Liberal Arts and Social Studies International Journal

Volume 1 Number 2 (May-August 2025)

Unveiling Character Networks in Literary Texts

Eunjeong Park
Department of English Education, Faculty, Sunchon National University,
Suncheon, SOUTH KOREA
Corresponding Author E-mail Address: parkej@scnu.ac.kr

Abstract
This paper takes a fresh approach to analyzing the musical Wicked, utilizing text mining techniques. Through the use of a text analysis tool, we delve deeper into the narrative, meticulously examining the subtleties found in the characters' dialogue, actions, and relationships. By harnessing this advanced technology, our study seeks to uncover hidden patterns, reveal underlying attitudes, and unravel the thematic threads interwoven throughout the characters' interactions in the entire literary piece. Our goal is not only to deepen our understanding of literature but also to provide novel perspectives on its acclaim and popularity. The findings shed light on the primary protagonists, their characteristics, and the central themes conveyed through the structure of sentences. Moreover, we suggest that text mining methods can offer valuable insights into literary techniques and aid in the development of a semantic markup for literary works.

Keywords: character analysis, literary text analysis, text mining

1. INTRODUCTION

Literary character analysis, a cornerstone of literary studies, has traditionally relied on methodologies of close reading and qualitative interpretation to unravel the complexities of character traits, motivations, and developmental arcs. Frameworks such as S.T.E.A.L. (Speech, Thought, Effect on others, Actions, Looks) have provided structured approaches for the manual deconstruction of characters, guiding readers to synthesize information to substantiate an interpretive claim. Historically, these analyses often centered on ethical evaluations or categorized characters as broadly “positive” or “negative,” sometimes underemphasizing the inherently constructed nature of these literary figures within the narrative. The evolution of literary theory, however, has seen a shift towards analyzing “characterization” as a dynamic process, with theorists like Greimas and Ginzburg redirecting focus towards characters' functional roles and their multidimensional attributes. This established the groundwork from which computational methods have more recently departed or built upon, expanding the toolkit available to literary scholars.

The contemporary era, marked by the proliferation of digital texts and exponential growth in computational power, has ushered in a new wave of analytical methodologies. These computational approaches augment traditional techniques by enabling the analysis of large-scale textual data, thereby uncovering patterns, relationships, and nuances that were previously difficult, if not impossible, to discern through manual reading alone. This “computational turn,” often situated within the broader interdisciplinary field of Digital Humanities, represents more than a mere

technological overlay; it signifies a fundamental methodological shift in how literary texts are approached and understood. The capacity of corpus technologies to allow scholars to query vast textual repositories, revealing macro-level trends and intricate relationships—a practice Franco Moretti termed "distant reading"- exemplifies this transformation. Computational literary studies now involve "obtaining varying degrees of assistance from computers to analyze literary works," treating text as data to rigorously test existing hypotheses and formulate new research questions.

This evolution does not necessarily entail a replacement of established methods but rather an expansion of analytical capabilities. The sheer volume of digitized literary texts often exceeds the processing capacity of traditional human-centric reading practices. Computational tools directly address this limitation, facilitating novel inquiries into extensive literary corpora. Consequently, the current research landscape increasingly seeks to bridge the qualitative depth of nuanced interpretation with the quantitative breadth and empirical rigor afforded by computational methods. This synthesis is particularly pertinent for complex analytical tasks such as the examination of character networks. For instance, the metric text analysis capabilities of software like KH Coder are explicitly designed to *"bridge the gap between quantitative and qualitative analyses"*. KH Coder has gained traction as a tool within digital humanities for content analysis, adeptly converting verbal texts into quantitative information suitable for statistical examination. This integration necessitates a re-evaluation of what constitutes "reading" and "interpretation" in literary studies, fostering a more data-informed hermeneutic where statistical patterns are brought into dialogue with qualitative insights. The increasing reliance on such computational tools may also suggest a need for new forms of training and interdisciplinary collaboration within language and literature education.

