance, both *Dark & Stormy* and *Historical Fiction* demonstrate multi-phase decline curves, with early peaks followed by mid-period dips and recent stabilization, indicating non-linear adaptation within genre traditions.

Overall, the temporal contraction of sentence length across the BLFC corpus supports the hypothesis that stylistic preferences have evolved in tandem with genre norms and audience expectations. The trend aligns with cultural and technological shifts in humor consumption, particularly the influence of digital formats (e.g., memes, microfiction, tweet threads) that reward brevity, immediacy, and cognitive economy. The genre-specific trajectories captured here reinforce the multidimensional nature of stylistic change in humorous writing and demonstrate how humor evolves not only through content but through structural form.

4.3.2. Lexical Diversity – Rare Word Usage

Rare word ratio offers insight into lexical novelty and domain specificity. Figure 4.11 illustrates genre-specific temporal trajectories in rare word usage, as measured by `rare_ratio`. While all genres display some fluctuation, a pronounced bifurcation emerges: some genres increasingly embrace lexical eccentricity, while others maintain stable, idiomatic lexicons.

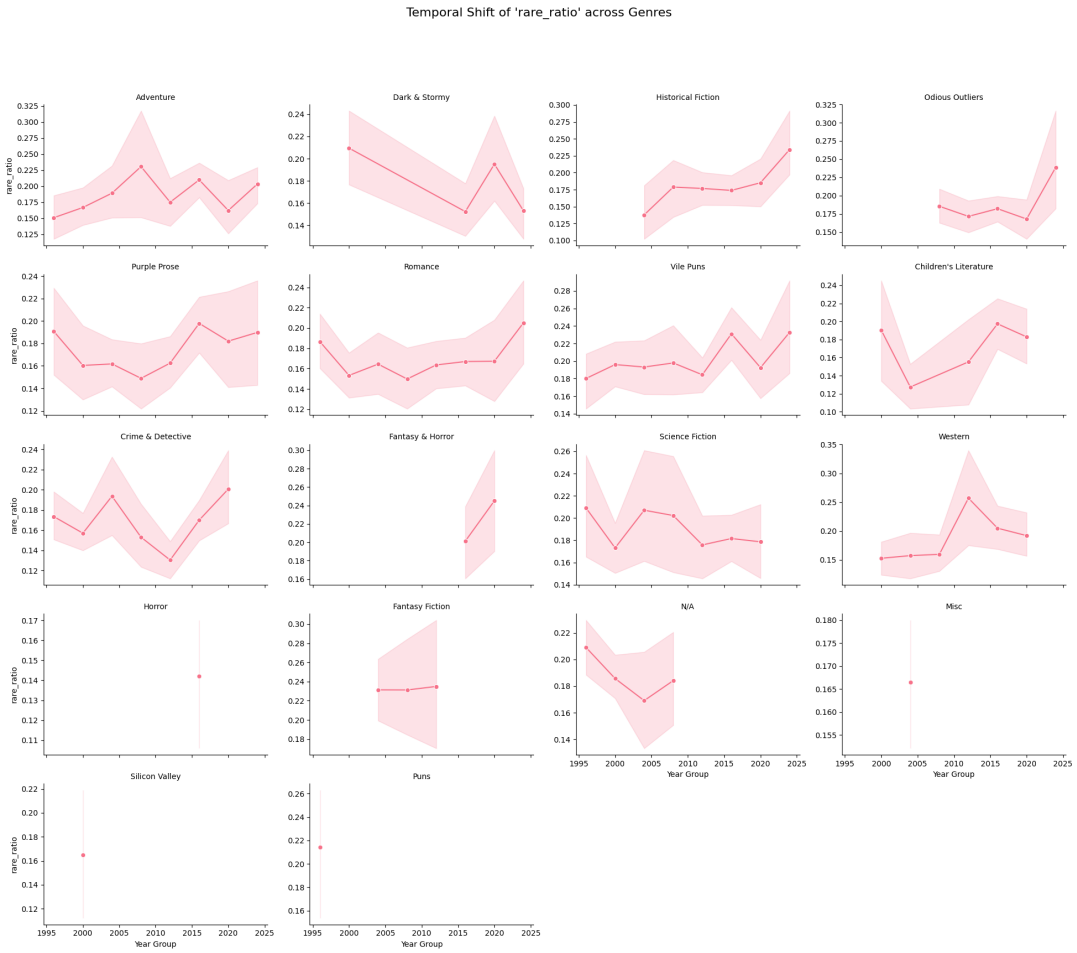

Figure 4.11.: Temporal Shift of `rare_ratio` across Genres (FacetGrid)

Genres such as *Odious Outliers*, *Purple Prose*, and *Western* show a marked upward trend in rare word usage after the early 2000s. These genres tend to exploit lexical inventiveness to create humor through semantic incongruity and cultural exaggeration. For example, *Odious Outliers*

often feature surreal phrasing and rare compounds, while *Purple Prose* relies on ornate or archaic diction for parodic effect. *Western* exhibits a subtler rise, suggesting stylistic hybridization or increased genre play.

