A Corpus-Based Critical Discourse Analysis of News Reports on the 2023 Israel-Hamas War

Yankai Liu[1]

[1] City University of Hong Kong, China
Correspondence: Yankai Liu, City University of Hong Kong, China.

Abstract

This study analyzed news reports of the 2023 Israel-Hamas war in three media outlets (Al-Jazeera, China Daily, and CNN). An armed conflict between Israel and Hamas (Palestinian militant groups) has been taking place mainly in the Gaza Strip since 7 October 2023. The 2023 Israel-Hamas war has been the most violent and deadliest for Palestinians in the history of the conflicts between Israel and Palestine. Such a severe global issue has attracted significant international concerns and has been reported extensively. However, due to different ideologies, media outlets used different language to describe the war. This study compiled three corpora with 1,586,438 words for a horizontal comparison and analysis. Combining corpus linguistics (CL) and critical discourse analysis (CDA), this study aimed to reveal how these three media outlets constructed language to depict the war, demonstrating their political stances and revealing their ideologies from a more comprehensive perspective. Based on the findings from analyzing the linguistic features of the three corpora, this study hopes to provide a framework and methodological implications for future studies on news reports, especially on war topics.

Keywords: the Israel-Hamas war, critical discourse analysis, corpus linguistics, keyword, frequency, collocation, lexical distribution

1. Introduction

1.1 Research Background

The Israel-Palestine conflict, which dates back to the end of the nineteenth century, is one of the most protracted and ferocious conflicts in modern history, and a surprising attack by Hamas on Israel on 7 October 2023 escalated the conflict. The Israel-Hamas war, centered in the Gaza Strip, has been the deadliest for Palestinians in the history of the Israeli–Palestinian conflict and the most significant military engagement in the region since the Yom Kippur War. Global issues related to this war, like humanitarian crises, economic disruption, political realignment, international relations, and world order, all pop up. The severity of the Israel-Hamas war has attracted significant international concerns and has been reported extensively by media outlets all around the world. All media outlets are supposed to uphold the principles of fairness, objectivity, and inclusivity, ensuring all voices are heard, all stories are reported, and all truths are displayed. However, with different ideologies and political stances, languages used by some media outlets

show significant biases and undermining, which further leads to misinterpretation among readers and audiences.

Studies have extensively discussed the languages used in the news reports of previous Israel-Palestine conflicts and interpreted them from multiple linguistics perspectives, especially from a critical discourse analysis perspective. Nevertheless, none of the studies focus on the 2023 Israel-Hamas war, and critical discourse analysis is not a suitable method to analyze large amounts of texts nowadays. Therefore, this study aims to fill this gap by implementing the corpus linguistics (CL) method in critical discourse analysis (CDA), which is irreplaceable in analyzing large-scale texts (new reports in this case). The author built three corpora, comprising news reports in Arabic countries, American and Asian media, represented by Al Jazeera Media Network (Al-Jazeera), Cable News Network (CNN), and China Daily, to systematically analyze the languages they used and explore the ideologies and political stances buried behind the language.

1.2 Research Questions and Significance

The general research question addressed in this study is: How do media outlets use different language patterns to convey their ideologies and political stances in the news reports about the same topic (the Israel-Hamas war)? By answering the question, the author provides a replicable analysis framework that leverages the strengths of both quantitative and qualitative analyses and provides theoretical and empirical language evidence. Such a framework will be helpful and efficient in future studies on the language used by media outlets. Theoretically, the author further verified that combining corpus linguistics and critical discourse analysis is feasible, and such methodological synergy could further extend to analyze discourses and ideologies of large-scale texts covering various topics.

1.3 Structure

This study is organized as follows: Section 2 reviews previous studies on corpus linguistics and critical discourse analysis and discusses the theoretical feasibility of combining these two methodologies; section 3 provides detailed methodological information on this study; section 4 demonstrates the preliminary results and corresponding analysis; section 5 concludes this study and discuss the implications for the future studies.

2. Literature Review

2.1 Critical Discourse Analysis

Since its inception, CDA has been applied to multiple domains, including media, law, politics, and education, to investigate language usage. Fairclough (2013) notes that critical discourse analysis combines social analysis with language studies and emphasizes the connection between discourse and social elements, including power relations, ideologies, social identities, and so forth. CDA has also been regarded as "a tool for deconstructing the ideologies of the mass media and other elite groups" (Henry & Tator, 2002, p. 72). Therefore, the language used by media outlets that contain ideologies like political stances can be analyzed through critical discourse analysis. In this study, the war scene of the Israel-Hamas war was different in the three media outlets as each represents a different interest group and ideology, which led to political bias. Such bias can be identified through CDA as it studies "the way social-power abuse and inequality are enacted, reproduced, legitimated, and resisted by text and talk in the social and political context." (Van Dijk, 2004, p. 466). Considering all factors, the language used by media outlets can be analyzed through CDA, and the hidden ideological presuppositions can be revealed from language choices in texts.

Fairclough (1992) sketches a three-dimensional framework for analyzing discourse: discourse-as-text as the first dimension, discourse-as-discursive-practice as the second, and discourse as social practice as the third. From vocabulary, the basic unit of the text, to a broader picture of society, CDA analyzes every detail of the social wrongs. Such CDA-based methodology can be formulated as multiple stages when it comes to analyzing social wrongs. Fairclough (2013) then further formulates four stages to fully emerge CDA methodology into particular research, including selecting a research topic that relates to a social wrong from semiotic aspects, identifying obstacles that prevent social wrong from being addressed, considering whether the social wrong is a natural part of the social order and whether it can be addressed within it or by changing it; finding possible approaches to address the obstacles in social wrong.