This research is dedicated to conducting an in-depth analysis of the characters in Wicked using KH Coder 3, a cutting-edge text analysis software. With this tool, we aim to delve beneath the surface of the narrative, carefully examining the complexities of characters' dialogues, actions, and relationships. By leveraging this advanced technology, our objective is to unveil hidden patterns, attitudes, and themes embedded within the characters' interactions, offering fresh insights into the depth of the story. Our study is particularly focused on exploring the multidimensional character of the protagonist, Elphaba, delving into her life, challenges, and significant role in the overarching plot. Additionally, we will also scrutinize other pivotal characters such as Glinda, along with supporting characters like Fiyero, Nessarose, and Boq. Through gaining a comprehensive understanding of the characters in Wicked, we aim to uncover sophisticated methods for analyzing literary works using text mining techniques.

Through literary exploration, this article aims at exploring the traits of protagonists and main themes in literature. Guiding research questions are as follows:

1. Who appears to be protagonists in literature?
2. What are the main themes of the plot?

2. LITERATURE REVIEW

The integration of computational methods, particularly text mining, into literary analysis has revolutionized the study of literature, enabling researchers to uncover hidden patterns, thematic structures, and character dynamics in texts. This literature review critically examines the current state of research on text mining in literary

analysis, focusing on its application to character networks and thematic exploration in literary texts, as exemplified by the study of *Wicked*. It discusses foundational concepts, recent technological and theoretical developments, gaps and limitations, and prominent trends shaping the field. The review draws on a robust theoretical framework for analyzing character networks and themes in literary works. The details as following.

2.1. Corpus Linguistics: Unveiling Patterns in Language

Corpus linguistics is fundamentally characterized by the empirical analysis of language, grounded in the study of large, principled collections of electronic texts, known as corpora. The seminal work of McEnery & Wilson (2001), "Corpus Linguistics: An Introduction," laid much of the groundwork for contemporary understandings in the field, detailing the methodologies for corpus construction and the analytical techniques that can be applied to them. Their work underscored the versatility of corpus methods, predicting their utility across a wide spectrum of linguistic theories and extending their potential application to various social science domains and, by extension, the digital humanities. The epistemological stance often associated with corpus linguistics, which tends towards positivist or naturalist perspectives, can sometimes create a degree of friction when these methods are integrated into humanities disciplines like literary studies, which traditionally favor more interpretive epistemologies. This highlights that the adoption of corpus methods is not merely a technical decision but also involves an epistemological alignment that warrants careful consideration and justification within the research context.

In literary analysis, corpus linguistic techniques such as keyword analysis (identifying words that appear with statistically significant frequency), collocation analysis (examining words that frequently co-occur), and concordance analysis (scrutinizing words in their immediate textual contexts) enable a systematic investigation of stylistic features, thematic evolution, and character-specific discourse. For example, by analyzing the characteristic vocabulary, collocations, or grammatical patterns in the speech attributed to different characters, or in passages describing them, researchers can uncover insights into their personalities, their relationships with other characters, and their functional roles within the narrative structure. The evolution of corpus linguistics itself, from disparate, center-specific markup schemes to the widespread adoption of standards like XML, reflects a broader movement within the digital humanities towards interoperability and standardization. This technological maturation has been crucial in facilitating the kind of large-scale, often collaborative, and tool-driven research that characterizes much of contemporary computational literary study.

The application of corpus linguistics is particularly relevant to the study of character networks. By identifying and quantifying patterns in how characters speak, how they are spoken about by the narrator or other characters, or how their names and associated concepts co-occur, corpus methods can provide a rich dataset for constructing and interpreting these networks. For instance, the frequent co-occurrence of specific character names within particular lexical or thematic contexts might indicate strong relational ties or shared thematic significance, offering empirical grounding for network structures. The "democratizing" potential of corpus tools further empowers researchers to interrogate established literary canons or dominant interpretations by

furnishing empirical evidence derived from extensive text collections. However, this potential is coupled with the responsibility of critically engaging with the construction of these corpora and understanding what linguistic realities they genuinely represent.