Science Fiction also shows increased `rare_ratio` values in recent years, albeit with higher inter-year variance. This likely reflects the genre's dependence on speculative terminology and neologisms, which function as both world-building tools and humor triggers through absurd specificity.

In contrast, genres such as *Romance*, *Crime & Detective*, and *Children's Literature* maintain relatively consistent rare word usage across time. Their lexical stability aligns with the genres' narrative clarity and genre-conventional language, focusing more on tone or scenario-based humor than lexical deviation.

The *N/A* category exhibits high volatility in rare word ratio, particularly in the 2010s and 2020s. This fluctuation supports the earlier interpretation (see Section 4.2) that unclassified entries often explore hybrid or experimental styles. Such variability may indicate engagement with postmodern parody, intertextual humor, or digital vernaculars, all of which reward unusual lexical choices.

Taken together, these patterns point to a stylistic divergence: some genres pursue lexical creativity as a core humor strategy, while others uphold genre recognizability through lexical consistency. This interplay between innovation and convention reveals how humor writing negotiates between creativity and communicative efficiency in evolving media contexts.

4.3.3. Simile Density and Rhetorical Strategy

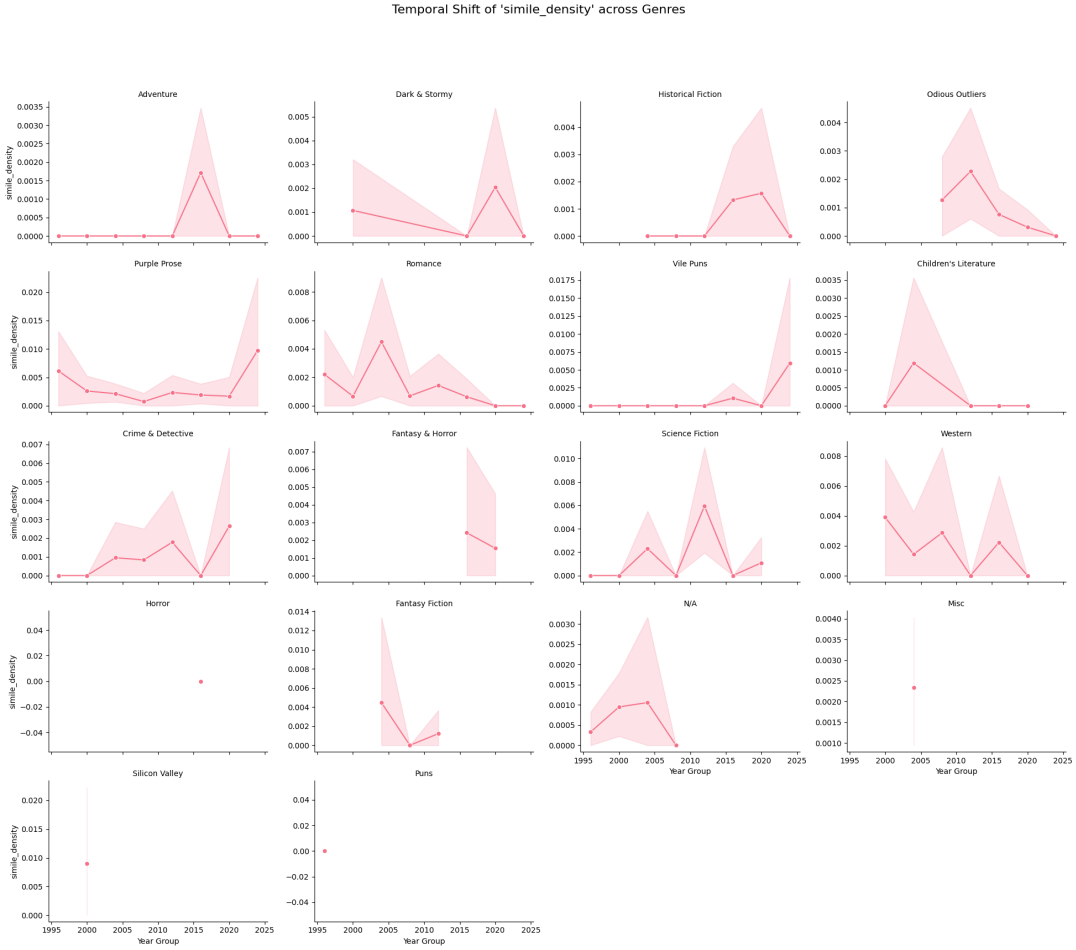

Figure 4.12.: Temporal Shift of `simile_density` across Genres (FacetGrid)

