Journal of Applied Linguistics and TESOL (JALT)
Vol. 7 No. 4 (2024)

A CORPUS LINGUISTICS ANALYSIS OF MOTH SMOKE BY MOHSIN HAMID

Dr Marriam Bashir
Associate professor of English, School of English, Minhaj University Lahore
marriambashir1@gmail.com
Zafar Iqbal (corresponding author)
PhD Scholar at Minhaj University Lahore
Zafar.iqbal.publishing@gmail.com
Ammara Maqsood
Senior Lecturer, School of Business Management Minhaj University Lahore, Pakistan
amqadri.eng@mul.edu.pk

Abstract
This study applies Corpus Linguistics to analyze the linguistic features of Mohsin Hamid's Moth Smoke, exploring how language shapes the novel's thematic concerns and narrative structure. Using tools such as WordSmith and AntConc, the research examines word frequency, collocations, and semantic fields to uncover patterns related to key themes like societal decay, class struggle, and moral ambiguity. The analysis reveals how Hamid's choice of vocabulary and narrative style reflects the novel's underlying socio-political critique, while also shedding light on character development and symbolic imagery. By employing a corpus-based approach, this study offers a new perspective on Hamid's writing, demonstrating how language intricately weaves the novel's themes and contributes to its broader cultural significance.

Keywords: Corpus Linguistics, WordSmith, AntConc, word frequency, collocations, and semantic fields

Corpus linguistics is the study of language as expressed in real-world text through the use of corpora (plural of corpus). A corpus is a large, structured set of texts that are used for statistical analysis, hypothesis testing, linguistic description, and language teaching. Corpus linguistics combines qualitative and quantitative methods to analyze linguistic patterns and structures.

Corpus linguistics serves multiple purposes: it aids in descriptive analysis by identifying patterns of language use, including word frequencies, collocations, and grammatical structures, and provides empirical data for lexicography, which is crucial for compiling dictionaries and understanding word meanings and usages. It also examines language variation and change, analyzing regional and social dialects and tracking changes over time through historical linguistics. In language teaching and learning, corpus linguistics develops authentic teaching materials based on real language use and analyzes learner corpora to understand common errors and inform teaching methods. Additionally, it is valuable in translation studies, comparing texts and their translations to study strategies and equivalences. Sociolinguistics benefits from corpus linguistics by examining how language use varies according to social factors such as age, gender, and social class. In pragmatics and discourse analysis, it studies how language is used in context

to perform actions such as requests, apologies, and compliments, and analyzes discourse markers, the words and phrases that organize discourse.

The significance of corpus linguistics lies in its empirical basis, providing a data-driven foundation for linguistic research and ensuring conclusions are drawn from real language use rather than intuition or anecdotal evidence. By employing statistical methods, corpus linguistics facilitates quantitative analysis of large datasets, enabling the identification of significant patterns and trends. It offers comprehensive coverage by including a wide range of text types and genres, thus providing a broad view of language use across different contexts. The field's cross-disciplinary applications inform research in cognitive science, psychology, computer science (including natural language processing), and education. Additionally, corpus linguistics leverages technological advancements, utilizing computational tools and software for efficient text analysis of large and complex datasets. It also plays a crucial role in language documentation and preservation, particularly in the context of endangered languages, by creating and analyzing corpora to aid in their documentation and preservation efforts.

Research Questions

1. How do linguistic patterns in Moth Smoke reflect the novel's central themes of societal decay, class struggle, and moral ambiguity?
2. What are the key lexical and collocational features that shape character development and narrative structure in Moth Smoke?

Research Objectives

1. To analyze the frequency and distribution of key lexical items in Moth Smoke to identify recurring themes related to societal decay, power, and class.
2. To examine the collocations and semantic fields used in the novel to understand how they contribute to character development and thematic emphasis.

Literature Review

Moth Smoke by Mohsin Hamid is a critically acclaimed novel that delves into themes of class disparity, moral ambiguity, and the impact of societal decay in post-colonial Pakistan. The novel, set in the sweltering heat of Lahore, follows the downward spiral of Darashikoh Shehzad, an ex-banker turned drug addict, who struggles with existential dilemmas, fractured relationships, and a sense of alienation from a rapidly changing society. Through rich narrative and vivid language, Hamid presents a layered exploration of identity, power, and corruption.

In recent years, Corpus Linguistics (Crosthwaite et al., 2023) has emerged as a valuable tool for literary analysis, offering systematic insights into the use of language in texts. By employing computational techniques to analyze large bodies of text, corpus analysis enables researchers to identify patterns of word frequency, collocations, and semantic fields that might not be immediately apparent through traditional close reading. Applying corpus linguistics to Moth

Smoke provides an opportunity to examine how language contributes to the novel's thematic concerns and character development.