2.2. Thematic Analysis: Identifying Meaningful Patterns

Thematic analysis serves as a crucial method for identifying, analyzing, and reporting patterns—or themes—within qualitative data, including literary texts. Ryan & Bernard (2003) have provided influential techniques for theme identification, which include systematic searches for word repetitions, key-indigenous terms, and keywords-in-contexts (KWIC), alongside the careful scrutiny of larger segments of text to discern underlying patterns of meaning. While Ryan and Bernard initially conceptualized thematic coding as a process embedded within broader analytic traditions such as grounded theory, subsequent scholars, notably Braun & Clarke (2006), have advocated for thematic analysis to be recognized as a distinct and flexible methodological approach in its own right. This ongoing discussion about whether thematic analysis constitutes a "tool" or a "method" underscores a broader debate within qualitative research concerning methodological rigor and definition, a debate that gains particular relevance as these techniques are adapted for computational environments.

In the realm of literary interpretation, thematic analysis facilitates the systematic extraction and interpretation of recurring ideas, topics, motifs, or discourses related to characters, their interpersonal relationships, plot developments, and the overarching messages conveyed by a literary work. This process can be undertaken through meticulous manual coding of the text or can be significantly supported by computational tools capable of identifying word frequencies, lexical patterns, and co-occurrence phenomena, which can then serve as pointers to potential themes. Ryan & Bernard's (2003) emphasis on word-based techniques, such as the formal analysis of computer-generated word frequency lists as initial clues for thematic development, naturally lends itself to computational assistance. This characteristic forms a conceptual bridge between traditional qualitative thematic identification and more quantitative text mining approaches, making their framework particularly adaptable for digital humanities projects.

The relevance of thematic analysis to the study of character networks is profound. Themes identified through this systematic process can imbue the structural representations of character networks with qualitative depth and explanatory power. For instance, if a computationally identified cluster of characters within a network is found, through thematic analysis of their dialogues or associated narrative segments, to frequently engage with themes of "ambition," "loyalty," or "deception," this adds a significant layer of meaning to their structural interconnectedness. It moves the analysis beyond simply mapping connections to understanding the substantive nature of those connections. The inherent flexibility of thematic analysis, allowing its application across diverse theoretical and epistemological frameworks, further enhances its utility for interdisciplinary research, such as that undertaken in the field of Education Integrated through Language and Literature. This adaptability makes it a valuable component in the methodological toolkit for computational literary studies.

The foundational contributions of McEnery & Wilson (2001), Ryan & Bernard (2003), and Koichi Higuchi (2017) are pivotal in shaping the landscape of computational literary analysis, particularly in the context of character network studies. Their distinct yet complementary approaches provide a robust theoretical and methodological scaffold for such research, as summarized in Table 1.

Table 1: Key Contributions of Foundational Scholars to Computational Literary Analysis

Scholar(s)	Core Theory/Methodology	Key Contributions to Literary/Text Analysis	Relevance to Character Network Analysis
McEnery & Wilson (2001)	Corpus Linguistics	Established principles of corpus construction and analysis (e.g., keywords, collocations, concordances), enabling empirical study of language patterns in large texts.	Provides empirical data on character language, interaction patterns (e.g., through collocation of names), and contextual mentions, which can inform network construction and edge definition.
Ryan & Bernard (2003)	Thematic Analysis	Developed techniques for systematic identification of themes in qualitative data (e.g., word repetitions, KWIC, scrutiny of texts).	Enriches network nodes (characters) and edges (relationships) with qualitative thematic meaning, helping to explain *why* characters are connected and the nature of their interactions.
Koichi Higuchi (2017)	Quantitative Content Analysis / Text Mining (KH Coder)	Proposed a two-step approach combining automated word extraction/statistical analysis with researcher-defined coding rules for concept extraction and analysis.	Directly facilitates character network generation through features like co-occurrence networks and offers tools (e.g., clustering, correspondence analysis) for thematic exploration of character-related text segments.

This table synthesizes how these scholars' core ideas provide the building blocks for computationally assisted literary analysis. McEnery & Wilson's work allows for the systematic gathering of linguistic evidence about characters from large texts. Ryan & Bernard's techniques guide the interpretation of this evidence into meaningful themes associated with characters and their interactions. Higuchi's development of KH Coder offers a practical toolkit that integrates many of these principles, enabling researchers to operationalize the analysis of character networks and associated themes from textual data. Together, their contributions underscore the movement towards integrating quantitative evidence with qualitative interpretation in literary studies.

