Jordan Journal of Modern Languages and Literatures Vol.18, No. 1, 2026, pp301-310

JJMLL

A Textual Analysis of the Musical *Wicked* Using Text Mining Techniques

Eunjeong Park*
Department of English Language Education, Sunchon National University, Suncheon, Jeollanado, South Korea

Received on: 29-5-2024 **Accepted on: 10-2-2025**

Abstract

This study aims to investigate the potential of text mining use to analyze the musical piece *Wicked*. It delves into the complex concerns and relationships of the characters, uncovering hidden plot patterns. The study shows the main characters, their personalities, and the major themes by using text minig techniques. The findings provide insight into the primary themes, as illustrated by word frequency tools, which can be applied to various fields of education and applied linguistics. The use of text mining will contribute to developing meaningful tagging systems for second and foreign language education, designed to be user-friendly for learners and practitioners alike. Through various textual analytical tools, researchers and educators gain a richer understanding of major themes, advancing academic discourse. Such research will, moreover. Influence educational disciplines and help inform society about the potential of technology in analyzing and interpreting literature and art.

Keywords: Character Analysis, Literary Text Analysis, Technology, Text mining, wicked.

1. Introduction

Natural language processing (NLP) technology has advanced significantly, prompting shifts in the methods and approaches used by literary scholars. Various studies (Elson et al. 2010; Kokkinakis & Malm 2011) have applied NLP to analyze literature across different languages and genres. These works generally aim to reveal character relationships through lexico-syntactic patterns and co-occurrence metrics. This study demonstrates how researchers use advanced text-mining techniques to examine literature, showing that analyzing syntactic patterns and character expressions can deepen our understanding of both the narrative and individual character traits.

Text mining, a form of NLP, has seen rapid development and has influenced the methods used by literary scholars. Text mining allows researchers to process and analyze large volumes of text, unveiling patterns and insights that might be difficult to capture through traditional analysis. By examining lexico-syntactic structures, co-occurrence patterns, and character expressions, text mining has been used to reveal underlying relationships between characters, themes, and narrative structures within literary works. Studies by Elson et al. (2010) and Kokkinakis & Malm (2011) illustrate the potential of NLP in literature, applying these methods to analyze texts across diverse languages and genres. Such approaches make it

* Doi: https://doi.org/ 10.47012/jjmll. 18.1.14
* Corresponding Author: **parkej@scnu.ac.kr**

possible to explore character dynamics, plot development, and thematic elements quantitatively. Through text mining, researchers can detect subtle syntactic patterns and recurring phrases that reflect deeper character traits and contribute to a fuller understanding of the narrative.

This paper relies on character analysis in the musical *Wicked (music* composed by Stephen Schwartz and songs written by Winnie Holzman) by applying a text mining approach. It uses this modern method to identify patterns, attributes, and themes in character interactions, highlighting the key traits of main characters and the overarching themes within the story. Thus, guiding research questions are as follows:

(1) Who are the main characters in literature as identified through text mining?

(2) What are the major themes of the plot as identified through text mining?

(3) How are the characters described with adjectives?

2. The Musical *Wicked*

2.1 The History of Musical Wicked

Wicked, based on Gregory Maguire's 1995 novel *Wicked: The Life and Times of the Wicked Witch of the West*, reimagines L. Frank Baum's *The Wizard of Oz* by telling the backstory of the "Wicked" Witch of the West, Elphaba, and her unlikely friendship with Glinda the Good Witch. The musical explores the themes of friendship, prejudice, and the blurry line between good and evil. The original Broadway cast starred Idina Menzel as Elphaba and Kristin Chenoweth as Glinda, both of whom received critical acclaim for their performances. Since its debut, *Wicked* has become one of Broadway's most iconic musicals, celebrated for its powerful songs such as "Defying Gravity," "Popular," and "For Good." It has toured worldwide and been adapted for numerous international productions, reaching audiences from London's West End to Australia, Japan, and beyond. Its theatrical impact includes complex stagecraft and special effects, particularly the famous "Defying Gravity" scene where Elphaba appears to fly, showcasing technological advancements in musical staging. The production's success, critical acclaim, and cultural impact have made it a milestone in musical theatre history, and it continues to be performed in major cities globally. Its exploration of identity, societal expectations, and friendship has resonated with diverse audiences, solidifying Wicked as a staple in modern musical theatre.