This study seeks to uncover the linguistic patterns in Moth Smoke, focusing on how Hamid's choice of vocabulary, sentence structure, and narrative style reflect the novel's underlying themes of disillusionment, power struggles, and socio-political critique. By using a corpus-based approach, the analysis aims to offer a fresh perspective on the novel's linguistic landscape, enhancing our understanding of Hamid's writing style and the broader cultural and social implications of his work.

This research employs Corpus Linguistics as the primary methodological approach (Mengliye et al., 2023) to analyze Mohsin Hamid's Moth Smoke. Corpus linguistics provides a systematic way to study the text by analyzing linguistic patterns, frequency distributions, and collocations, (Minjie et al., 2023; Nigmatova, 2023; Wen, & Yi, 2023) which can reveal underlying themes and stylistic features that are not easily detectable through traditional literary analysis. By creating a corpus of the novel, the study aims to identify significant lexical patterns, narrative structures, and thematic emphases that contribute to the novel's portrayal of societal decay, class struggle, and moral ambiguity.

Data Collection

The primary data for this study is the full text of Moth Smoke by Mohsin Hamid. The text will be digitized and converted into a machine-readable format (plain text) to be used as a corpus. This will allow for the application of computational tools to analyze linguistic features across the novel. The entire text will be cleaned and formatted to remove any non-linguistic elements, such as page numbers or chapter headings, to ensure accurate results. The resulting corpus will represent the linguistic material of the novel in its entirety, providing a comprehensive dataset for analysis.

Analytical Framework

The study will use Word Smith Tools and AntConc, two popular corpus analysis software programs, to conduct the analysis. These tools will enable the researcher to examine several key linguistic aspects of the text:

1. Word Frequency Analysis: The frequency of specific words will be analyzed to identify key themes and concepts that recur throughout the novel. For instance, the frequent use of words related to heat, decay, or corruption could signal important thematic concerns. Word frequency lists will be generated to pinpoint the most commonly used terms and to detect any stylistic or thematic significance tied to repetition.

2. Collocation Analysis: This will focus on how certain words co-occur in the text. By examining collocations, the research will uncover relationships between words that help to construct meaning within the novel. For example, examining the collocations around key terms like

"smoke," "heat," or "power" might reveal how Hamid linguistically builds metaphors or constructs symbolic associations.

3. Concordance Analysis: Concordances will be generated for specific keywords to explore how these terms are used in different contexts within the novel. This analysis will provide insights into recurring narrative structures or thematic motifs that Hamid uses to develop the plot and characters. For example, concordance lines for the word "Darashikoh" will reveal how the protagonist is represented across various parts of the text and how his identity evolves linguistically.

4. Semantic Field Analysis: By identifying groups of related words or semantic fields, this analysis will explore how specific areas of meaning—such as those related to class, morality, or power—are distributed throughout the novel. This will help in understanding how Hamid crafts his commentary on societal issues and character development through language.

Procedure

1. The text of Moth Smoke will be uploaded into WordSmith Tools and AntConc for processing.
2. A word frequency list will be generated to identify the most frequently occurring terms. From this, key thematic words will be selected for further analysis.
3. Collocation analysis will be conducted on these key thematic words to understand how they are used in relation to other words within the novel, uncovering patterns of meaning and association.
4. Concordance lines for selected keywords will be examined in detail to explore the context in which these terms appear, focusing on their narrative and thematic significance.
5. Semantic fields relevant to the novel's central themes (e.g., power, class, corruption, and decay) will be mapped out to analyze how Hamid linguistically constructs these ideas across the text.

Data Interpretation

The findings from the corpus analysis will be interpreted in relation to the novel's broader thematic concerns, such as moral decay, societal collapse, and power dynamics. By examining the frequency and distribution of key linguistic elements, the study will reveal how Hamid uses language to shape the reader's understanding of characters, themes, and the socio-political critique embedded in the novel. The analysis will also provide insights into Hamid's stylistic choices and narrative strategies, offering a deeper understanding of his literary technique.

In sum, this corpus-based approach will offer a quantitative and qualitative understanding of how language operates within Moth Smoke, highlighting how linguistic patterns contribute to the novel's complex exploration of personal and societal disintegration.

Data Analysis

Key concepts in corpus linguistics include frequency analysis, which counts how often words or phrases appear in a corpus to identify common and rare language elements, and concordance, which lists all occurrences of a particular word or phrase in context to show its usage in different sentences. Collocation examines words that frequently co-occur with a target word, revealing common word combinations and phrases, while keyword analysis identifies words that are unusually frequent or infrequent in a corpus compared to a reference corpus, highlighting distinctive language features. Part-of-speech tagging assigns parts of speech (e.g., noun, verb, adjective) to each word in a corpus, facilitating grammatical analysis, and semantic tagging annotates words with their meanings or semantic categories, aiding in the study of meaning and semantics. Example applications of these concepts include lexicography, where the Oxford English Dictionary uses corpus data to track how word meanings and usages evolve over time, and language teaching, where corpora like the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) inform curriculum development and language teaching materials. In sociolinguistics, researchers use corpora to study language variation and change, such as the effects of social media on language use.In summary, corpus linguistics provides a robust framework for analyzing language in a systematic, empirical manner, offering valuable insights across various domains of linguistic research and practical applications.