The Musical Wicked

Wicked explores the age-old theme of the conflict between good and evil, showcasing how these opposing forces can coexist within individuals. Unlike traditional tales where heroes battle villains, Wicked offers a fresh perspective with

characters who defy simple labels of good or evil. The narrative delves into philosophical contemplations on evil, giving more emphasis to its darker aspects than to reflections on goodness. It presents a nuanced view of morality, challenging conventional notions by portraying characters in complex, ambiguous shades. The story suggests that the line between good and evil isn't always clear, and individuals may be influenced by their circumstances and experiences. For example, Elphaba is depicted sympathetically, driven by a desire to confront injustice rather than malice. Conversely, characters typically seen as "good" may have flaws and make morally questionable choices. While the novel doesn't claim that evil triumphs over good, it explores the gray areas within characters' moral compasses, offering a deeper understanding of morality. Ultimately, goodness is depicted as elusive and difficult to define, contrasting with the nuanced portrayal of evil beyond the stereotypical image of a wicked witch in a black hat. Goodness requires deliberate intention and awareness, while evil can manifest subconsciously or unintentionally.

These foundational works establish text mining as a rigorous approach to literary analysis, combining computational precision with interpretive depth. However, they primarily focus on methodological development rather than theoretical implications, leaving room for further exploration of how text mining reshapes literary theory

2.3 Computational Approaches to Character Network Analysis

The advent of computational methods has provided powerful new avenues for exploring the intricate webs of relationships between characters in literary texts. Social Network Analysis offers a formal framework for mapping these connections, while Quantitative Content Analysis and Text Mining provide tools for extracting the data needed to build and interpret these networks. Software like KH Coder integrates many of these functionalities, offering a comprehensive platform for such investigations.

2.3.1 Social Network Analysis (SNA) in Literary Studies

Social Network Analysis (SNA) provides a robust theoretical and methodological framework for mapping, visualizing, and analyzing the relational structures between entities within a narrative, where characters are typically represented as nodes and their interactions or relationships as edges. This approach allows for the quantification and systematic examination of social structures embedded in literary texts, transforming qualitative descriptions of character relationships into measurable network properties. The definition of an "interaction" or an "edge" is a critical methodological decision in literary SNA and can vary widely depending on the research question and the nature of the text. Interactions might be defined by characters' co-occurrence in the same scene or chapter, direct dialogue exchanges, mutual mentions, or even shared thematic concerns. This flexibility, however, also presents a challenge, as superficial definitions of interaction, such as mere co-appearance, may yield networks that do not accurately reflect the true depth or significance of character relationships, potentially leading to misleading interpretations.

Once a character network is constructed, a range of metrics can be employed to glean insights into the narrative's social architecture. Measures such as degree centrality

(indicating the number of direct connections a character has), betweenness centrality (reflecting a character's role in bridging different parts of the network), closeness centrality (measuring how easily a character can reach others), network density (the overall level of interconnectedness), and clustering coefficient (the tendency of characters to form tight-knit groups) help to identify influential figures, character communities or cliques, and the overarching structural properties of the fictional social world. These quantitative measures can illuminate power dynamics, reveal social gaps or fragmentation, and objectively assess the relative importance of different characters within the narrative, moving beyond purely intuitive assessments.

Despite its quantitative power, the application of SNA in literary studies is not without its challenges. A primary difficulty lies in meaningfully defining interactions in a way that captures the nuances of literary relationships and subsequently interpreting the derived network metrics within a rich literary and theoretical context. Automated methods for extracting interactions from text, while efficient, can often lack the subtlety to distinguish between significant and trivial connections, or to capture implicit relationships. The true value of SNA in this domain, therefore, often lies not just in the generation of quantitative measures or visualizations, but in the rigorous interpretation of these findings in conjunction with close reading and established literary theory. As Moretti, a proponent of quantitative literary study, has argued, network analysis in literature must strive to go beyond “just showing” structural characteristics and should be actively used to engage with and deepen existing literary understanding. This necessitates greater transparency in methodological choices-particularly in how interactions are defined and extracted and the development of more robust frameworks for linking quantitative network features to qualitative literary interpretations, a concern that aligns with broader discussions of tool criticism within the Digital Humanities.