2.2 The Plot of Wicked[1]

Wicked begins with Act I, where the citizens of Oz celebrate the death of the Wicked Witch of the West, Elphaba, and the arrival of Glinda the Good Witch. The story then flashes back to Elphaba's origins, involving her mother's affair and a mysterious green elixir, eventually focusing on her tumultuous college years with Glinda at Shiz University. As Elphaba's magical talents develop and her friendship with Glinda deepens, the Wizard's darker intentions come to light, leading to rumors branding Elphaba as "wicked."

In Act II, these rumors intensify as Fiyero sets out to find Elphaba, and we learn of her supposed weakness to water. Tensions and betrayals unfold, leading to a climactic confrontation between Elphaba and Glinda after Nessarose's death. Unable to save Fiyero, Elphaba embraces her "wicked" identity,

determined to retrieve the slippers from Dorothy. A twist reveals Fiyero's survival, sparking a reconciliation between Elphaba and Glinda. Elphaba and Fiyero's escape plan emerges, allowing them to leave Oz together. The story weaves themes of friendship, betrayal, and self-discovery against the magical setting of Oz, transforming characters into ways that challenge notions of good and evil. Wicked artfully combines fantasy, morality, and growth, offering an engaging perspective on the hidden story behind *The Wizard of Oz*.

2.3 Good vs. Evil

Wicked explores the timeless theme of good versus evil, highlighting how these forces coexist within each person. Departing from typical narratives where a heroic figure fights a clear villain, Wicked presents characters who defy simple labels of good or evil. Notably, the story leans more toward philosophical reflections on evil than on goodness, casting a spotlight on the darker aspects of its themes (Shih 2014). This approach offers a nuanced look at morality, challenging traditional views by showing that the line between good and evil is often blurred, shaped by each character's circumstances and choices. Elphaba, for example, is portrayed sympathetically, her actions driven more by a fight against injustice than by malicious intent. Meanwhile, characters traditionally viewed as "good" display flaws and make morally ambiguous decisions. The story doesn't argue that evil surpasses good; rather, it reveals the complex shades of the characters' moral choices. Goodness, depicted as elusive and difficult to define, contrasts with a more layered portrayal of evil, which is shown to go beyond the stereotype of a wicked, green-skinned witch (Shih 2014). While goodness requires conscious intention, evil can sometimes emerge subconsciously and even without intent.

This paper may extend existing scholarship on *Wicked* by applying advanced text mining techniques, offering new insights that go beyond traditional literary analysis. While previous studies have examined themes, character relationships, and linguistic patterns in the musical, this study will leverage data-driven methods to uncover patterns, recurring motifs, and subtle shifts in language that reflect deeper narrative structures and character development.

3. Research Method and Tools

3.1 Data Analysis

This article notably focuses on content analysis and text mining, providing a unique opportunity to explore the characters' narrations in multiple dimensions, patterns of speech, and the intricacies of character dynamics in ways that traditional literary analysis may overlook.

The detailed procedures of data analysis consist of several steps. Step 1 is to prepare texts to be processed. The corpus contains 27,140 tokens (i.e., an arithmetical count of items) and 2,597 types (i.e., a classification in a text) (see McEnery and Wilson 2001). Step 2 is to start a new project and configure stopwords. The function of stopwords is to exclude particular functional words, such as prepositions and contractions, to ensure that the results of the analysis become more meaningful. Step 3 is to extract the word frequency list of the text. The compiled list is categorized based on both the part of speech and its occurrence rate. This procedure aims to pinpoint the principal protagonist and supporting characters

within the text, while also unveiling the primary themes portrayed. In Step 4, an analysis of word co-occurrence is undertaken. This helps to discover connections and relationships among words or expressions, subsequently facilitating the identification of latent themes and patterns. Step 5 involves conducting correspondence analysis. This phase allows us to scrutinize distinct characteristics throughout the narrative's progression.

3.2 Analysis Tool

The primary analytical tool used in this study is KH Coder 3, a free software designed for text mining, content analysis, and computational linguistics (http://khcoder.net/). KH Coder 3 allows users to enhance a language model through two main processes. First, it enables associations between a plain text and a language model, using statistical methods to identify common patterns in the target language. The second step involves creating guidelines to code themes and topics that emerge as the analysis progresses (Koichi 2017). KH Coder 3 offers several analysis features, including word frequency lists, word co-occurrence networks, correspondence analysis, and hierarchical cluster analysis. Through its text mining capabilities, researchers can identify topic elements within the text. This computational approach is especially effective for analyzing large sources of unstructured data, revealing hidden topic structures (Chen et al. 2021). The study also utilizes Latent Dirichlet Allocation (LDA), a technique that identifies underlying topics by analyzing word occurrences to uncover major themes, treating each text as a mixture of distinct topics.