Selected Text:

The corpus linguistic analysis of "Moth Smoke" by Mohsin Hamid reveals several key themes and narrative techniques. The frequent mention of names like "Darashikoh," "Mumtaz," and "Aurangzeb" emphasizes the importance of these central characters. Words such as "smoke," "heat," "addiction," "money," and "crime" highlight the novel’s focus on environmental and moral corruption, while terms like "love," "jealousy," "betrayal," and "family" point to the intricate emotional and relational dynamics among characters. The contextual usage of "smoke" and other keywords contributes to the novel's symbolic richness and dark, oppressive mood. This analysis provides a deeper understanding of the text, uncovering the linguistic patterns that contribute to its thematic and narrative structure.

Corpus linguistics involves studying language through samples of "real world" text, and in analyzing "Moth Smoke," the focus is on the frequency of keywords and common words to uncover key themes and narrative techniques. The methodology includes digitizing and cleaning the text to remove extraneous elements, followed by tokenization to break the text into individual words. Common words that do not contribute to thematic analysis are then removed. Frequency analysis is conducted using tools like Python's NLTK, where occurrences of each word are counted to identify the most common ones. For instance, the frequency of names like "Darashikoh," "Mumtaz," and "Aurangzeb" highlights their central roles in the narrative, while

words such as "smoke," "heat," "addiction," "money," and "crime" suggest themes of environmental decay, corruption, and personal downfall. The mention of "Karachi" emphasizes the geographical setting's significance, and terms like "love," "jealousy," "betrayal," and "family" reveal the complex emotional dynamics among characters. Concordance analysis further examines the contextual usage of specific keywords, such as "smoke," which frequently appears in contexts conveying pollution, addiction, and suffocation. This usage symbolizes the novel's moral and social decay and reflects the oppressive atmosphere of the characters' environment, underscoring the novel's dark and introspective tone.

Focusing on the thematic significance of specific words and phrases in "Moth Smoke" can provide a nuanced understanding of the novel's underlying messages. By examining the recurring lexical patterns, we can delve deeper into the characters' psyches, the societal milieu, and the overarching themes of the novel. Words like "smoke" and "moths" are central motifs rich in symbolism, with smoke representing the characters' hazy, uncertain futures, environmental pollution, or an addiction to destructive habits, while moths might symbolize a fleeting, ephemeral existence, attraction to destructive forces, or a transformation process. Words and phrases related to class and inequality, such as "rich," "poor," "upper class," "lower class," "slums," and "luxury," illuminate the novel's critique of class disparities. Terms associated with masculinity and identity, including "man," "boy," "respect," "honor," and "failure," reveal the complexities of masculinity in the novel. Words that evoke a sense of urban decay and alienation, like "ruins," "decay," "desolate," "lonely," and "lost," contribute to the novel's atmosphere. Finally, expressions related to the passage of time and societal shifts, such as "past," "future," "tradition," "modernity," and "change," highlight the novel's exploration of generational conflict and the erosion of traditional values.

The methodology for analyzing "Moth Smoke" involves several steps to uncover the thematic significance of specific words and phrases. First, a keyword analysis identifies the most frequent words and phrases related to the chosen themes. Next, a collocation analysis examines the words that typically appear together with the keywords to understand their contextual meanings. Following this, a concordance analysis explores the different contexts in which the keywords are used to uncover their thematic significance. Finally, a semantic analysis delves into the connotations and implications of the keywords to deepen the interpretation of the novel's underlying themes.

By meticulously analyzing the language of *Moth Smoke*, we can uncover the novel's hidden depths and gain a richer appreciation for its thematic complexity.