2.3.2 Quantitative Content Analysis and Text Mining for Literary Insights

Quantitative Content Analysis (QCA) and text mining offer systematic approaches to extracting and analyzing information from literary texts, providing empirical data that can illuminate character attributes, thematic concerns, and narrative dynamics. QCA involves the methodical coding of textual content into predefined categories to quantify the presence, frequency, meanings, and relationships of specific words, themes, or concepts. Text mining, a broader field, encompasses the process of extracting valuable, often previously unknown, insights and knowledge from unstructured or semi-structured text data. It employs computational techniques to identify patterns, trends, and relationships that would be arduous or impossible to uncover through manual analysis alone. The efficiency offered by text mining is particularly transformative for literary studies, where the sheer volume of both primary literary works and secondary critical literature can be overwhelming for traditional, non-computational methods. This allows for research at new scales, addressing questions that were previously unfeasible.

In the context of literary analysis, QCA and text mining can be applied to character dialogue, narrative descriptions, and characters' actions to extract features related to their personalities, emotional states, primary concerns, and developmental trajectories throughout a story. This may involve simple frequency counts of specific

terms associated with characters, sentiment analysis of their speech to gauge emotional polarity, or topic modeling of character-centric textual segments to identify dominant themes in their discourse or in passages related to them. The process of "coding" in QCA, whether performed manually by a researcher or assisted by computational tools, remains a critical interpretive act. The development of robust coding rules and meaningful categories is where significant scholarly judgment resides, shaping the subsequent quantitative output and its interpretation.

These methodologies are particularly well-suited for analyzing dramatic texts and musical librettos, such as the focus of the current research, 'Wicked'. In such texts, QCA and text mining can be employed to dissect song lyrics, spoken dialogue, and even stage directions to quantify thematic presence, analyze character sentiment as expressed through their words, and map patterns of interaction. For instance, studies have utilized Natural Language Processing (NLP) to analyze song lyrics for evolving topics, affect, and narrative structure, techniques directly applicable to the lyrical components of a musical. Liu, M., Yan, J., & Yao, G. (2023) study, which specifically applies text mining to “Wicked,” aims to explore characters’ concerns, relationships, and underlying plot patterns by examining word frequencies, demonstrating the direct relevance of these methods. As text mining tools become increasingly sophisticated, incorporating advanced techniques like sentiment analysis and topic modeling, it becomes imperative for literary scholars to develop a critical understanding of the underlying algorithms' assumptions and limitations. This allows for a more informed evaluation of their outputs, preventing the treatment of these tools as infallible "black boxes" and fostering a more critical engagement with computational results.

2.3.3 KH Coder for Integrated Text Analysis

KH Coder, a free, open-source software package developed by Koichi Higuchi, stands as a significant tool for quantitative content analysis, text mining, and computational linguistics, widely adopted across various disciplines including literary studies. Higuchi’s (2016) methodology, central to KH Coder’s design, advocates a two-step approach to textual analysis. The first step involves the automatic extraction of words from the text and their statistical analysis (e.g., frequency counts, distribution) to explore the data's prominent features, aiming to minimize initial researcher bias. The second step entails the researcher specifying coding rules or dictionaries to extract predefined concepts or themes from the data, which are then subjected to further statistical analysis to deepen the investigation. This structured, two-step process inherently promotes a mixed-methods research design, encouraging a progression from data-driven exploration to more hypothesis-driven conceptual analysis, effectively bridging quantitative discovery with qualitative interpretation.

The software offers a suite of functionalities particularly relevant for character network analysis and thematic exploration. These include word frequency lists, Keywords-In-Context (KWIC) concordance displays, collocation statistics, correspondence analysis, hierarchical cluster analysis, and, crucially for this research, co-occurrence networks. Co-occurrence networks can visually represent relationships between entities based on their proximal appearance in the text; for instance, characters whose names frequently appear together, or characters frequently associated with specific thematic words or concepts. Such visualizations provide an intuitive way to

map out the relational landscape of a literary work. The availability of these diverse analytical features within a single, freely accessible tool like KH Coder democratizes access to complex computational techniques for literary scholars, particularly those who may not possess extensive programming expertise.