However, several criticisms have questioned the

theoretical basis and methodology adopted by CDA. It has been criticized for its objectivity since it pays too much attention to social and political topics. Widdowson (1995) views CDA as an ideological interpretation instead of an analysis, given that CDA demonstrates language controlled by the privileged to uncover societal iniquities. Sharrock and Anderson (1981) even ironically state that Kress and his critical linguistics college "look in the wrong place for something, then complain that they can't find it, and suggest that it is being concealed from them." Such vague analysis is caused by a limited analysis sample since language patterns summarized from small-scale texts are not convincing and cannot even be regarded as "patterns." Stubbs (1997) is concerned that CDA's analysis of texts is not based on standards that can be replicated and tested for reliability. Breeze (2011) also expresses his criticism of the limited text samples used for CDA, which lacks representativeness; he also points out that critical discourse analysts sometimes jump too quickly from the language data to the interpretations of the data, which makes readers question the availability and objectivity of the data.

2.2 Corpus Linguistics

Corpus linguistics (CL), which studies "real language," is a methodology rather than an aspect of language use explanation and description, and it is used to tackle some aspects across the areas of linguistics such as lexicography, pedagogy, sociolinguistics, and discourse analysis (McEnery, 2019). Therefore, the corpus-based approach is suitable to be applied in multiple areas of linguistics. The strengths of the CL methodology, summarized by Bednarek (2009), lie in its use of representative data and empirical and systematic evidence, and it uncovers language features that cannot be discovered in small-scale texts. Biber et al. (1998) also conclude three features of the corpus-based approach that can be recognized as strengthens of CL methodology, namely:

1) It analyzes language in natural texts and provides empirical evidence;

2) It uses an extensive collection of natural texts, also known as corpus, as its analysis basis;

3) It depends on the extensive usage of computers for both quantitative and qualitative analysis.

The features mentioned above indicate that the corpus-based approach enables the conclusion of convincing, frequent, and repetitive linguistics patterns through conducting quantitative analysis and providing solid linguistics evidence. As McEnery and Gabrielatos (2006) point out, the corpus-based approach emphasizes quantitative information, including frequency counts and statistical measures, which allows the researchers to replicate the studies and check the statistical reliability of the analysis results. However, the weakness of CL mainly lies in its negligence of social-cultural contexts and language usage patterns observed from decontextualized examples, especially for large-scale size corpus (Baker, 2023). Due to the collection of large amounts of texts and the unfolding of semantics meaning in texts, researchers may find it unfeasible to provide in-depth interpretations, and 'important features of the context of production may be lost when using such techniques' (Clark, 2007). Such weakness is also reinforced by Bednarek (2009), who points out that "large scale corpus linguistics usually has less to say about context, and the unfolding of meaning in texts."

2.3 Needs for Synergy: Corpus-Based Critical Discourse Analysis

The limitations mentioned above of CDA and CL create a dilemma where the researchers find that CDA methodology only provides unrepresentative data through analyzing limited text samples, and it is unfeasible to conduct in-depth analysis based on large-scale corpora. However, such limitations motivate combining them to form a more robust methodological framework. Stubbs (1997) suggests strengthening CDA by applying a large corpus to generalize reliable and convincing patterns about language use. The corpus-based CDA approach finds a perfect balance to compensate for weaknesses, and such methodological synergy has become increasingly popular in recent studies. Through a meta-analysis, Nartey and Mwinlaaru (2019) make the assumption that corpus-based CDA research is essential in advancing political and social topics since such methodological synergy helps to observe the dynamics and nature of ideologies related to race, gender, and prejudice issues by analyzing large amounts of data. It pinpoints specific political and social domains for in-depth

investigation by narrowing down a large corpus. The synergy of CDA and CL mainly focuses on the grammatical and lexical choices that carry ideologies (Orpin, 2005). In this study, the author analyzed news reports related to the Israel-Hamas war by focusing on the first dimension of CDA, the linguistics features, and more specifically, vocabulary choices and patterns since the commonly used CL methods that can be efficiently applied to CDA include keywords, frequency, dispersion, collocation, clusters and concordance analysis (Baker et al., 2008). From a lexical perspective, the author hopes to explore the in-depth ideologies hidden in the media discourse, thus further verifying that the synergy of CL and CDA provides an efficient methodological framework.

3. Methodology

3.1 Corpus-Building and Reference Corpora

3.1.1 CNN, Al-Jazeera, and SCMP as Target Corpora

Instead of using multi-million-word and pre-built corpora such as the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC), the author built his corpora for this study. Such purpose-built 'do-it-yourself' (DIY) corpora

focus on investigating specific research questions and domains (Almut, 2010). In this study, the interest in finding media outlets' political stances towards the Israel-Hamas war and their ideologies carried in the news reports motivates the author to build related corpora. The texts collected for this study were news reports from Al-Jazeera English (the first global English-language news channel to be headquartered in the Middle East), CNN (the first and biggest all-news television channel in the United States), and China Daily (the most prominent English portal in China). Generally, Arab media are considered as pro-Palestine for cultural and political reasons, American media are considered as pro-Israel since they are strategic allies and Chinese media are considered somewhere in between for China's peaceful diplomatic policy. Each media outlet represents different political stances, which leads them to use different language to portray this war and convey their ideologies. By selecting these three media outlets, the author could conduct a horizontal comparison to examine how they use different language patterns to express different political stances and ideologies.

Table 1. General Statistics of the Israeli-Hamas War Study Corpora

	Al-Jazeera English	China Daily	CNN
Number of Words	643,304	160,044	783,090
Number of news reports	1,192	287	1,070
Number of days collected	248	248	248
Average new reports per day	4.806	1.157	4.319
Average report length by words	540	558	732

The author used Factiva, a business intelligence platform that includes content from 33,000 news, data, and information sources from 200 countries and 32 languages, to collect news reports related to the Israel-Hamas war from 7 October 2023 (the beginning of the war) to 11 June 2024 (an UN-backed ceasefire plan was purposed). The author focused on the political stances of the media outlets. Thus, he only collected news reports tagged as "Global/World Issues" and "Politics/International Relations" subjects. For accurate collection, the author used "Gaza," "Hamas," "Israel," and "Palestine" as search terms and only selected news reports

with the above terms in their headline and lead paragraph. The collected texts were sorted by timeline, with the oldest reports at the beginning of the corpus and the newest at the end and cleaned to build the corpus. As shown in Table 1, there were 1,192 news reports collected from Al-Jazeera, 287 from China Daily, and 1,070 from CNN over the same period. The total number of words was 643,304 in the Al-Jazeera English corpus, 160,044 in the China Daily corpus, and 783,090 in the CNN corpus (see Table 1 for more details).