Simile density (`simile_density`) serves as a rhetorical indicator of figurative creativity. Figure 4.12 traces temporal shifts in simile usage across genres. While some fluctuations are visible, particularly around the early 2000s, the Kruskal–Wallis test confirms that these differences are not statistically significant ($H = 6.71$, $p = 0.35$). This suggests that, unlike features such as sentence length or rare word usage, simile frequency has remained relatively stable across decades.

Certain genres, however, continue to rely on figurative elaboration as a stylistic hallmark. *Purple Prose* consistently exhibits the highest simile density across time, with a recent uptick after 2018. Similarly, *Romance* and *Adventure* show moderate usage during earlier phases, reflecting a parody-oriented use of exaggerated comparisons. In these genres, similes often function performatively, invoking overwrought imagery to satirize sentimental tropes.

Genres such as *Science Fiction*, *Crime & Detective*, and *Western* maintain low and steady levels of simile usage. Their rhetorical restraint likely reflects stylistic norms prioritizing clarity, pacing, or lexical novelty rather than decorative elaboration.

Temporary spikes in genres like *Historical Fiction*, *Dark & Stormy*, and *Odious Outliers* during the early 2000s may correspond to a period of stylistic experimentation coinciding with the contest's rising visibility. These, however, are followed by declines or stabilizations, indicating a general convergence toward subtler or structurally integrated rhetorical strategies.

Overall, the genre-specific trends in simile density suggest that while similes remain a genre-sensitive stylistic marker, their expressive prominence has been modulated over time. Later entries tend to employ figurative comparisons more economically or allusively, aligning with broader aesthetic shifts favoring brevity, intertextuality, and subversive understatement in digital-era humor.

Summary The genre–year interaction analysis demonstrates that the stylistic evolution of humorous writing within the BLFC corpus is neither linear nor uniform, but strongly mediated by genre conventions. Although certain corpus-level trends, such as syntactic simplification and increased lexical deviation, are visible over the past three decades, their trajectory varies substantially across genres. Features like simile usage, for example, remain relatively stable over time, with significant variation observed only within specific genres rather than across the corpus as a whole.

Genres like *Purple Prose* have largely preserved their rhetorical intensity, continuing to rely on elaborate figurative structures and syntactic excess as essential components of their parodic function. In contrast, genres such as *Science Fiction* and *Odious Outliers* have increasingly engaged in lexical innovation, deploying brevity and semantic incongruity rather than overt rhetorical ornamentation. Other genres, including *Romance*, *Crime & Detective*, and *Western*, show more moderate adjustments, balancing new stylistic tendencies with established narrative conventions.

These findings confirm that the stylistic construction of humor is both genre-sensitive and historically dynamic. The observed changes reflect a convergence of cultural, aesthetic, and technological pressures, including evolving reader expectations, the influence of internet-based humor, and broader shifts in literary taste. Genres respond to these pressures in different ways, reshaping their stylistic repertoires to remain effective within changing communicative environments.

In summary, this section provides detailed empirical answers to KRQ2 and KRQ3. It suggests that genre not only shapes the linguistic realization of humor along syntactic, lexical, and rhetorical dimensions (KRQ2), but also interacts dynamically with temporal factors to produce genre-specific trajectories of stylistic change (KRQ3). These results reinforce the value of a multidimensional, temporally aware computational approach to understanding how humor adapts across time and genre.

4.4. Limitation

While this study provides a robust multidimensional computational account of stylistic variation in humor writing that captures both genre sensitivity and diachronic variation, it is not without

its limitations, particularly in terms of structural imbalances in the dataset, the granularity of the temporal modeling, the limitations of the particular tool, and the broader scope of interpretation.

First, the **genre taxonomy** provided by the Bulwer-Lytton Fiction Contest is inherently imbalanced and loosely defined. Certain categories (e.g., Odious Outliers, N/A) function as catch-all labels for unconventional or uncategorized submissions, leading to heterogeneity within genre boundaries. Moreover, the frequency of submissions varies markedly across genres and time periods, potentially introducing sampling bias that affects statistical robustness and the generalizability of genre comparisons.