4. Results and Discussion

4.1 The Frequency of Keywords of Musical Wicked

The first research question is: (1) Who are the main characters in literature as identified through text mining? Analysing word frequency statistics in text mining proves to be an effective and valuable method for describing themes within a text. This approach is grounded in the idea that words with higher frequencies offer more crucial indications for uncovering literary subjects compared to less frequently occurring words (Ryan and Bernard 2003). Initially, Table (1) displays the most frequently occurring ten words across the four lexical categories along with their respective frequencies. Interestingly, not many adjectives emerged, but 'good' (30 occurrences), 'wicked' (24 occurrences), 'happy' (15 occurrences), 'perfect' (8 occurrences), and 'wrong' (7 occurrences) were found through the analysis.

The results indicate that the main character in this work is Elphaba, who appears as the most frequently mentioned proper noun, with 384 occurrences. The second most prominent character is 'Glinda', appearing 244 times. It's clear that 'Elphaba' and 'Glinda' are the most frequently referenced characters in the text. Furthermore, the analysis of adjectives reveals that "good" and "wicked" appear 30 and 24 times, respectively, suggesting these words emphasize the central theme of *Wicked*. Through keyword frequency analysis, 'Elphaba' and 'Glinda' stand out as the principal characters in the musical. While this conclusion aligns with initial assumptions, text mining provides a precise quantitative basis for this insight, enhancing the study's accuracy. Rawlins (2009) explains that "telling stories together

explores the points of view and particularities of each friend's individuated life" (p. 47). In *Wicked*, 'Elphaba' and 'Glinda' engage in a dialogue that alternates between speaking and listening, delivering a lesson on the importance of conversation and attentively listening to friends (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		

4.2 Topic Modelling

The second research question is: (2) What are the major themes of the plot as identified through text mining? In this section, Co-Occurrence Network of Words, Hierarchical Cluster Analysis, and Correspondence Analysis of Words were employed to identify topical groups. Table 2 shows four topical groups. In the first group (Topic-1), 'Glinda', 'Elphaba', 'know', 'heart', and 'wicked' were grouped. The topic of Elphaba as a witch arose from this. In the second group (Topic-2), the words 'Wizard', 'Oz', 'see', 'voice', and 'door' were arranged. Topic-2 provides contextual information only. The third group (Topic-3) contained 'Glinda', 'Fiyero', 'follow', 'Boq', and 'make', indicating conflicts among the characters. In the fourth group (Topic-4), the words 'Witch', 'Elphaba', 'guard', 'make', and 'just' emerged. This is related to the plot that Elphaba would become a witch who pursued justice.

Table 2: Topic Information

Topic	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

To confirm the prior results, Co-Occurrence Network of Words in Figure 1, Hierarchical Cluster Analysis in Figure 2, and Correspondence Analysis in Figure 3 were conducted. First, Co-Occurrence Network of Words, a network diagram with high degrees of co-occurrence through connect lines or edges, shows the following. The first subgraph included 'Dorothy' and 'Scarecrow'. The second subgraph displayed 'Oz', 'Wizard', 'see', 'get', and 'now'. This is about contextual information corresponding to Topic-2 as above. The third subgraph exhibited 'Glinda', 'Fiyero', 'Boq', and 'Nessarose,' revealing conflicts or incidents among the characters. The fourth subgraph contained 'Elphaba', 'do', 'know', and 'think'. This may convey the main character's complexity through the exploration of various thoughts and inner turmoil. The character's internal struggles and conflicting emotions contribute to a nuanced and intricate portrayal, revealing the depth and intricacy of her personality. Hierarchical Cluster Analysis in Figure 2 also generated similar ideas to the results of Co-Occurrence Network of Words, but

Correspondence Analysis in Figure 3 was not helpful in revealing the characters' conflict or struggles. This type of text mining analysis seems beneficial for understanding and developing personalities within a narrative. Exploring the challenges and inner turmoil that the characters face adds depth to the plot of the stories and connects with them on a more profound level (Bonch-Osmolovskaya and Skorinkin 2017).