3.1.2 Reference Corpus

Mautner (2009) argued that for critical discourse

analysts who use smaller study corpora, validating the interpretations by checking them against comparative evidence collected from larger reference corpora is essential. Keywords, which inflect the initial interpretation of the study corpus and lay the foundation for further critical discourse analysis, are generated based on the comparison with the word lists of the study corpus and reference corpus. Keyword results are highly likely influenced by the size and contents of the reference corpus researchers chose. According to the study conducted by Sardinha (2000), utilizing a large reference corpus was unnecessary since employing a corpus more significant than five times the size of the study corpus did not appear to affect the keywords yielded. Thus, more attention and resources should be emphasized on the contents compiled for the reference corpus. In this study, to better calculate and identify keywords of study corpora, each one has its corresponding reference corpus collected from the same source and built based on the same standard of study corpora. Reference corpora contained all news reports tagged with "Global/World Issues" and "Politics/International Relations" subjects over the same period. The total number of words was 2,814,085 in the Al-Jazeera reference corpus, 1,882,321 in the China Daily reference corpus, and 5,722,839 in the CNN reference corpus.

3.2 Research Toolkits

In this study, the author used AntConc, a freeware, multiplatform tool for carrying out corpus linguistics research developed by Anthony (2005). AntConc is a Windows-user-friendly application that provides multiple toolkits for linguistics research, including KWIC (Key-Word-In-Context) tool, plot tool, file tool, cluster tool, N-gram tool, collocate tool, word list tool, keyword list tool and word cloud tool. It offers a free platform for users to conduct in-depth corpus analysis by introducing corpus methods and data-driven language evidence. In this study, the author used five toolkits mentioned above—keyword list tool, word list tool, collocate tool, KWIC tool, and plot tool to analyze the study corpora step by step.

3.2.1 Keyword List

This tool displays words that are unusually frequent in the study corpus compared to the words in the reference corpus, allowing the user to identify characteristic words in the corpus

(Anthony, 2011). A number of corpus-based ideology studies followed Stubbs's (1996) approach to keyword analysis, which has become an indispensable tool for discourse analysis and is being applied to identify theme words or recurrent topics of the texts in the target discourse domain (Egbert & Biber, 2019). Therefore, through keyword analysis, the author could select words that represent the themes of each corpus, which could be further analyzed to reveal the ideologies of texts.

Given that this study only focused on the semantic meaning of language, all grammatical words were excluded. Each study corpus had its corresponding reference corpus to ensure the keywords could be representative. After generating three keyword lists, the author manually categorized the top 100 keywords of each corpus into three categories: participants of the war, political aspects of the war, and militant aspects of the war. Keywords in each categorized table were ranked in the order of their keyness. Words that were key in all three corpora were in bold type, and those that only appeared in one corpus were underlined. The author then analyzed each table and selected keywords that could carry each media outlet's political stances and ideologies for further analysis.

3.2.2 Word List

According to Gries (2010), from the corpus linguistics perspective, one of the most fundamental statistics is the observed absolute frequency of some phenomenon or frequency of word occurrence. From the critical discourse analysis perspective, lexical choice is ideologically based, consciously and unconsciously principled, and systematic (Sheyholislami, 2001). Therefore, the high frequency of certain words highlights the theme of the texts, thus conveying the ideologies behind them to the target readers. By contrast, the relatively low frequency of certain words also conveys the corresponding ideologies since the texts try to downplay the importance of specific themes. With the help of AntConc, the authors were able to analyze large-scale corpora and focus on the frequency of specific keywords since they can better accentuate the themes of the news reports and convey the ideologies of Al-Jazeera, CNN, and China Daily.

The word list tool counts all the words in the corpus and presents the corresponding lexical

frequency. In this step, the author only focused on the frequency of selected keywords. Given that these three corpora have different sizes, it is necessary to compare the importance of target words by converting raw frequency into normalized frequency (Anthony, 2011). The basis of the norming standard was 100,000 words, as all three study corpora contained less than one million words. By comparing the normalized frequency results, the authors could make the comparison analysis more reasonable and draw a more comprehensive conclusion.

3.2.3 Collocate and KWIC

Firth (1957) famously summarized that "you shall know a word by the company it keeps," which can be achieved through collocation analysis of a large-scale corpus nowadays. Collocation is "a lexical relation between two or more words which tend to co-occur within a few words of each other in running text" (Stubbs, 2001, p. 24). Since critical discourse analysis focuses on revealing the ideology coded implicitly behind the overt proposition (Fowler, 1996, p. 3), collocation can be a valuable research method since it shows how a word acquires meanings by combining with other words in different contexts.

The collocate tool allows users to search for collocates of the search term and investigate non-sequential patterns in language (Anthony, 2011). In this study, the author focused on the collocation of selected keywords, aiming to find out how exactly Al-Jazeera, CNN, and China Daily used words to company keywords. More specifically, the authors paid extra attention to both commonly and specifically used words in three study corpora, which contained information that could be interpreted from multiple perspectives in different contexts. With a span of five words to the left and five to the right of the node words, all collocated words are ranked by the collocation strength as measured by the log-likelihood statistic. Since the authors only considered words with semantic meanings for later critical discourse analysis, all grammatical words were excluded. In addition, the KWIC tool was used, which allowed the authors to see how selected keywords were commonly used in the original contexts, and corresponding sentences were listed as examples for detailed demonstration and analysis.

3.2.4 Plot

According to Wodak (2012), language contains a power that draws clear boundaries between "us" and "others," and its speakers only speak the language for their various vested interests. Therefore, investigating how words are distributed in the texts could reveal the changing attitude and attention of the media since it only reports and accentuates the theme that is conducive to their interests and fits its ideologies. In this study, the author focused on the distribution of selected keywords since the changing attention on the theme represented by the keywords reflects the changing attention and attitude of the three study corpora, which may reveal their political stances and ideologies.