Second, the use of **aggregated temporal bins** (i.e., four-year intervals) was necessary to address year-wise data sparsity, but this strategy may obscure more granular stylistic developments. Additionally, the Kruskal–Wallis test, while appropriate for detecting global differences across groups, does not model interaction effects or intra-group variance. More sophisticated statistical frameworks, such as linear mixed-effects models or Bayesian hierarchical modeling, could better account for nested dependencies between genre, year, and stylistic features.

Third, the analysis is constrained by the operational definitions of stylistic metrics, which are in turn dependent on the capabilities and limitations of current NLP tools. For instance, **simile density** is estimated using regular expressions targeting explicit comparative patterns (e.g., "like," "as…as"), with additional POS-based filtering to retain likely figurative constructions. However, this extraction-based approach cannot guarantee full accuracy: a manual review of the simile candidates identified a non-negligible number of false positives. To illustrate this, a sample of simile candidates drawn from 145 texts in our corpus was manually annotated for validity. The full list of these candidates, together with the underlying dataset and processing code, is available on GitHub[5]. This highlights the difficulty of identifying similes computationally, especially given the creative and unconventional constructions often found in humorous texts. Future improvements could involve computing semantic distance between target and vehicle using word embeddings, or training transformer-based models (e.g., BERT classifiers) to predict the presence of figurative comparison in context. More accurate simile detection would not only enhance rhetorical feature extraction but also enable more nuanced analyses of figurative style in humor.

Similarly, the **rare word ratio** is computed using Zipf-based frequency thresholds, which may not always align with genre-specific expectations or audience perception of lexical salience. Other stylistic metrics, such as POS diversity or clause ratio, are also sensitive to annotation quality and parsing accuracy, which may vary depending on sentence complexity or genre-specific syntax.

Fourth, the study focuses on **formalist linguistic features** as proxies for humor construction, without directly modeling **humor perception or reception**. While genre- and time-sensitive patterns were identified, the actual comedic effect, its perceived success, cultural resonance, or audience alignment, remains unmeasured. Future work could integrate human judgments of humor quality or reader-based metrics (e.g., annotation, laughter ratings, online engagement data) to validate or refine the stylistic correlates of humor effectiveness.

Finally, although the multidimensional framework captures syntactic, lexical, and rhetorical facets of humorous writing, other potentially influential factors, such as **pragmatic context**, multimodal cues (e.g., visual layout in digital formats), and the role of intertextuality, remain outside the scope of this study. Incorporating these elements would allow for a more ecologically valid account of how humor operates across modalities and communicative settings.

Despite these limitations, the present analysis offers clear evidence that stylistic variation in humor is both **genre-conditioned** and **historically dynamic**, underscoring the value of computational stylistics in capturing the evolving linguistic signatures of comedic writing.

[5] https://github.com/YuzuZxy/Humor-Analysis

5. Conclusion

This study presents a multidimensional computational analysis of humorous writing in the Bulwer-Lytton Fiction Contest (BLFC) corpus. Combining literature-informed design with NLP-based feature extraction, statistical testing, and data visualization, it investigates how linguistic features, particularly syntactic complexity, lexical creativity, and rhetorical strategies, vary across genres and evolve over time. The results reveal that humor writing is neither structurally uniform nor temporally static, but instead mediated by genre conventions and shifting cultural and communicative trends.

5.1. Summary of Findings and Answers to Research Questions

Guided by the three central research questions (KRQ1–KRQ3), the analysis unfolds across three perspectives: genre-based profiles, diachronic stylistic shifts, and genre–year interactions.

KRQ1: What computational frameworks and linguistic features have been prioritized in prior research on humor understanding? As discussed in Chapter 2.2, the field of computational humor has progressed from early rule-based methods toward transformer-based architectures such as BERT and RoBERTa. While early models prioritized surface-level features like sentiment polarity, antonymy, and phonetic contrast, recent approaches increasingly draw on contextual embeddings to model incongruity, a central mechanism in humor theory. However, few studies have systematically accounted for genre or stylistic variation. This thesis addresses this gap by operationalizing a suite of linguistically interpretable features, covering syntax, lexicon, and rhetoric, and applying them to a genre-rich, longitudinal dataset.