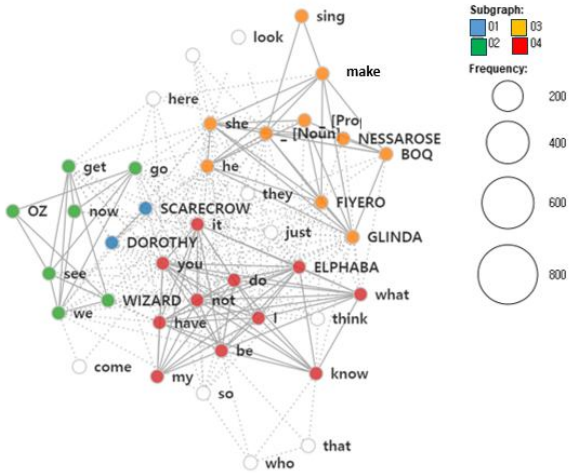

Figure 1: Co-Occurrence Network of Words

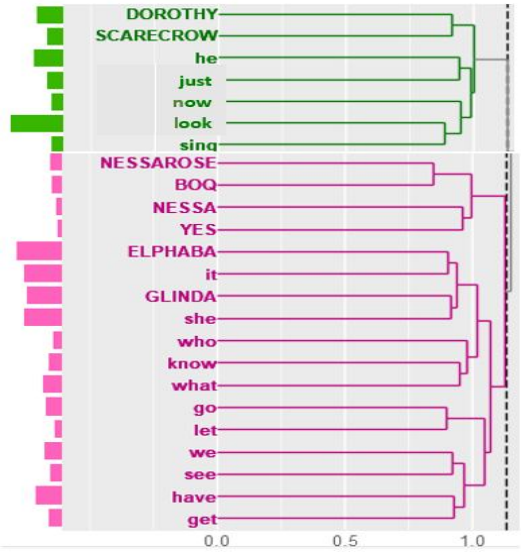

Figure 2: Hierarchy Cluster Analysis

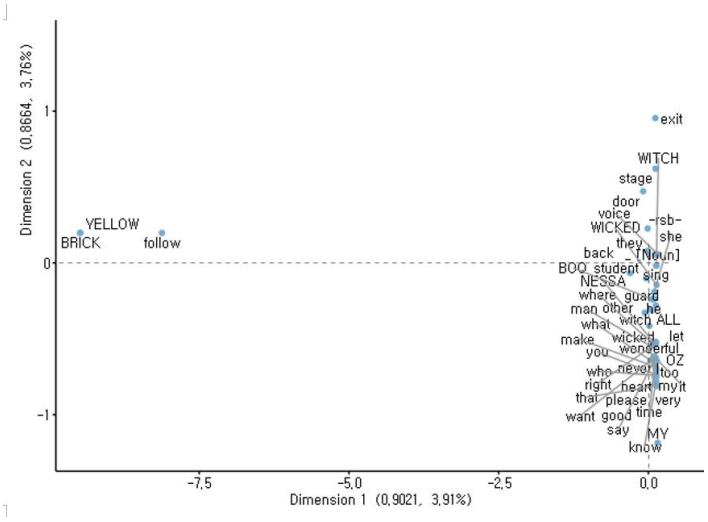

Figure 3: Correspondence Analysis

4.3 Characters Associated with Adjectives

The third research question is: (3) How are the characters described with adjectives? The utilization of Word Association enables researchers to identify words closely linked to a specific word. According to Koichi (2017), these associations are deduced from data through calculations involving conditional probabilities. Upon specifying the condition that a particular word must be present, the system internally retrieves documents that meet this criterion. Subsequently, the frequently occurring words within these documents are displayed in a window. Word Association (see Figure 4) encompasses various features, with the "POS" column indicating the part-of-speech classification for each word. In Table (3), the "unconditional" column presents details regarding the frequency of each word's occurrence in the target file and its probability of appearing across all documents, referred to as unconditional probability. On the other hand, the "conditional" column displays the number of documents that meet specific conditions in which each word is present, along with the probability of its appearance in documents that satisfy those conditions, known as conditional probability. "Jaccard" coefficient function is used to calculate distance, indicating the associations between words and emphasizing whether or not specific words co-occur (Koichi 2017).