The plot tool shows search results plotted in a 'barcode' format, allowing users to locate the position where search results appear in target texts (Anthony, 2011). The left edge of the plot is the beginning of the corpus, and the right edge is the end, and the search result is shown as a vertical line within the plot. In this study, news reports collected for all three corpora were organized by time order, which covers all news reports from the beginning of the war. The density of the vertical line in the plot demonstrates the frequency of the search word, and the density change implies the change in the theme and media's attitude. Therefore, by producing a plot showing the distribution of the selected keywords through the text, the authors could observe the frequency change of the keyword's usage as the war proceeded, thus making reasonable speculation on the changing theme.

4. Data Collection and Analysis

4.1 Keywords Categorization and Analysis

Table 2 demonstrates the keywords related to the participants of the Israel-Hamas war, and those words can be further subcategorized under three themes: countries and places involved, organizations involved, and people involved. Locations (like the Gaza Strip and enclave) and cities (like Rafah, Shifa, Jerusalem, and Tel Aviv) constantly appeared in all three study corpora, indicating that they were potential targets in the Israel-Hamas war. "Hospital(s)" is also in the top 100 keywords list of all three study corpora, which is disturbing since its usage either means a large number of casualties or the hospital was the militant target. Noticeably, "ICJ" (International Court of Justice) only appears in Al-Jazeera corpus, which indicates that Al-Jazeera may quote resolutions

decided by "ICJ" to call for international attention, seek justice, and accentuate the severity of the war. "IDF" (The Israel Defense Forces) only shows up in the CNN corpus, indicating that CNN may constantly convey its political stances and ideologies from an Israeli perspective, which is expected considering the particular relationship between America and Israel. For keywords related to people who were involved in the war, it is worrying to see that "civilian," "captive," and "hostages" appear in all three corpora since they were directly impacted by the war and suffered. However, which people these words referred to were unknown to the author since either Israeli or Palestinian or both are possible collocations.

Table 2. Keywords related to the participants of the Israel-Hamas war

	Al-Jazeera	China Daily	CNN
Countries and places involved	**Israel**, **Gaza**, **strip**, **hospital**, **Palestine**, **Lebanon**, **enclave**, **Iran**, **Rafah**, **Egypt**, **Qatar**, southern, **middle**, **Jerusalem**, **Tel Aviv**, bank, **East**, **Syria**, <u>Shifa</u>, <u>Jordan</u>, **ground**, **Saudi**, **hospitals**, West, Maghazi, United (States), <u>Aqsa</u>, Arabia	**Israel**, **Gaza**, **Palestine**, **strip**, **Rafah**, **middle**, **Iran**, **East**, **Egypt**, United (States), **Lebanon**, **enclave**, southern, **Saudi**, **Tel Aviv**, **hospital**, **Syria**, **Qatar**, **Jerusalem**, Arabia, <u>Ukraine</u>, **ground**, <u>Cairo</u>	**Israel**, **Gaza**, **Iran**, **Egypt**, **strip**, **Rafah**, **enclave**, **Qatar**, **East**, **Lebanon**, <u>region</u>, **middle**, **ground**, **hospital**, **Tel Aviv**, **Palestine**, <u>regional</u>, **Syria**, <u>northern</u>, **Saudi**, **Jerusalem**, <u>Tehran</u>, bank
Organizations involved	**Hamas**, **Hezbollah**, **Al (Jazeera)**, **UNRWA**, <u>ABU</u>, council, group, <u>ICJ</u>, **United (Nations)**	**Hamas**, **Al (Jazeera)**, **United (Nations)**, council, **UNRWA**, **Hezbollah**, <u>Houthi</u>,	**Hamas**, <u>IDF</u>, **Al (Jazeera)**, **Hezbollah**, <u>ministry</u>, **UNRWA**, group, **United (Nations)**
People involved	**Israeli**, **Palestinian**, **Palestinians**, **Netanyahu**, **civilians**, **Arab**, **Israelis**, **captives**, **Benjamin**, **hostages**, <u>Younis</u>, children, Jewish, **Blinken**, <u>Guterres</u>, refugee, <u>Mohammed</u>, people, <u>Lebanese</u>, **captive**, Biden	**Israeli**, **Palestinian**, **Palestinians**, **civilians**, **Netanyahu**, **Arab**, **hostages**, **Benjamin**, **Israelis**, <u>Islamic</u>, **civilians**, minister, children, Egyptian, Iranian, **Blinken**, Biden	**Israeli**, **Palestinian**, **hostages**, **Netanyahu**, **Palestinians**, **civilians**, **Arab**, **Blinken**, **hostage**, **Israelis**, Jewish, **Benjamin**, **civilian**, minister, Iranian, people, Prime (minister), <u>officials</u>, Egyptian, <u>Gazans</u>, <u>Tlaib</u>, <u>official</u>, refugees

Table 3 listed all keywords related to the political aspects of the war, and those words were further subcategorized under two themes: peace process and global issues. Peace process includes all keywords that describe efforts made and actions taken for achieving peace, demonstrating that all three study corpora had reported peace-seeking news and called for "pausing" the war, "releasing" hostages, and providing "aid" and "supplies." The top keyword in the global issue subcategory is "humanitarian," which strongly suggests that humanitarian issues in the Israel-Hamas war were the top concerns of the three study corpora. Keywords like "genocide," "food," "medical," "famine," and "health" were all consequences of the humanitarian issues arising from the war. Remarkably, the keyword "genocide" was in all three study corpora, but its keyness was different, which means that these three study corpora emphasized the "genocide" theme at different levels. "Genocide" ranked 70th in the keywords list of Al-Jazeera and 52nd in China Daily, but it only ranked 97th in CNN's list, which strongly indicated that CNN may have the intention to weaken the "genocide" related theme. As a severe accusation, media outlets should use "genocide" cautiously since whoever conducts such a crime would be condemned morally and punished legally. The author speculated that Al-Jazeera and China Daily accentuated the "genocide" theme in their new reports, suggesting that they charged this crime upon one side without a

doubt and firmly against such crime. CNN used this keyword more prudently, suggesting that it remained unsure which side to blame or purposely avoid this theme. However, the reason behind such a difference was unknown, so more solid evidence and further in-depth research were needed.