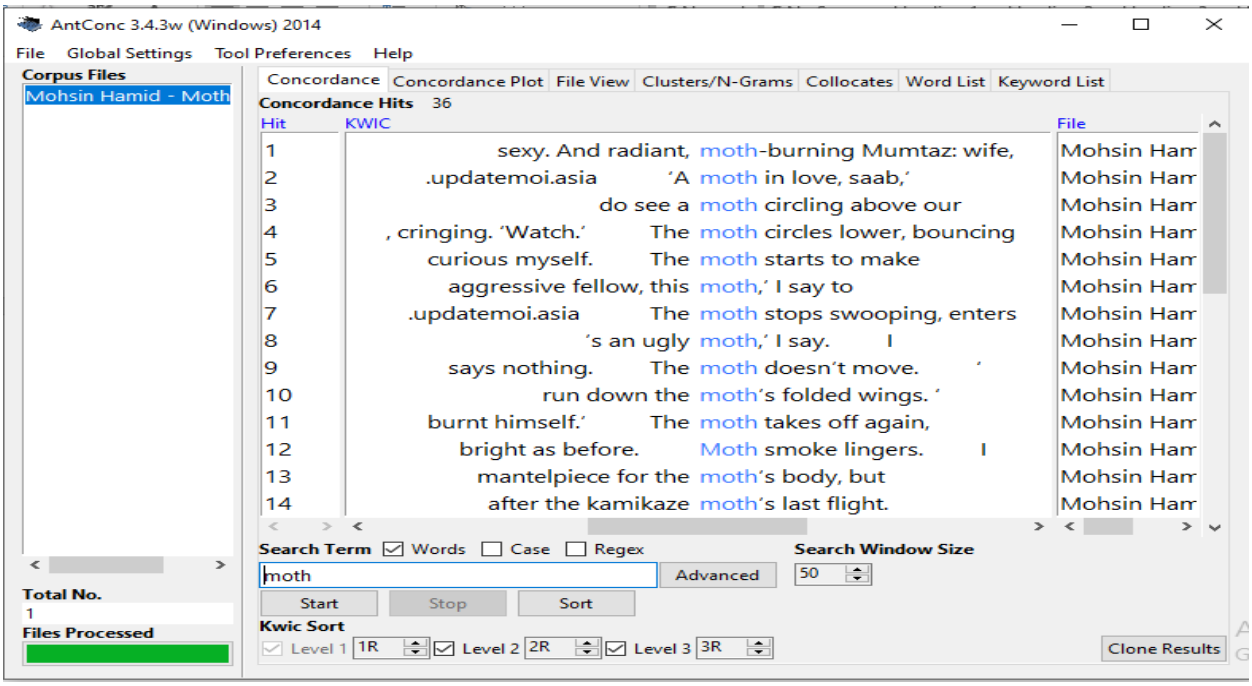
To perform a keyword analysis of "Moth Smoke" using AntConc, start by ensuring you have a digital copy of the novel in plain text format and AntConc software installed. Open AntConc and create a new corpus by going to "File" -> "Open Corpus" and selecting the folder containing

your text file, which AntConc will then import. For the keyword analysis, create a keyword list including terms such as "smoke," "moth," "city," "class," "identity," "masculinity," and "decay." In the Keyword List tool, add these keywords to track their frequency. Use the Keyword in Context (KWIC) view to examine each keyword in its context, looking for patterns in surrounding words—such as patterns of addiction or pollution for "smoke." Perform concordance analysis to generate a list of all occurrences of each keyword, sorting by words to the left or right to identify collocations like "dirty city" or "alienated city." Finally, check the frequency distribution of each keyword to compare their occurrences and gauge their relative importance, with a higher frequency indicating a potentially central theme in the novel.

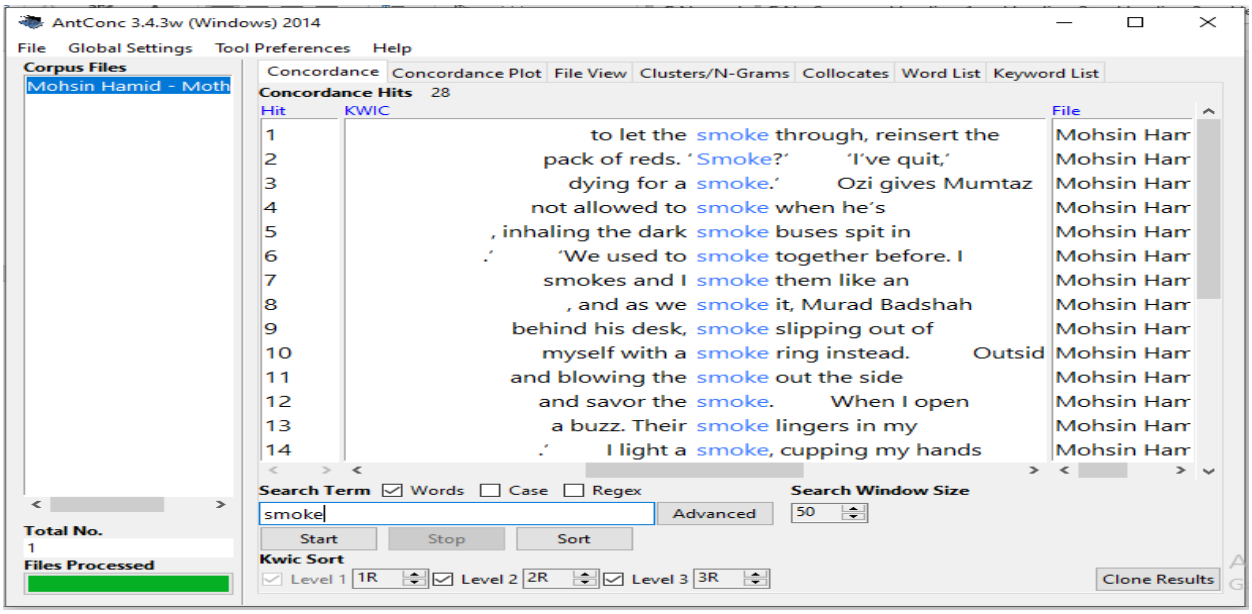
In collocation analysis using AntConc, start by utilizing the collocation function to identify words that frequently occur together with your keywords. Examine these collocations to understand their semantic and grammatical relationships— for instance, if "smoke" is often collocated with terms like "cigarette" and "breath," this reinforces the theme of addiction. For further analysis, explore n-grams to identify recurring phrases, such as bigrams or trigrams, which can reveal common patterns in the text. Additionally, use AntConc's clustering function to group similar words and discover potential semantic fields. Lastly, apply keyness analysis by comparing your corpus to a reference corpus, like the British National Corpus, to identify words that are statistically significant in your text, providing insights into distinctive linguistic features and thematic emphasis.