The application of KH Coder to analyze a musical libretto like 'Wicked' aligns well with its documented uses in exploring character concerns, interpersonal relationships, and plot patterns through the analysis of word frequencies, thematic clusters, and network visualizations. Steinhall, N., et al's (2024) study on "Wicked" is a direct example of this application. The software's capacity to process English language texts and perform a variety of statistical analyses makes it a suitable instrument for a detailed investigation of the rich textual data within a musical, encompassing both dialogue and lyrical content.

Among KH Coder's strengths are its user-friendly interface, which has contributed to its increasing popularity, its open-source nature allowing for methodological transparency and verification, and its explicit design to facilitate the integration of quantitative findings with qualitative analytical depth. However, effective use of the tool necessitates careful data preparation, including text cleaning and formatting. Furthermore, a general challenge in computational literary studies, applicable here, is the crucial step of interpreting statistical outputs and visualizations within the nuanced context of the literary work. While KH Coder provides powerful means for quantitative analysis and pattern discovery, the ultimate validity and richness of the findings in literary studies depend heavily on the researcher's qualitative interpretive skills and their ability to connect computational results back to the text's specific meanings, aesthetic qualities, and broader literary or cultural contexts. The tool, therefore, serves as an aid to, rather than a replacement for, scholarly interpretation.

2.4 Recent Developments, Challenges, and Future Directions

The field of computational literary analysis has witnessed rapid evolution, particularly since 2020, driven by advancements in artificial intelligence and a growing critical engagement with the methodologies employed. This period is characterized by the increased sophistication of analytical tools, the emergence of specialized subfields, and a heightened awareness of the limitations and ethical dimensions of computational approaches to literature.

2.4.1 Advancements in Computational Literary Studies

The years from 2020 to the present have been marked by a significant surge in the application of advanced Natural Language Processing (NLP), Artificial Intelligence (AI), and especially Large Language Models (LLMs) like GPT variants, to textual analysis within literary studies and the broader digital humanities. These sophisticated tools offer enhanced capabilities for a range of tasks, including nuanced sentiment analysis, complex topic modeling, automated narrative generation, and even the simulation of dynamic character interactions in dramatic scenarios. The rapid development of LLMs, in particular, presents both a substantial opportunity for more refined automated analysis of literary texts and a considerable challenge due to their

often opaque "black box" nature and their documented potential for generating biased or fabricated content (hallucinations), thereby demanding even more rigorous critical oversight and validation from researchers.

Alongside these technological advancements, specialized subfields are beginning to coalesce, such as "Computational Narratology." This emerging area seeks to systematically integrate data-driven computational methods with established narratological theories to model, analyze, and interpret narrative structures, plot progressions, character networks, and thematic developments across diverse media and cultural contexts. The formalization of such subfields signifies a maturation of computational literary studies, indicating a move from the general application of computational tools to the development of more specialized, theory-informed computational approaches tailored to specific aspects of literary inquiry, like narrative itself.

Concurrent with this expansion of technical capabilities is a growing scholarly emphasis on "tool criticism". This involves a critical and reflexive evaluation of the underlying assumptions, inherent biases, and functional limitations of the computational tools and methods being deployed in literary research. New theoretical frameworks are also emerging that aim to better integrate computational findings with traditional hermeneutic practices, addressing concerns about representation, algorithmic bias, and the overall validity of computationally derived interpretations. This dual movement-towards increasing technological sophistication on one hand, and deepening critical self-reflection on the other-suggests a field that is actively expanding its analytical power while simultaneously engaging in a crucial examination of its methodological and epistemological foundations.

3. RESEARCH METHODOLOGY

This study focuses on content analysis through text mining techniques using KH Coder 3 (Koichi, 2017), offering a unique opportunity to delve into the narrations of characters from various perspectives, including patterns of speech and the complexities of character interactions that traditional literary analysis might miss. The detailed data analysis procedure involves several steps. Step 1 involves preparing the texts for processing. The corpus consists of 27,140 tokens (i.e., individual items) and 2,597 types (i.e., distinct classifications in the text) (McEnery & Wilson, 2001). Step 2 entails initiating a new project and configuring stop-words. Stop-words are functional words like prepositions and contractions that are excluded to ensure more meaningful analysis results. Step 3 focuses on extracting the word frequency list from the text, categorized by part of speech and occurrence rate. This step aims to identify the main protagonist, supporting characters, and primary themes portrayed. In Step 4, word co-occurrence analysis is conducted to uncover connections and relationships among words or expressions, aiding in the identification of underlying themes and patterns. Step 5 involves performing correspondence analysis, allowing for a more comprehensive understanding of the narrative's progression by examining words with distinct characteristics.