KRQ2: How do linguistic patterns vary across genres in humorous writing? Chapter 4.1 provides a detailed comparative analysis of genre-specific stylistic tendencies. Three major dimensions were examined:

- **Syntactic complexity:** Genres such as *Romance* and *Purple Prose* feature deeply nested syntactic structures and higher clause ratios, reflecting a tradition of verbose and emotionally charged narrative. In contrast, *Science Fiction* and *Crime & Detective* favor shorter, structurally flatter constructions.
- **Lexical creativity:** *Purple Prose* and genre-ambiguous entries (N/A) exhibit high adjective density and elevated rare word ratios, consistent with parodic maximalism. *Odious Outliers* and *Science Fiction* show high POS diversity, suggesting lexical innovation and genre-specific jargon.
- **Rhetorical strategy:** Simile usage is especially prominent in *Romance* and *Purple Prose*, where figurative exaggeration supports genre parody. By contrast, genres such as *Western* and *Historical Fiction* adopt a more restrained rhetorical profile.

These patterns confirm that genre functions not only as a thematic category but as a stylistic system shaping linguistic choices in humor construction.

KRQ3: How has humorous style evolved over time in the BLFC corpus? The temporal analysis in Chapters 4.2 and 4.3 reveals clear diachronic shifts. From 1996 to 2024, average sentence length steadily declined across the corpus, indicating a broader move toward syntactic economy and punchline-oriented delivery. Simile usage peaked in the early 2000s but declined in later years, signaling a shift away from overt rhetorical embellishment toward more embedded or implicit figurative strategies. The rare word ratio followed a U-shaped trend, with a decline in the mid-2000s followed by a resurgence in the last decade, suggesting renewed interest in lexical play and stylistic idiosyncrasy.

Importantly, the genre-year interaction analysis demonstrates that these stylistic changes are not evenly distributed. For instance, *Romance* shows a trajectory from ornate to pragmatic phrasing, while *Science Fiction* exhibits increasing lexical eccentricity and structural minimalism. Such trends reflect the influence of evolving cultural aesthetics, technological mediation (e.g., internet humor), and audience preference for brevity, incongruity, and intertextuality.

However, it is important to note that rhetorical metrics such as simile density are harder to capture with precision. The current detection method, based on regex and part-of-speech filters, while effective for prototypical similes, may miss more implicit, creative, or structurally unconventional figurative comparisons. Manual annotation revealed a non-trivial proportion of false positives, highlighting the limitations of surface-pattern detection and suggesting future improvements through semantic modeling or classifier-based approaches.

In conclusion, the findings provide empirical support for the core assumption that humor writing is genre-sensitive and historically dynamic. By operationalizing genre features across structural, lexical, and rhetorical domains, this study contributes to computational stylistics in three ways:

First, it provides an interpretable, genre-aware framework for tracking the stylistic evolution of non-classical and deliberately hyperbolic prose, thereby extending the empirical foundations of humor linguistics beyond joke datasets and stand-up comedy factoids.

Second, it emphasizes genre as both a constraint and a creative resource, showing how genre-specific expectations are shaped and subverted by stylistic constructions in humor writing.

Third, it highlights the importance of understanding humor from a diachronic perspective, not only as a timeless cognitive phenomenon, but also as a temporal communicative strategy embedded in evolving cultural, technological, and rhetorical contexts.

5.2. Implications and Future Work

The findings of this study underscore the inherently dynamic nature of humorous writing and its adaptation to both genre-specific conventions and broader temporal developments. By integrating corpus-based linguistic feature extraction with genre-aware and diachronic analysis, this research advances the methodological scope of computational stylistics and offers novel insights into the linguistic construction of humor.

From a theoretical perspective, the study highlights that humor is not merely a surface-level stylistic embellishment but a structured phenomenon shaped by genre traditions, rhetorical expectations, and shifting cultural aesthetics. The observed stylistic divergences across genres, ranging from florid figurative saturation to lexical eccentricity and structural minimalism, suggest that humor operates as a genre-indexical communicative strategy, contingent upon both textual form and sociocultural context.

The findings from the temporal analysis reflect how comedic writing evolves in tandem with changes in media practices, audience preferences, and cultural production. The transition from elaborate, sentence-heavy constructions toward more concise and lexically novel formulations parallels the rise of internet-based humor genres, including meme culture, microblogging, and multimodal satire.