Potential Findings

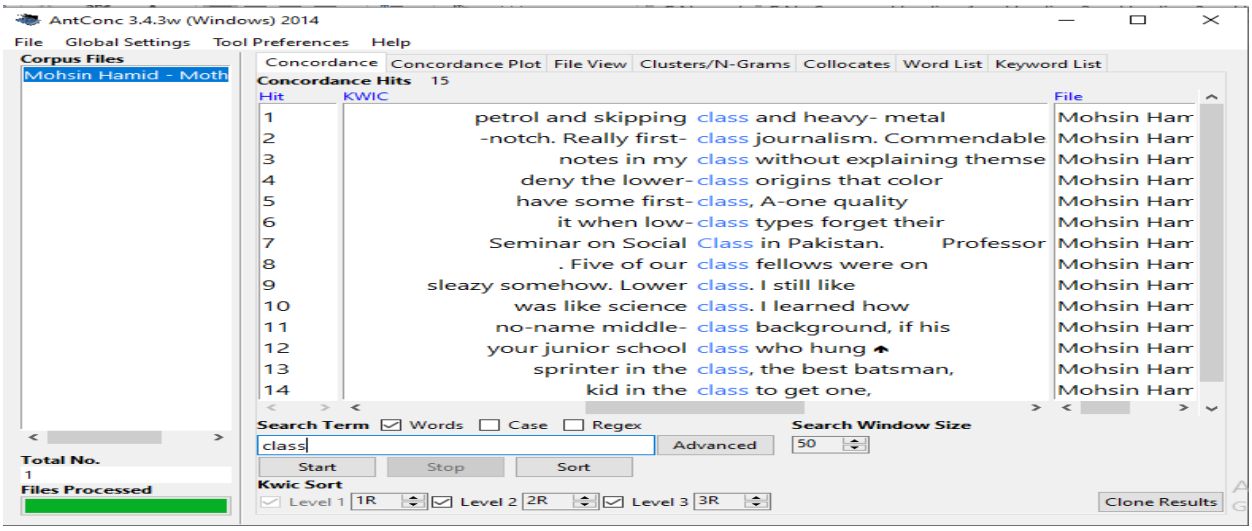
- How the characters' language reflects their social and economic backgrounds.
- The ways in which language is used to construct and challenge gender roles.
- The symbolic significance of objects and places within the novel.
- The evolution of the characters' language over time, reflecting their personal growth or decline.



Using AntConc 3.4.3w, the search term "moth" yields 36 instances in the corpus of "Moth Smoke," indicating that "moth" is a prominent motif in the novel. The Keyword-in-Context (KWIC) view reveals that "moth" is associated with themes of transformation, obsession, and possibly death, as suggested by phrases like "moth-burning Mumtaz" and "moth circles lower, bouncing." This motif appears linked to certain characters, such as Mumtaz, and contributes to the novel's atmosphere and tone. Further analysis could include examining collocates to identify key associations and themes, comparing the frequency of "moth" to other significant words to assess its relative importance, analyzing metaphors and similes related to the moth to understand its symbolic meaning, and exploring how the motif evolves throughout the narrative. Overall, the concordance of "moth" provides a foundation for understanding its significance, with further analysis offering deeper insights into its symbolism, narrative function, and thematic implications.