4. RESEARCH RESULTS & DISCUSSION

To address Research Question 1, we conducted an analysis of keyword frequencies. Examining the statistics of word frequency proves to be a useful and insightful method for identifying the underlying themes within a text. This approach is based on the premise that words with higher frequencies provide more significant insights into literary themes compared to less frequently occurring words (Ryan & Bernard, 2003). Initially, Table 1 presents the top ten most frequently occurring words across the four lexical categories, along with their respective frequencies. Interestingly, while not many adjectives surfaced, the analysis revealed occurrences of 'good' (30 times), 'wicked' (24 times), 'happy' (15 times), 'perfect' (8 times), and 'wrong' (7 times). Through the keyword frequency analysis, it became evident that 'Elphaba' and 'Glinda' emerge as the central characters in the musical Wicked. Although this observation may have been apparent initially, the results obtained through text mining provide more precise information, making a substantial contribution to this study. Rawlins (2009) emphasizes that "telling stories together explores the points of view and particularities of each friend's individuated life" (p. 47). In their dialogues, 'Elphaba' and 'Glinda' take turns speaking and listening, highlighting the importance of dialogue and active listening, as demonstrated in Wicked (Schrader, 2013).

Table 1 Frequency of Keywords

Proper Noun	Freq	Noun	Freq	Verb	Freq	Adv	Freq	Adj	Freq
Elphaba	384	way	41	do	252	so	74	good	30
Glinda	244	one	39	have	188	just	70	wicked	24
Dorothy	183	man	39	go	106	here	56	happy	15
Wizard	131	door	32	know	84	now	54	perfect	8
Fiyero	114	student	29	get	83	back	44	wrong	7
Scarecrow	105	guard	26	see	74	well	38		
Oz	81	heart	23	sing	71	then	34		
Nessarose	74	thing	22	come	53	too	27		
Boq	66	time	21	look	53	never	26		
Witch	50	stage	20	think	50	very	24		

For Research Question 2, we utilized the Co-Occurrence Network of Words, as illustrated in Figure 1, to identify topical groups. Table 2 presents four distinct topical groups that emerged from our analysis. In the first group (Topic-1), words such as 'Witch,' 'Elphaba,' 'know,' 'guard,' and 'wicked' were clustered together, leading to the topic of Elphaba's portrayal as a witch. The second group (Topic-2) consisted of words like 'Wizard,' 'Oz,' 'get,' 'see,' and 'door,' primarily providing contextual information. The third group (Topic-3) included 'Glinda,' 'Fiyero,' 'follow,' 'Boq,' and 'make,' indicating conflicts among the characters. Lastly, the fourth group (Topic-4) featured words such as 'Elphaba,' 'Glinda,' 'heart,' 'make,' and 'just,' highlighting the close relationship between Elphaba and Glinda.

Table 2. Topic Information

	1st Keyword	2nd Keyword	3rd Keyword	4th Keyword	5th Keyword
Topic-1	Witch	Elphaba	know	guard	wicked
Topic-2	Wizard	Oz	get	see	door
Topic-3	Glinda	Fiyero	follow	Boq	make
Topic-4	Elphaba	Glinda	heart	make	just