As shown in Table (1), five adjectives are present: 'good', 'happy', 'perfect', 'wicked', and 'wrong'. Based on these, Table (3) exhibits the adjectives closely related to the main characters. Because the "unconditional" column offers the total number of the occurring word, the "conditional" column will mainly be used for analysis. As expected, 'Elphaba' is associated with 'good' and 'happy', which are positive words. Interestingly, 'Glinda' is also seen as the positive one with seven occurrences of 'good' in the conditional column. Hence, the finding also confirms that 'Elphaba' and 'Glinda' are the protagonists of the story. 'Fiyero,' who becomes a supportive partner to 'Elphaba,' is described with three occurrences of 'perfect' and 'wrong' respectively. It may be understandable that Fiyero's characteristic transforms from a carefree character to one deeply connected to 'Elphaba' symbolizing the potential for change and

growth. ‘Fiyero’ acts as a catalyst by devising scenarios that erode and strengthen women's companionship (Schrader 2013).

Table 3: Word Association Focusing on Adjectives Connected to Characters

Character	Word	Unconditional	Conditional	Jaccard
Elphaba	good	30 (0.012)	6 (0.019)	0.0176
	happy	15 (0.006)	4 (0.013)	0.0122
Glinda	good	30 (0.012)	7 (0.029)	0.0268
Fiyero	perfect	8 (0.003)	3 (0.027)	0.0261
	wrong	7 (0.003)	3 (0.027)	0.0263

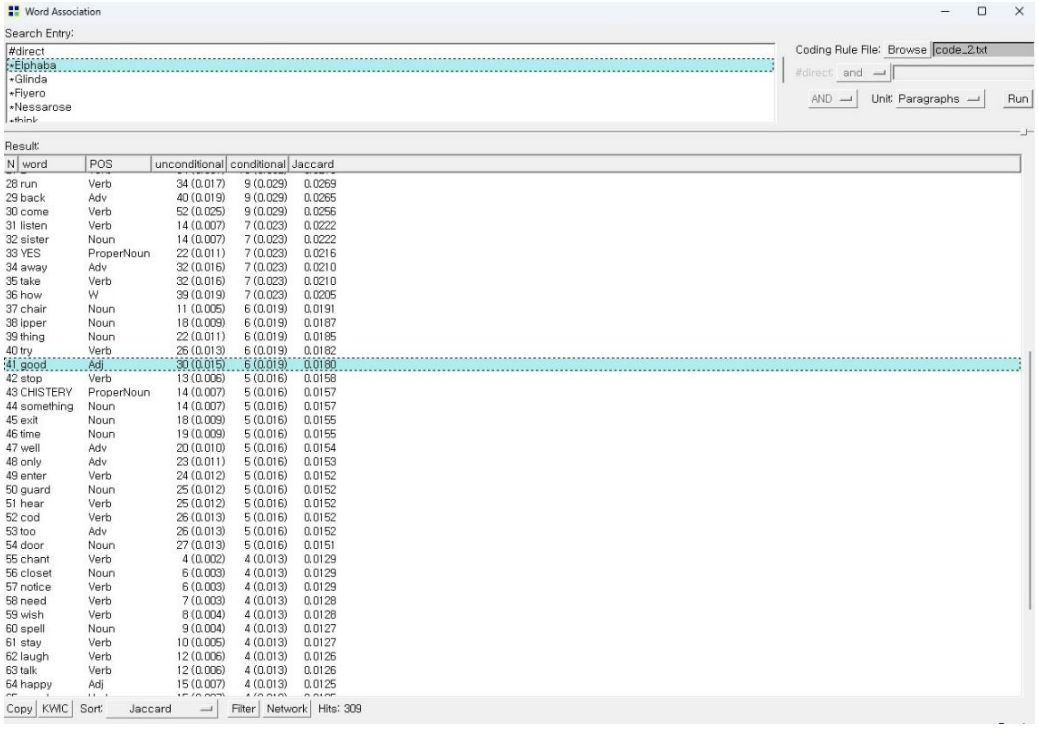

Figure 4: Word Association Screenshot

5. Conclusion

This study has examined the use of text mining within literary studies and education. It has shown that through text mining, researchers and educators can uncover themes and insights that might not be immediately apparent in the text, offering a data-driven approach to understanding rhetorical and linguistic features. While the strengths and limitations of text mining for literary analysis are evident, the challenge lies in capturing subtleties, such as tone, and interpreting the social and historical context reflected in the language of a specific period.

This study integrates machine learning to recognize evolving literary trends and the shifting roles of authorship. As this technology progresses, educational and literary fields will gain from the combined insights of various disciplines. The rise of digital tools and specialized knowledge presents opportunities for educators to develop curricula and teaching strategies that foster engagement. Text mining advances digital approaches in literary research, underscoring technology’s growing impact on research methods.