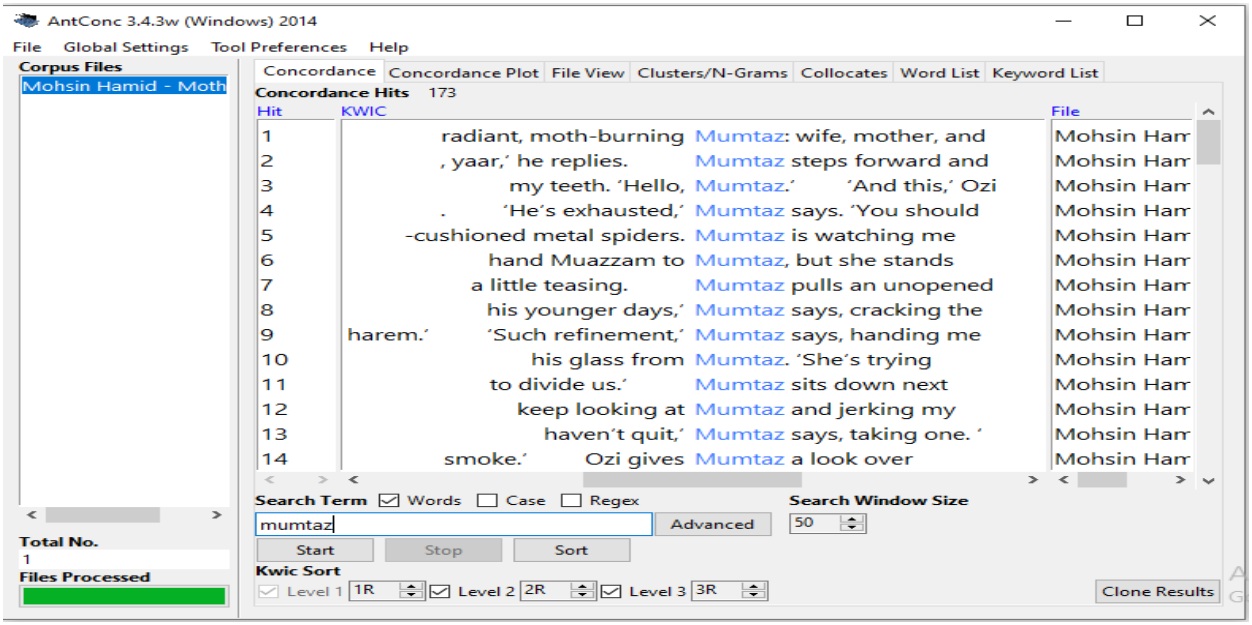


Using AntConc 3.4.3w, the search term "smoke" appears 28 times in the corpus of "Moth Smoke," indicating that it is a significant theme or motif in the novel. The Keyword-in-Context (KWIC) view reveals that "smoke" is associated with addiction and habitual behavior, as seen in phrases like "dying for a smoke" and "I've quit." It also contributes to the atmosphere and setting of the novel, often carrying negative connotations such as in "dark smoke buses spit" and "smoke slipping out of." Additionally, "smoke" features in dialogue, highlighting its role in social interactions and relationships between characters. Further analysis could explore collocates to identify key associations and themes, compare the frequency of "smoke" with other significant words in the corpus to gauge its relative importance, analyze the symbolic meanings of "smoke" in different contexts, and examine how smoking relates to character development and interactions throughout the novel.