Future research could build upon this foundation in several directions:

- **Semantic enrichment**: Incorporating semantic frame analysis, metaphor detection, or narrative schema extraction could deepen the interpretive scope of stylistic features beyond surface-level lexis and syntax.
- **Cultural grounding:** Linking entries to real-world cultural references, intertextual echoes, or historical events may illuminate how humor negotiates topicality, cultural memory, and sociopolitical resonance.
- **Audience reception:** Introducing human judgment data, such as humor ratings, reaction time studies, or reader-based annotationswould allow triangulation between formal stylistic markers and perceived comedic impact.
- **Model augmentation:** Employing transformer-based language models (e.g., BERT, RoBERTa, GPT) could improve the detection of subtle figurative cues (e.g., sarcasm, irony) and enable more nuanced genre- and time-sensitive classification models.
- **Figurative language modeling:** Improve the precision of simile detection by incorporating semantic similarity measures (e.g., target–vehicle embedding distance) or by training transformer-based classifiers (e.g., BERT) on annotated datasets of figurative comparisons. This could expand rhetorical analysis beyond literal simile patterns and capture a wider spectrum of creative constructions.
- **Cross-linguistic extension:** Applying the analytical framework to multilingual or culturally diverse humor corpora could test the generalizability of genre-stylistic patterns and offer insights into culturally variable humor constructions.

To conclude, while the Bulwer-Lytton corpus is intentionally parodic and stylized, it provides a valuable case study for understanding how language, genre, and time work together to construct humor. Its exaggerated textuality, time span, and genre diversity provide a rare opportunity to explore the intersection of computational linguistics, literary analysis, and humor theory. Future research that combines context-sensitive semantic analysis, cultural data, and reader-centered evaluation will further advance our understanding of humor as a linguistic artifact and dynamic socio-cultural manifestation.

A. Appendix A: Corpus Overview and Sampling

A.1. Corpus Summary

This study analyzed a curated subset of the Bulwer-Lytton Fiction Contest (BLFC) corpus comprising 1,633 entries spanning from 1996 to 2024. The entries were manually checked for completeness and labeled according to genre. The most frequent genres included *Purple Prose*, *Vile Puns*, *Odious Outliers*, *Romance*, and *Science Fiction*.

A.2. Sampling Procedures

For statistical robustness, only genres with a minimum of 20 entries were retained in genre-level analyses. Temporal analysis grouped entries into 4-year intervals to address sparsity and uneven annual participation. Incomplete or duplicate entries were excluded. All preprocessed data were stored in tab-separated files and converted to UTF-8.

B. Appendix B: Annotation Guidelines and Feature Extraction

B.1. Feature Definitions

Each entry was analyzed using `Stanza` (v1.5.0) to extract the following linguistic features:

- `avg_sent_len`: Average number of tokens per sentence
- `tree_depth`: Maximum syntactic dependency tree depth
- `dep_distance`: Mean distance between heads and dependents
- `rare_ratio`: Proportion of tokens with Zipf frequency < 3.0
- `pos_diversity`: Ratio of unique POS tags to total POS tags
- `adj_count`: Average number of adjectives per sentence
- `simile_density`: Similes per token, based on regular expression patterns

C. Appendix C: Case Studies on Linguistic Features

This appendix presents representative entries from the BLFC corpus that exemplify stylistic constructions along three dimensions of analysis: syntactic complexity, lexical creativity, and rhetorical strategy. Each case study was selected to highlight salient features discussed in Chapter 4 and illustrates how humor emerges through genre-sensitive linguistic design.

C.1. Syntax Case Study: Syntactic Accumulation and Tonal Subversion in *Adventure*

The following sentence demonstrates the kind of syntactic accumulation and tonal deflation frequently observed in the *Adventure* genre:

> *"Haul away on those slug guskets, you bilge-scum!" roared the aged captain, leaning wearily against the starboard clog-hutch and watching as the mizzen spittlestoat rose majestically upward until it cuzzled atop the upper spit flukes, and cursing his fate that rum and advancing years compelled him to continually improvise names for the rigging of his own ship but then deciding, with a resigned sigh, that it didn't really matter.*

This sentence is characterized by chained participial constructions ("leaning," "watching," "cursing," "deciding"), extended syntactic dependencies, and recursive clauses. The progression builds both grammatical and narrative complexity before concluding with a deadpan anti-climax ("that it didn't really matter"), which subverts the syntactic intensity for humorous effect.

C.2. Lexical Case Study: Lexical Excess and Romantic Cliché Subversion in *Romance*

The following example illustrates lexical saturation and affective exaggeration, a hallmark of parodic *Romance* writing:

> *"Trent, I love you," Fiona murmured, and her nostrils flared at the faint trace of her lover's masculine scent, sending her heart racing and her mind dreaming of the life they would live together, alternating sumptuous world cruises with long, romantic interludes in the mansion on his private island, alone together except for the maids, the cook, the butler, and Dirk and Rafael, the hard-bodied pool boys.*

This entry is densely packed with adjectives (*sumptuous, romantic, masculine*), affective verbs, and evaluative phrases. The elaborate buildupdriven by excessive imagery and fantasy tropesis undercut by the absurd climax ("Dirk and Rafael, the hard-bodied pool boys"), producing humor through parodic exaggeration of genre conventions.