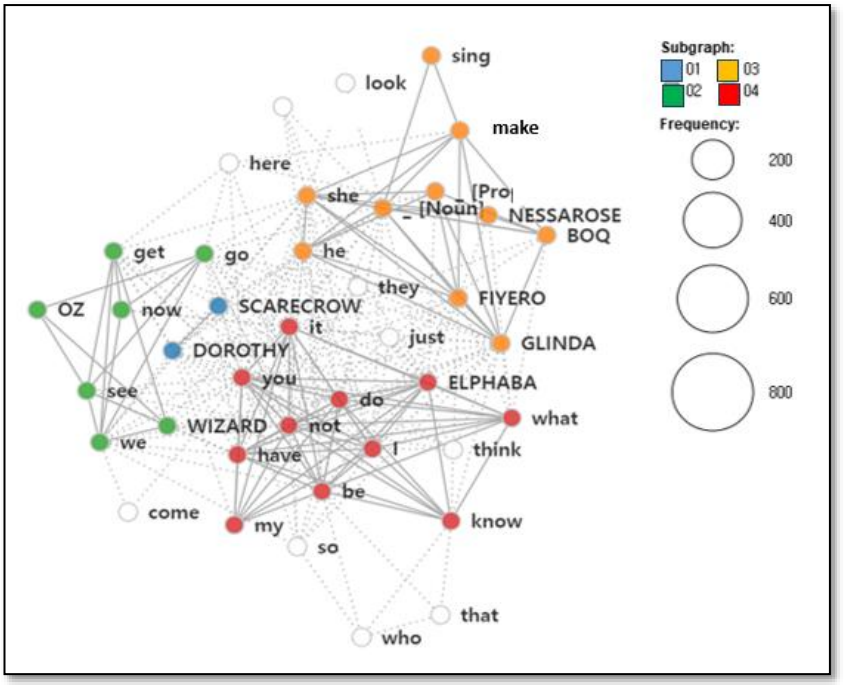

Figure 1. Co-Occurrence Network of Words

5. CONCLUSION

This research sheds light on the promising application of text mining in literary analysis, offering a valuable tool to unveil the intricate techniques authors employ to construct narrative complexity in a linear fashion. Integrating text mining methods into literary analysis represents a significant advancement, providing scholars and researchers with robust tools to uncover hidden patterns, themes, and insights within vast and intricate literary collections. By utilizing advanced computational techniques, this approach enables a nuanced exploration of texts, surpassing traditional methods to uncover deeper meanings and trends that may have otherwise gone unnoticed.

In essence, the fusion of text mining techniques with literary analysis not only enriches scholarly endeavors but also fosters a deeper comprehension of the diverse narratives that shape our cultural landscape. The ongoing evolution of these methods presents an intriguing avenue into the essence of literary inquiry, urging researchers to unveil the mysteries within texts and redefine the boundaries of literary scholarship.

REFERENCES

Algee-Hewitt, M., et al. (2021). Sentiment and character networks in 19th-century fiction. *Digital Scholarship in the Humanities*, 36(2), 245–263. https://doi.org/10.1093/llc/fqaa012.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bode, K. (2020). A world of fiction: Digital collections and the future of literary history. *University of Michigan Press*. https://doi.org/10.3998/mpub.8784777

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Higuchi, K. (2016). KH Coder 3 reference manual. *Kioto (Japan): Ritsumeikan University*.

Jockers, M. L., & Mimno, D. (2020). Text mining and the humanities: A critical introduction. *Journal of Digital Humanities*, 3(1), 45–60.

Koichi, H. (2017). A two-step approach to quantitative content analysis: KH Coder tutorial using Anne of Green Gables (Part II). *Ritsumeikan Social Sciences Review*, *53*(1), 137-147.

Liu, M., Yan, J., & Yao, G. (2023). Themes and ideologies in China's diplomatic discourse–a corpus-assisted discourse analysis in China's official speeches. *Frontiers in Psychology*, *14*, 1278240.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.

Moretti, F. (2013). *Distant reading*. Verso Books.

Rawlins, W. K. (2009). *The compass of friendship: Narratives, identities, and dialogues*. Sage.

Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes. *Field Methods*, *15*(1), 85-109. https://doi.org/10.1177/1525822X02239569

Schrader, V. L. (2013). Friends "for good" wicked: A new musical and the idealization of friendship. *Communication and Theater Association of Minnesota Journal*, *36*, 7-19.

Shih, P. S. H. (2014). The metamorphosis of the witch: Evilness and the representation of the female body in the Wizard of Oz and Wicked. In M. Hedenborg-White & B. Sandhoff (Eds.), *Transgressive womanhood: Investigating vamps, witches, whores, serial killers and monsters* (pp. 27-33). Brill.

Steinhall, N., McPettit, R., Bond, J., Parks, M., Khan, M., Sharfarz, D., ... & Cabrera, D. (2024). Wicked solutions for wicked problems: Misalignment in public policy. *Journal of Systems Thinking*, *4*(3), 1-68.