Table 3. Keywords related to the political aspects of the Israel-Hamas war

	Al-Jazeera	China Daily	CNN
Peace process	ceasefire, **aid**, resolution, truce, **pause**, **release**, calling, **supplies**, support	<u>cease</u>, **aid**, resolution, truce, peace, <u>reported</u>, <u>condemned</u>, calling, calls, **release**, <u>talks</u>, <u>pressure</u>, **supplies**, **pause**, <u>end</u>	**aid**, ceasefire, <u>said</u>, **release**, **supplies**, <u>assistance</u>, <u>response</u>, **pause**, <u>negotiations</u>, <u>deal</u>, peace
Global issues	**humanitarian**, **genocide**, international, food, <u>medical</u>, health	**humanitarian**, situation, **genocide**, territory, famine, health	**humanitarian**, international, food, <u>security</u>, territory, situation, <u>diplomatic</u>, **genocide**, famine

Table 4 shows the keywords that refer to the militant aspects of the Israel-Hamas war, which are further categorized into three types. Descriptions of the militant activity include keywords that describe the definition of the Israel-Hamas war, time, weapons, and militant forces. Compared to Al-Jazeera and CNN, China Daily seemed to prefer to use "conflict" and "crisis" to define the militant activity between Israel and Hamas, which were less severe words than "war." Such cautious lexical choice reflects China Daily's cautious attitude in defining the Israel-Hamas war. Notably, the highly negative words "terror" and "terrorist" were only on the list of CNN, which strongly indicates that CNN preferred to define the war as a "terror" attack initiated by a "terrorist." Such lexical choice difference indicated that compared to the coverage of Al-Jazeera and China Daily, CNN coverage emphasizes the terror-related theme when reporting the Israel-Hamas war, and its readers naturally were more prone to blame one side of the participants as terrorists. How these three study corpora defined this war demonstrated their political stances and ideologies, and their lexical choices highly influenced their readers' viewpoints of this war.

Keywords in the militant action subcategory describe how this war proceeded. "Killed," "attack(s)," and "strike(s)" were commonly used keywords in all three study corpora, suggesting the ferocity of the war. Al-Jazeera also used "occupied" and "occupation" to indicate that one participant in the war was invaded and controlled by another, thus further implying the unjust nature of the war. As for the third subcategory, it is evident that CNN devoted much less attention to describing the dreadful outcomes of the war, while Al-Jazeera and China Daily depicted a more detailed war scene, which implies that those two media outlets condemned the war crime and called for international attention toward this war.

Table 4. Keywords related to the militant aspects of the Israel-Hamas war

	Al-Jazeera	China Daily	CNN
Descriptions of this militant activity	**war**, **October**, **trucks**, <u>camp</u>, **conflict**, immediate, <u>fighters</u>, <u>rubble</u>, fuel	**conflict**, **oct**, immediate, military, offensive, <u>crisis</u>, <u>Sunday</u>, **war**, **trucks**, militants, <u>Saturday</u>, <u>violence</u>, <u>hostilities</u>	**war**, **conflict**, **October**, military, <u>forces</u>, **trucks**, <u>terror</u>, militant, militants, offensive, <u>terrorist</u>, fuel
Militant action	**killed**, **attacks**, **attack**, bombardment, <u>occupied</u>, <u>besieged</u>, <u>air</u>, <u>occupation</u>, <u>bombing</u>, killing, **strikes**, siege, **strike**, <u>assault</u>	**killed**, **attacks**, **attack**, fighting, **strikes**, airstrikes, killing, crossing	**attack**, **killed**, **attacks**, crossing, fighting, airstrikes, **strike**, bombardment, siege, **strikes**

Outcomes	<u>wounded</u>, toll, escalation, **displaced**	<u>casualties</u>, escalation, **displaced**, toll, <u>deaths</u>, <u>death</u>	**displaced**

In sum, through keyword analysis, the author could make his initial assumption of these three study corpora's attitudes towards this war: China Daily used less aggressive words to describe the war and was the most conservative compared to the other two, who were relatively radical regarding lexical choices. However, keyword analysis is insufficient to make further conclusions and analysis regarding their political stances and ideologies. As mentioned above, the evident differences in terms of the usage and the implicit information of "genocide" may be worth exploring. The usage of genocide indicates the initiator of this conflict and arouses people's anger against such unlawful acts of violence. It also suggests the victim of the war and resonates with the public's sympathy. The study corpora, with different ideologies and political standpoints, may use different language to depict and distinguish "initiators" and "victims" of this conflict. Therefore, the author selected the keyword "genocide" for further research, hoping to

explore more related language usage patterns to verify his speculation.

4.2 Lexical Frequency Data and Analysis

The keyness statistic only demonstrates the importance of "genocide" in three study corpora respectively; for horizontal comparison, the normalized frequency of "genocide" should be considered. The statistical frequency information on keyword genocide can be observed in Table 5, thus showing the relative importance of the theme in the coverage of the Israel-Gaza war across Al-Jazeera, China Daily, and CNN, and further revealing their respective ideologies. The most obvious observation in the normalized frequency data is that genocide occurs much more frequently in the Al-Jazeera corpus—1.398 times higher than the China Daily corpus and 2.155 times higher than the CNN corpus. Based on the data, it is safe to say that Al-Jazeera mentioned the genocide theme more frequently than China Daily and CNN.

Table 5. Frequency of genocide in Al-Jazeera, China Daily, and CNN corpora

Corpus	Total number in the corpus	Normalized frequency (per 100,000 words)
Al-Jazeera	354	550.284
China Daily	63	393.642
CNN	200	255.398

As mentioned before, "genocide" indicates the victim and initiator of this war, which stimulated the readers to take a political stance, influenced the public's attitude towards the participants in the event, drew more attention and sympathy to the vulnerable side, and called for condemnation towards the opposite. Al-Jazeera highlighted the genocide topic, while such a topic was weakened on CNN, which further influenced their target readers' viewpoints of the Israel-Hamas war. On the other hand, China Daily maintained its conservativeness and demonstrated no tendentious standpoints. However, lexical frequency only reflected the relative importance

of the target keyword, and more detailed evidence was needed to draw a more complete picture.