C.3. Rhetorical Case Study: Rhetorical Accumulation and Tonal Overload

This final example features a long sentence composed of deeply embedded clauses, formal diction, and cumulative rhetorical buildup, parodying melodramatic introspection:

> *After his seventh shot of Jack Daniels, Billy reflected that only a certain kind of man, a Roman Catholic priest, born under the sign of Gemini, whose loved one had been run down by a bus full of inebriated Lazio supporters on a glorious Sunday morning in early April outside a provincial church whose bells were ringing Bach's Toccata and Fugue in B minor, would truly be able to understand the abyss of despair in which he was drowning.*

The sentence combines cultural references, solemn tone, and extended subordination to generate rhetorical excess. Humor arises from the dissonance between the syntactic grandeur and the trivial absurdity of the scenario. The cumulative style ultimately collapses under its own weight, revealing the comic machinery of parodic solemnity.

D. Appendix D: Glossary

D.1. Glossary

- **Simile:** A comparison using “like” or “as” (e.g., “like a foghorn in heat”)
- **TTR:** Type-Token Ratio; a measure of vocabulary diversity
- **POS Diversity:** Variation in part-of-speech tags normalized by sentence length
- **Clause Ratio:** Clauses per sentence, based on syntactic dependency types
- **Zipf Score:** Log-based frequency score; Zipf < 3.0 = rare word

E. Appendix E: Submitted Software and Data Files

This appendix documents the supplementary materials submitted with the thesis to ensure transparency, reproducibility, and accessibility. All resources, including the cleaned dataset, analysis scripts, compiled thesis, and usage instructions, are openly available on GitHub:
https://github.com/YuzuZxy/Humor-Analysis
The repository includes:

- `code/Humor_BL_analysis.ipynb` - Jupyter notebook for feature extraction, statistical analysis, and data visualization
- `data/Bulwer_HumorText.tsv` – Cleaned and genre-labeled dataset used in all analyses
- `data/similes.tsv` – Manually annotated list of simile candidates in 145 texts including year, genre, and label (valid/false positive)
- `thesis.pdf` - PDF version of this thesis
- `README.md` - Instructions for setting up the environment

Execution Instructions

```
# Required environment:
# Python 3.10 or higher
# Dependencies: stanza, pandas, matplotlib, seaborn, scipy

# Step 1: Install required libraries
pip install -r requirements.txt

# Step 2: Launch the analysis notebook
jupyter notebook code/Humor_BL_analysis.ipynb

# The notebook runs sequentially and includes:
# - Data loading and preprocessing
# - NLP pipeline with Stanza
# - Feature extraction (syntax, lexical, rhetoric)
# - Kruskal-Wallis tests
# - Genre and temporal visualizations
```

References

Khalid Alnajjar and Mika Hämäläinen. 2021. When a computer cracks a joke: Automated generation of humorous headlines.

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Jannis Androutsopoulos. 2014. Computer-mediated communication and sociolinguistics. In Alexandra Georgakopoulou and Tereza Spilioti, editors, *The Handbook of Language and Digital Communication*, pages 75–89. Wiley-Blackwell, Chichester, UK.

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Salvatore Attardo. 1993. Violation of conversational maxims and cooperation: The case of jokes. *Journal of Pragmatics*, 19(6):537–558.

Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Walter de Gruyter.

Salvatore Attardo. 2020. *Humor 2.0: Linguistic, Cognitive and Computational Approaches*. De Gruyter Mouton, Berlin. Comprehensive theoretical and computational framework for humor.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research*, 4(3-4):293–347.

Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps LLMs understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 9–17, Online. Association for Computational Linguistics.

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *International Conference on Innovative Computing and Cloud Computing*.

Benjamin K. Bergen and Kim Binsted. 2003. The cognitive linguistics of scalar humor. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25, pages 152–157.

Kim Binsted and Graeme Ritchie. 1994. An implemented model of punning riddles.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901.

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore. Association for Computational Linguistics.

Christian Burgers, Elly A. Konijn, and Gerard J. Steen. 2012. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology*, 31(3):290–310.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge, UK.

J. A. Cuddon. 1999. *The Penguin Dictionary of Literary Terms and Literary Theory*. Penguin Books, London. P. 383.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Faraz Faruqi and Manish Shrivastava. 2018. "is this a joke?": A large humor classification dataset. In *Proceedings of the 15th International Conference on Natural Language Processing*, pages 104–109, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Gilles Fauconnier and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York.