However, ethical considerations—such as privacy, intellectual property, and cultural respect—must remain central to its use. Overall, text mining for literary analysis deepens our understanding of narratives within their chronological contexts. The ongoing development of these techniques will enable researchers to explore new horizons in literature and expand the boundaries of literary studies.

تحليل النص الأدبي للشريرة الموسيقيَّة من خلال التنقيب في النص

إيون جونغ بارك
قسم تعليم اللغة الإنجليزيَّة، جامعة سون تشون الوطنيَّة، سون تشون، جيولانادو، كوريا الجنوبيَّة

الملخص

تهدف هذه الدراسة إلى استكشاف إمكانيَّة استخدام التنقيب في النص لتحليل القطعة الموسيقيَّة الشريرة، إذ إنه يتعمق في الاهتمامات والعلاقات المعقدة بين الشخصيات، ويكشف عن أنماط الحبكة المخفيَّة، وأظهرت الدراسة الشخصيات الرئيسيَّة وشخصياتهم والموضوعات الرئيسيَّة باستخدام تقنيات تصغير النص، وتوفر هذه النتائج نظرة ثاقبة للمواضيع الأساسيَّة، كما هو موضح من خلال أدوات تكرار الكلمات، التي يمكن تطبيقها على مختلف مجالات التعليم واللغويَّات التطبيقيَّة، إذ سيساهم استخدام التنقيب في النصوص في تطوير أنظمة وضع العلامات ذات المغزى لتعليم اللغة الثانية والأجنبيَّة، التي صُمِّمت لتكون سهلة الاستخدام للمتعلمين والممارسين على حد سواء، من خلال أدوات التحليل النصيَّة المختلفة، إذ يكتسب الباحثون والمعلمون فهمًا أكثر ثراءً للموضوعات الرئيسيَّة، مما يؤدي إلى تقدم الخطاب الأكاديمي، وستؤثر مثل هذه الأبحاث في التخصصات التعليميَّة وتساعد في تعريف المجتمع بإمكانيَّات التكنولوجيا في تحليل الأدب والفن وتفسيرهما.

الكلمات المفتاحية تحليل الشخصيَّة، تحليل النص الأدبي، التكنولوجيا، التنقيب في النص، الأشرار.

Endotes

[1] The information was extracted from McDonald's (2020) thesis paper, "A Textual Analysis of Songs from Wicked."

References

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993-1022.

Bonch-Osmolovskaya, Anastasia and Daniil Skorinkin. 2017. Text Mining War and Peace: Automatic Extraction of Character Traits from Literary Pieces. *Digital Scholarship in the Humanities* 32(Supplement 1): 17-24. https://doi.org/10.1093/llc/fqw052

Chen, Ye, Bei Yu and Yihan Yu. 2021. Analyzing Preservice Teachers' Reflection Journals Using Text-Mining Techniques. *International Journal of Innovation in Education* 7(2):122-143.

Elson, David, Nicholas Dames and Kathleen McKeown. 2010. Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 138-147). Uppsala, Sweden: Uppsala University.

Koichi, Higuchi. 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.

Kokkinakis, Dimitrios, & Malm, Mats. 2011. Character Profiling in 19th Century Fiction. In *Proceedings of the workshop on Language Technologies for Digital Humanities and Cultural Heritage*: 70-77. Hissar, Bulgaria.

McDonald, D. 2020. *A Textual Analysis of Songs from Wicked*. Middle Tennessee State University master's thesis.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction* (2nd ed.). Scotland: Edinburgh University Press.

Rawlins, William K. 2009. *The Compass of Friendship: Narratives, Identities, and Dialogues*. New York: Sage.

Ryan, Gery W. and Bernard H. Russell 2003. Techniques to Identify Themes. *Field Methods* 15(1): 85-109. https://doi.org/10.1177/1525822X02239569

Schrader, Valerie Lynn. 2013. Friends "For Good" Wicked: A New Musical and the Idealization of Friendship. *Communication and Theater Association of Minnesota Journal* 36: 7-19.

Shih, Paris Shun-Hsiang. 2014. The Metamorphosis of the Witch: Evilness and the Representation of the Female Body in the Wizard of Oz and Wicked. In Manon Hedenborg-White and Bridget Sandhoff (eds.), *Transgressive Womanhood: Investigating Vamps, Witches, Whores, Serial Killers and Monsters*, 27-33. Brill.