4.3 Collocation and Analysis

Table 6 demonstrated words collocated with "genocide" in Al-Jazeera, China Daily, and CNN, and all were ranked based on their collocation strength. From the list, the author could initially speculate the potential initiator or victim of the genocide in Al-Jazeera (including Gaza and Israel), China Daily (including Palestinians, Israel, and Gaza), and CNN (including Jews, Jewish, and Palestine).

Table 6. Collocation list of *genocide* in Al-Jazeera, China Daily, and CNN corpora

Corpus	
Al-Jazeera	convention, committing, prevent, case, Africa, acts, crimes, risk, ICJ, Gaza, accused, south, actions, grave, crime, complicity, charges, court, unfolding, Israel, Ione, hearing, plausibly, plausible, complicit, committed, accusing, commit, engaging, assisted, textbook, patterns, cleansing, signatories, allegedly, AI, ethnic, incitement, ecocide
China Daily	committing, convention, crime, complicit, Africa, Palestinians, Israel, accusing, case, alleged, prevent, accuses, allegations, amounted, Gaza, acts, south
CNN	convention, prevent, Jews, committing, violate, calling, Jewish, calls, signatory, Joe, act, defines, supporting, case, Africa, ruling, committed, incitement, crime, allegations, violating, punish, acts, plausible, ordering, ICJ, accusing, punishment, Palestinian, ho, afoul, ensure, millennia, obliged, foul, stops, complicit, necessarily, south, Lieblich, constitute, people

Through the examples shown in Table 7, by repeatedly collocating "Gaza" and "Israel" with "genocide," Al-Jazeera depicted Israel as the one who started the genocide and Gaza as the victim. The same collocation pattern also appeared in the coverage of China Daily (see example in Table 8), condemning Israel for committing genocide in Gaza. Such collocation choice convinced the author that Al-Jazeera and China Daily expressed their concerns in Gaza and condemned Israel for carrying out such crime, thus demonstrating pro-Palestine political stances to some extent.

Table 7. Examples of *Gaza* and *Israel* collocated with *genocide* in Al-Jazeera

… Albanese of being an "accessory to **genocide** in **Gaza**" and asked the International Criminal Court…

… ongoing case accusing Israel of state-led **genocide** in **Gaza** after the October 7 Hamas attacks on…

… South Africa accused Israel of carrying out **genocide** in **Gaza** and demanded that the court order an…

… filed the lawsuit at the end of December, accusing **Israel** of **genocide** in its war on Gaza and seeking a halt…

… adding that South Africa's accusation against **Israel** of **genocide** in Gaza could only happen in a world…

Table 8. Example of *Gaza* and *Israel* collocated with genocide in China Daily

… Israel for committing what amounted to "gross **genocide**" in **Gaza**, the Biden administration continued…

…. sales to Israel, highlighting warnings of "**genocide**" in its war in **Gaza**. The resolution marked the…

… ICJ to rule on whether there is an ongoing **genocide** in **Gaza** and to clarify the duties of all states to…

… the ICJ, where the country accuses **Israel** of committing **genocide** against Palestinians. On the Israel-…

… United Nations' top court accusing **Israel** of **genocide** against Palestinians in Gaza. Others, including…

CNN revealed its opposite political stance by reporting different stories. "Jews" and "Jewish" were used by CNN to collate with "genocide" (see examples in Table 9), which seemed to

indicate that the victims of genocide might be Israel. By reviewing the suffering history of Israel and depicting the war from the Israeli perspective, CNN's language contained the opposite ideology that implied Israel was the victim of genocide.

Table 9. Examples of *Jews* and *Jewish* collocated with *genocide* in CNN

… leaders explicitly said that calling for the **genocide** of **Jews** would necessarily violate their code of conduct…
… one of history's most vicious crimes — the Nazi **genocide** against **Jews**. But everyone understood the context…
… me be clear: Calls for violence or **genocide** against the **Jewish** community, or any religious or ethnic group…
… the phrase is seen by many as a call for **Jewish genocide** in Israel and has been used by Hamas as a rallying…
… left by millennia of anti-Semitism and **genocide** of the **Jewish** people, so in this moment, we must be crystal…

Surprisingly, "Palestinian" also collocated with genocide in CNN's reports, and related examples were demonstrated in Table 10. Such an inconsistent collocation pattern indicated the change in CNN's attitude, which may be a response to international concerns because as the Israel-Hamas war continued, it caused more civilian deaths than ever before. Based on the author's speculation, CNN may have used "Jewish" and "Jews" to collocate with "genocide" at the early stage of the war but then changed its attitude and regarded "Palestinians" as victims of the "genocide" as the war proceeded.

Table 10. Examples of *Palestinian* collocated with *genocide* in CNN

… Sunday defended accusing Israel of **genocide** against the **Palestinian** people and advocating for cuts to the US…
… to draw attention to the ongoing **genocide** of the **Palestinian** people and the people of Gaza," Madbak, 29, …
… the fighting, ending the killing, ending the **genocide** of the **Palestinian** people in the Gaza Strip." "The reason…
… previously accused Biden of supporting a **Palestinian** "**genocide**" and warned Americans will remember how…
… the president in November of supporting a **Palestinian** "**genocide**," straining her relationship with Biden. Her…

Notably, the collocation strength of "complicit" with "genocide" in Al-Jazeera and China Daily could be interpreted that some countries and organizations allowed, connived with, and even supported Israel to commit genocide in Gaza (see examples in Table 11). By accusing Western countries of assisting Israel and allowing genocide to happen in Gaza, Al-Jazeera and China Daily uncovered their hypocrisy, thus further demonstrating their pro-Palestine political stance.