Charles J. Fillmore. 1982. Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin, Seoul.

Sigmund Freud. 1960. *Jokes and Their Relation to the Unconscious*. W. W. Norton & Company.

Mayank Goel, Parameswari Krishnamurthy, and Radhika Mamidi. 2024. Automating humor: A novel approach to joke generation using template extraction and infilling. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 442–448, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

H.P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3, pages 41–58. Academic Press.

Charles R. Gruner. 1997. *The Game of Humor: A Comprehensive Theory of Why We Laugh*, 1st edition. Routledge.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Robert A. Heinlein, Cyril Kornbluth, Alfred Bester, and Robert Bloch. 1959. *The Science Fiction Novel: Imagination and Social Criticism*. Advent Publishers, Chicago.

Christian F. Hempelmann. 2005. Script opposition and logical mechanism in punning. *Humor*, 18(3):297–311.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Rosemary Jackson. 1981. *Fantasy: The Literature of Subversion*. Methuen, London.

Immanuel Kant. 1790. *Critique of Judgment*. Macmillan, London. Translated by J.H. Bernard, 1914.

Hao-Chuan Kao, Man-Chen Hung, Lung-Hao Lee, and Yuen-Hsien Tseng. 2021. Multi-label classification of Chinese humor texts using hypergraph attention networks. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 257–264, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Justine T. Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 78–88, Jeju Island, Korea. Association for Computational Linguistics. Used for stylistic and affective features in literary and humorous texts.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cognitive Science*, 40(5):1270–1285.

Junze Li, Mengjie Zhao, Yubo Xie, Antonis Maronikolakis, Pearl Pu, and Hinrich Schütze. 2022. This joke is [mask]: Recognizing humor and offense with prompting.

Jian Ma, Shuyi Xie, Haiqin Yang, Lianxin Jiang, Mengyuan Zhou, Xiaoyi Ruan, and Yang Mo. 2021. MagicPai at SemEval-2021 task 7: Method for detecting and rating humor based on multi-task adversarial training. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1153–1159, Online. Association for Computational Linguistics.

Victor De Marez, Thomas Winters, and Ayla Rigouts Terryn. 2024. Thinc: A theory-driven framework for computational humor detection.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005a. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005b. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

John Morreal. 1983. *Taking Laughter Seriously*. SUNY Press.

Ross Murfin and Supryia M. Ray. 2003. *The Bedford Glossary of Critical and Literary Terms*, 2nd edition. Bedford/St. Martin's.

Cheryl Nixon. 2008. *Novel Definitions*. Broadview Press.

Neal R. Norrick. 2003. Issues in conversational joking. *Journal of Pragmatics*, 35(9):1333–1359. The Pragmatics of Humor.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Victor Raskin. 1984. *Semantic Mechanisms of Humor*, volume 8 of *Studies in Linguistics and Philosophy*. Springer, Dordrecht.

Yishay Raz. 2012. Automatic humor classification on Twitter. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pages 66–70, Montréal, Canada. Association for Computational Linguistics.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1*, pages 293–306.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

Graeme Ritchie. 2004. *The Linguistic Analysis of Jokes*. Routledge.

Arthur Schopenhauer. 1819. *The World as Will and Representation*. Dover Publications, New York. Translated by E.F.J. Payne, 1958.

Walter Scott. 1992. Essay on romance. In Susan Manning, editor, *Prose Works, Volume VI*, page 129. Oxford University Press, Oxford.

Jonas Sjöbergh and Kenji Araki. 2008. A complete and modestly funny system for generating and performing Japanese stand-up comedy. In *Coling 2008: Companion volume: Posters*, pages 111–114, Manchester, UK. Coling 2008 Organizing Committee.

Răzvan-Alexandru Smădu, Dumitru-Clementin Cercel, and Mihai Dascalu. 2021. UPB at SemEval-2021 task 7: Adversarial multi-task learning for detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1160–1168, Online. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA.

Oliviero Stock and Carlo Strapparava. 2005. HAHAcronym: A computational humor system. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 113–116, Ann Arbor, Michigan. Association for Computational Linguistics.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*.

Tony Veale. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury.

Orion Weller and Kevin Seppi. 2020. The rJokes dataset: a large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6136–6141, Marseille, France. European Language Resources Association.

Owen Wister, John Fox Jr., Mary Austin, Ernest Haycox, Robert E. Howard, and August Nemo. 2020. *Big Book of Best Short Stories – Specials – Western 2: Volume 14*. Tacet Books.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015a. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015b. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

List of Figures

List of Tables