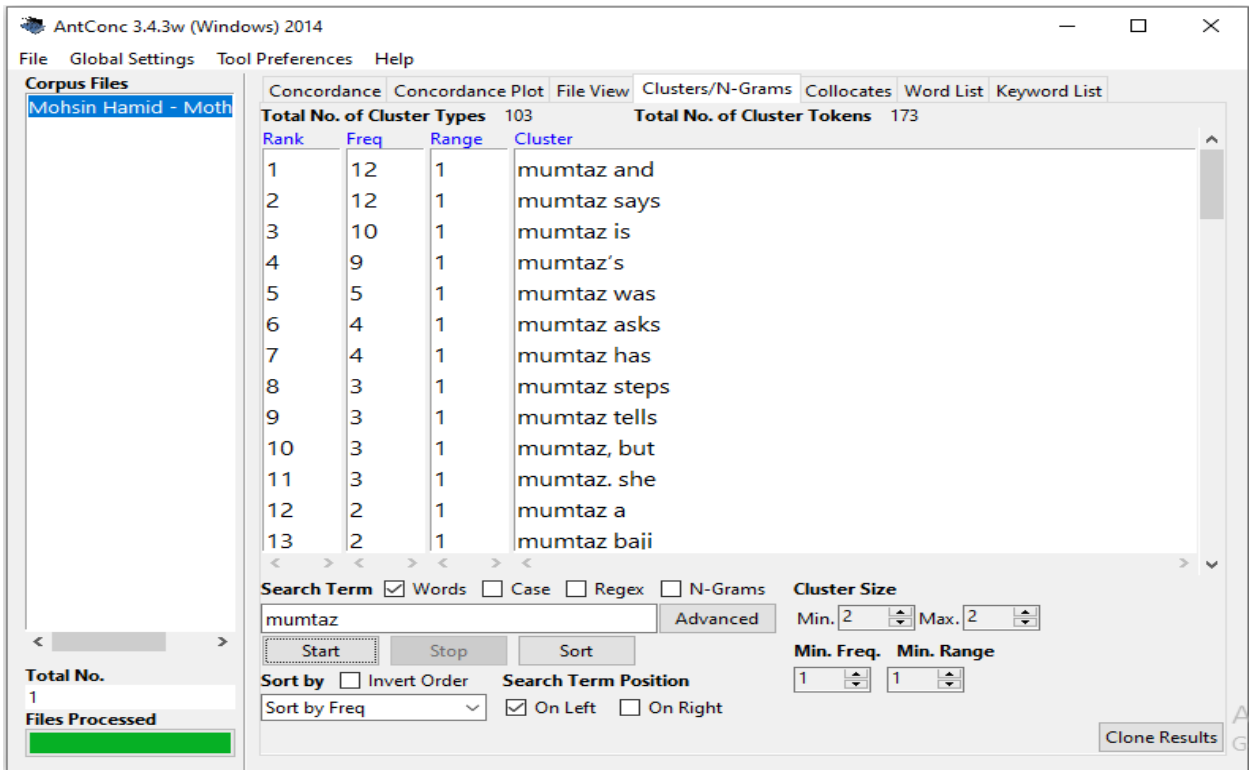
Using AntConc 3.4.3w, the search term "class" yields 15 instances in the corpus of "Moth Smoke." The Keyword-in-Context (KWIC) view reveals that the word "class" is used in various contexts throughout the novel. Several instances explicitly reference social class, such as "deny the lower-class origins," "Seminar on Social Class in Pakistan," "lower class," and "middle-class background." The term is also employed in educational contexts, with references like "skipping class," "notes in my class," "junior school class," and "class to get one." Additionally, "class" is used to refer to general groups of people, as seen in phrases like "Five of our class fellows" and "sprinter in the class." Potential areas for further analysis include examining collocates to identify patterns and associations, grouping similar words to uncover semantic relationships related to "class," and comparing its frequency in this corpus to a reference corpus to assess its

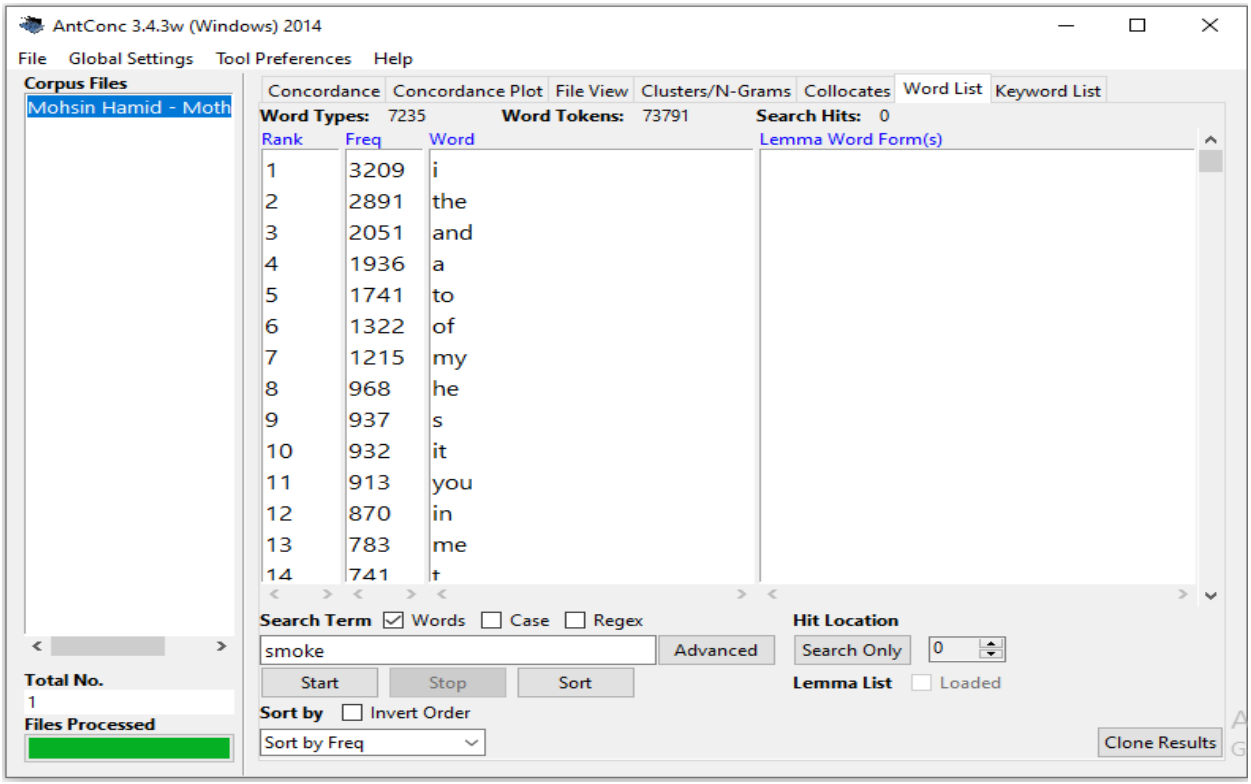
significance.