Table 11. Examples of *complicit* collocated with *genocide* in Al-Jazeera and China Daily

… after the Holocaust, Germany is accused of being **complicit** in an alleged Israeli **genocide** in Gaza. Nicaragua…
… halt arms sales to Israel, saying it could make Britain **complicit** in **genocide** in Gaza. Some

opposition political…

… Joe Biden and other senior officials of being **complicit** in Israel's "**genocide**" in Gaza has been dismissed by a…

… military assistance and material to Israel may render the UK **complicit** in **genocide** as well as serious breaches…

… ongoing support could make the Biden administration **complicit** in **genocide**, earning US President Joe Biden…

Similarly, "complicit" was also used to collocate with "genocide" in the CNN corpus, though the same lexical pattern contained different information. Given that "complicit" was not a strong collocation of "genocide" in CNN, only limited examples were shown in Table 12. Instead of directly showing its own attitude, CNN adopted the strategy of implying some countries were complicit in genocide by quoting others' speeches. In addition, the example also demonstrated that in some reports, CNN still pictured Hamas as the initiator of genocide by implying that some countries were helping Hamas to carry out genocide in Israel.

Table 12. Examples of *complicit* collocated with *genocide* in CNN

… the incident obtained by CNN that he would "no longer be **complicit** in **genocide**" and that his suffering was…

… state of Israel and weapons manufacturers that are **complicit** in the **genocide** that's happening in Gaza." Fisher…

… Eylon Levy said Pretoria is "criminally **complicit** with Hamas' campaign of **genocide** against our people." He…

It was worth mentioning that only Al-Jazeera used "cleansing" to collocate with "genocide." "Cleansing" indicated the lack of will to stop genocide, resulting in massive increases in civilian deaths and undermining international legal obligations to acknowledge genocide (Blum et al., 2008). Al-Jazeera used "cleansing" to express its concerns that Israel may continue to commit genocide in the future with no intention to stop, which was a much more serious accusation than genocide; examples were as follows:

Table 13. Examples of *cleaning* collocated with *genocide* in Al-Jazeera

… the Zionist aggression, crimes of **genocide** and ethnic **cleansing** against our people in the Gaza Strip…

… its coalition are calling for ethnic **cleansing** and even actual **genocide**. They are the ones who harm the…

… Israel is conducting in Gaza a sadistic ethnic **cleansing** campaign, a **genocide**, aimed at ridding the Strip…

… to putting an end to Israel's blatant ethnic **cleansing** efforts and **genocide**. The views expressed in this…

… 2023 Nakba. It is a terrifying act of **genocide** and ethnic **cleansing**, unlikely to be the last. There are still…

To sum up, Al-Jazeera and China Daily demonstrated pro-Palestine political stances by blaming Israel for carrying out genocide in Gaza, while CNN showed a seemingly balanced stance.

4.4 Lexical Distribution and Analysis

How the keyword "genocide" was distributed in Al-Jazeera, China Daily, and CNN was also a question worth exploring since it reflected the

change of political stances and attitudes. Based on Figure 1, it is noticeable that all three study corpora used "genocide" less frequently at the beginning of the Israel-Hamas war because they maintained a certain conservativeness at first. However, they increased the usage of "genocide" as the war proceeded and became more violent, which could be verified by the changes in the density of the vertical line. Differently, even at the early stage of the war, Al-Jazeera used the keyword "genocide" much more frequently than the other two study corpora.

The author also located the first time the word "genocide" appeared in three study corpora and traced it back to the original text; the original texts were as follows:

Al-Jazeera: Hamas accuses Israel of carrying out a "genocide" in Gaza.

China Daily: The Arab League and the African Union said in a joint statement on Sunday that it could "lead to a genocide of unprecedented proportions", the Arab News and various media reported.

CNN: … Biden said. "This attack has brought to the surface painful memories and the scars left by millennia of anti-Semitism and genocide of the Jewish people, so in this moment, we must be crystal clear, we stand with Israel."
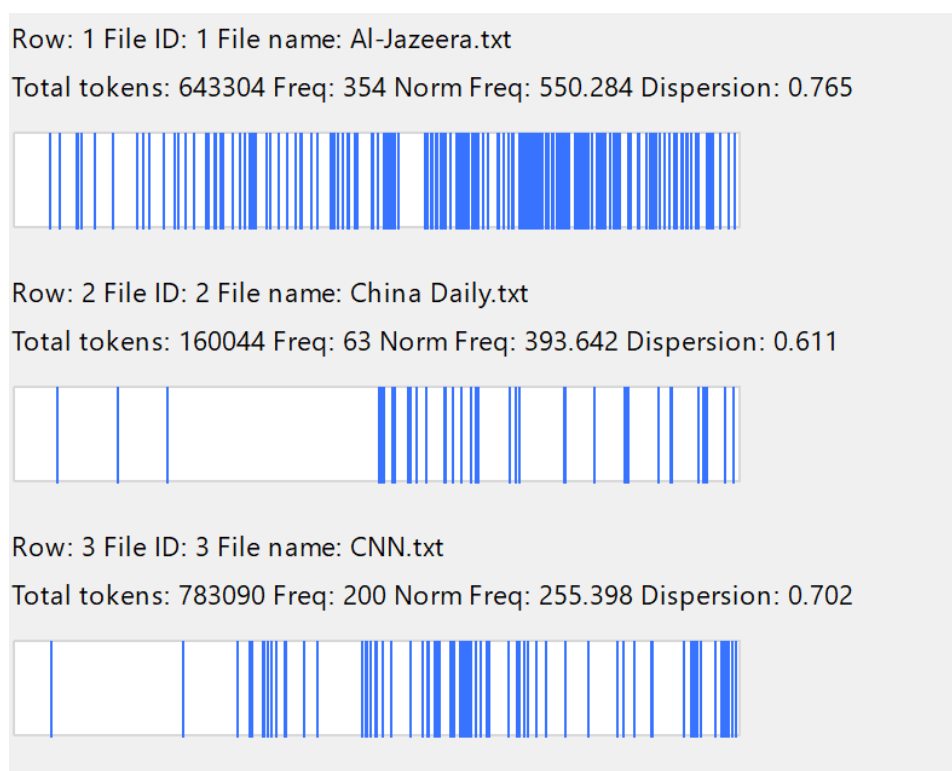

Figure 1. Distribution of keyword *genocide* in Al-Jazeera, China Daily, and CNN

Al-Jazeera accused Israel of conducting genocide towards Gaza at the very early stage of the war. In contrast, CNN reported this event in the opposite way by quoting President Biden's speech. China Daily, on the other hand, indirectly indicated that Israel was responsible for the genocide by quoting Arab news reports, which again reflected its cautiousness. In addition, such distribution trend indicated the attitude change of all three study corpora as the war became more violent and more casualties were reported — the severe humanitarian issues

caused by the war concern the international community. Increasing news reports started to highlight the genocide theme as the war proceeded. Interestingly, CNN, though affirmingly claimed that Hamas carried out genocide in Israel at first, changed its attitude and blamed Israel as the war proceeded. Such changes answered the collocation contradiction and verified the speculation mentioned above.

5. Conclusion

The results and analysis in Section 4 answered

how the same topic (the 2023 Israel-Hamas war) was presented differently on Al-Jazeera, China Daily, and CNN via different language patterns. The initial keywords analysis showed which themes were emphasized in the study corpus, and the author selected the most noteworthy word, "genocide," as the target keyword for in-depth analysis; normalized lexical frequency demonstrated and compared the importance of the theme "genocide" in different corpora; collocation analysis and lexical distribution further revealed how the same keyword used differently by tracing back to the original text. Overall, through such analysis framework, the author concluded that Al-Jazeera took a strong pro-Palestine political stance by condemning Israel's genocide activity, revealing the hypocrisy of Western countries, and accentuating the severity of the war; CNN took a wavering political stance by claiming the war was a terror attack initiated by Hamas, and blaming Hamas for carrying out genocide in Israel but later changed its attitude and accused Israel instead; China Daily maintained a consistent attitude towards this war and took a more balanced political stance, with slight political tendency toward Palestine.

This study only took one common keyword in all three study corpora for in-depth analysis and demonstration, and a more comprehensive study could be further conducted. Although the current limited data and limited analysis are insufficient to provide a more comprehensive conclusion, the analysis procedure demonstrated in this study suggests the potential and feasibility of combining the corpus linguistics method with critical discourse analysis to analyze language usage in news reports. This study is also the first step in the future large-scale project to provide an operative pattern of analyzing texts in social and political contexts with the corpus-based critical discourse analysis methodological framework.

References

Almut, K. (2010). Building small specialized corpora. In *The Routledge handbook of corpus linguistics* (pp. 66-79). Routledge.

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.* (pp. 729-737). IEEE.

Anthony, L. (2011). AntConc (Windows, Macintosh OS X, and Linux). Recuperado de: http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3, 2.

Baker, P. (2023). *Using corpora in discourse analysis*. Bloomsbury Publishing.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & society*, *19*(3), 273-306.

Bar-Siman-Tov, Y. (1998). The United States and Israel since 1948: a "special relationship"? *Diplomatic History*, *22*(2), 231-262.

Bednarek, M. (2009). Corpora and discourse: A three-pronged approach to analyzing linguistic data. In *Selected proceedings of the 2008 HCSNet workshop on designing the Australian National Corpus* (pp. 19-24). Somerville, MA: Cascadilla Proceedings Project.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Blum, R., Stanton, G. H., Sagi, S., & Richter, E. D. (2008). 'Ethnic cleansing' bleaches the atrocities of genocide. *European journal of public health*, *18*(2), 204-209. https://doi.org/10.1093/eurpub/ckm011

Breeze, R. (2011). Critical discourse analysis and its critics. *Pragmatics. quarterly publication of the international pragmatics association (IPrA)*, *21*(4), 493-525.

Clark, C., & DE BOHUN, M. A. R. Y. (2007). A war of words: a linguistic analysis of BBC embed reports during the Iraq conflict. *In Discourse Analysis and Contemporary Social Change* (pp. 119-140). Peter Lang.

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77-104.

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

Fairclough, N. (2001). Language and power second edition. London: Longman. and power relations in Nigerian newspaper headlines. *Nebula*, *4*(1), 218–245.

https://doi.org/10.4324/9781315838250

Fairclough, N. (2013). Critical discourse analysis. In *The Routledge handbook of discourse analysis* (pp. 9-20). Routledge.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10-32.

Fowler, R. (1996). *Linguistic criticism*. Oxford University Press.

Gries, S. T. (2010). Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches*, *66*, 269–291.

Henry, F., & Tator, C. (2002). *Discourses of domination: Racial bias in the Canadian English-language press*. University of Toronto Press. https://doi.org/10.1093/oso/9780192892614.001.0001

Mautner, G. (2009). Checks and balances: How corpus linguistics can contribute to CDA. *Methods of Critical Discourse Analysis*, 154–179.

McEnery, T. (2019). *Corpus linguistics*. Edinburgh University Press.

McEnery, T., & Gabrielatos, C. (2006). English corpus linguistics. *The handbook of English linguistics*, 33-71.

Nartey, M., & Mwinlaaru, I. N. (2019). Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis. *Corpora*, *14*(2), 203-235.

Orpin, D. (2005). Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. *International journal of corpus linguistics*, *10*(1), 37-61.

SARDINHA, Tony Berber. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? In: *The Workshop on Comparing Corpora*, p. 7-13.

Sharrock, W. W., & Anderson, D. C. (1981). Book review: Language, thought and reality, again. *Sociology*, *15*(2), 287-293.

Sheyholislami, J. (2001). Critical discourse analysis.

Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). *British studies in applied linguistics, 12*, 100-116.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. John Wiley & Sons.

Van Dijk, T. A. (2015). Critical discourse analysis. *The handbook of discourse analysis*, 466-485.

Van Dijk, T. A. (2015). Critical discourse analysis. *The handbook of discourse analysis*, 466-485. https://doi.org/10.1002/9781118584194.ch22

Widdowson, H. G. (1995). Discourse analysis: a critical view. *Language and literature, 4*(3), 157-172.

Wodak, R. (2012). Language, power and identity. *Language Teaching*, *45*(2), 215–233. doi:10.1017/S0261444811000048