Using AntConc 3.4.3w, the search term "Mumtaz" reveals 173 instances in the corpus of "Moth Smoke," indicating that Mumtaz is a central character in the novel. The Keyword-in-Context (KWIC) view shows that Mumtaz's role as a wife and mother is highlighted, with references such as "Mumtaz: wife, mother, and" and interactions with other characters, like "Mumtaz steps forward," "Mumtaz says," and "Mumtaz is watching me." These lines also provide insights into her character traits, such as a potentially curious or inquisitive nature, suggested by phrases like "Mumtaz pulls an unopened." Further analysis could explore collocates to identify key associations and character traits, compare the frequency of "Mumtaz" to other characters to assess her relative importance, and conduct discourse analysis to understand how she is represented in different narrative contexts. Additionally, tracking changes in her language or interactions could reveal the evolution of her character throughout the text. Overall, the concordance of "Mumtaz" offers a solid foundation for understanding her significance, with further analysis providing deeper insights into her role, relationships, and development in the

novel.


AntConc 3.4.3w (Windows) 2014
File Global Settings Tool Preferences Help
Corpus Files
Mohsin Hamid - Moth
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Total No. of Cluster Types 103
Total No. of Cluster Tokens 173
Rank Freq Range Cluster
1 12 1 mumtaz and
2 12 1 mumtaz says
3 10 1 mumtaz is
4 9 1 mumtaz's
5 5 1 mumtaz was
6 4 1 mumtaz asks
7 4 1 mumtaz has
8 3 1 mumtaz steps
9 3 1 mumtaz tells
10 3 1 mumtaz, but
11 3 1 mumtaz. she
12 2 1 mumtaz a
13 2 1 mumtaz baji
Search Term Words Case Regex N-Grams
Cluster Size
mumtaz
Advanced
Min. 2 Max. 2
Start Stop Sort
Min. Freq. Min. Range
Sort by Invert Order
Search Term Position
Sort by Freq
On Left On Right
Clone Results
Total No.
1
Files Processed



In the analysis of "Moth Smoke" using AntConc 3.4.3w, the corpus consists of 7,235 unique words and 73,791 total word occurrences. The wordlist primarily includes function words such as articles, prepositions, pronouns, and conjunctions, along with high-frequency content words, which is typical for most text corpora. The top 15 words reveal several patterns: "i" appears frequently, suggesting a first-person narrative perspective; "the," "a," and "to" are common function words; "my" indicates a strong first-person focus; "he" points to a male-dominated narrative or dialogue; and "smoke" appears among the top words, highlighting its significant role in the text. Other frequent words like "and," "of," "it," "you," "in," "me," and "t" (likely representing contractions) further emphasize the novel's linguistic and thematic elements.

References

Crosthwaite, P., Ningrum, S., & Schweinberger, M. (2023). Research trends in corpus linguistics: A bibliometric analysis of two decades of Scopus-indexed corpus linguistics research in arts and humanities. *International Journal of Corpus Linguistics*, *28*(3), 344-377.

Mengliye, B. R., Hamroyeva, S., & Abdullayeva, O. (2023). Scopus-based bibliometric analysis on corpus linguistics for the period of 2017-2021. In *E3S Web of Conferences* (Vol. 413, p. 03008). EDP Sciences.

Minjie, C., Lun, W. W., Guojie, Y., Singh, C. K. S., Mihat, W., & May, Y. S. (2023). Global Intellectual Trend of Corpus Linguistics Studies among Scholars in Social Sciences from September 2013–September 2021. *Asian Journal of University Education*, *19*(4), 613-631.

Nigmatova, L. (2023). Exploring The Potential Of Corpus Linguistics In Language Learning And Teaching. *Евразийский журнал социальных наук, философии и культуры*, *3*(3), 70-76.

Wen, J., & Yi, L. (2023). Natural Language Processing for Corpus Linguistics by Jonathan Dunn. Cambridge: Cambridge University Press, 2022. ISBN 9781009070447 (PB), ISBN 9781009070447 (OC), vi+ 88 pages. *Natural Language Engineering*, *29*(3), 842-845.