# *A Corpus-based Comparison of Lexical Features in English News: China Daily and CNN's Reports about Russia-Ukraine Conflict as Examples*

**Ruiyao Wang**[1,a,*]

*[1]School of Foreign Language, Zhejiang University of Finance and Economics, Qiantang District, Hangzhou City, Zhejiang, China*
*a. m15958042788@163.com*
**corresponding author*

***Abstract:*** The news plays a crucial role in disseminating information to the public. The use and preferences of vocabulary in news articles reflect the language proficiency, attitudes, and positions of news writers. Therefore, this study examines the differences in vocabulary characteristics between China Daily, a Chinese news organization, and CNN, a US news organization, in their news writing on the Russia-Ukraine conflict. Through corpus-based research methods, the study aims to explore the differences in lexical features and provide insights for journalists and English writing learners in terms of vocabulary selection and comprehension when writing and reading English news about the Russia-Ukraine conflict and related war and conflict events.

***Keywords:*** Lexical Features, Corpus, English News, Russia-Ukraine Conflict

## 1. Introduction

### 1.1. Research Background

Language style refers to the way in which individuals choose to express themselves through speaking or writing based on the circumstances. As an objective linguistic phenomenon, it is shaped by the differentiation of social domains and the diversification of communicative functions [1]. Media Stylistics, to be more specific, is a linguistic analysis that focuses on examining expressions in media discourse [2]. Among these, vocabulary is widely considered the most active factor in language [3], forming the basis and the core of language change [4].

Previous studies have shown notable variations in the lexical features of English news reports when covering the same event in different countries. Numerous studies have investigated the media discourse differences between news agencies in English-speaking and non-English-speaking countries. Nowadays, several countries have established their international news agencies, which primarily report news in English for the rest of the world. China Daily, CGTN, and CNS (China News Service) are typical examples of such news agencies. Thus, comprehending the fundamental principles of expressing information through precise word selection is critical for non-native English-speaking journalists and readers in conveying and understanding news more effectively.

Given that such understanding requires the examination of a broad range of media contents, a computer-based research method that can process large quantities of texts, is suitable for this study. Corpus-linguistic studies aim to reveal the relationships and similarities among specific lexical, syntactical, or pragmatic terms within certain text collections [5]. As a combination of both quantitative and qualitative analysis tools, corpus-linguistic studies are particularly favored in stylistic analysis. They are also widely applied in teaching, translation, and discourse analysis.

The Russian-Ukrainian Conflict is a newsworthy international event that has garnered significant media attention for an extended period. As a time-sensitive topic, it is a fitting subject for comparing English news vocabulary, as there is an abundance of written content available from various perspectives.

This study aims to provide an all-encompassing view of the lexical features in news reports and detect differences between reports from native and non-native English reporters. It also aims to provide valuable references for future news writing. To achieve these objectives, this study analyzed the lexical features of news coverage of the Russia-Ukraine conflict by China Daily and CNN, two representative mass media channels from China and the United States. The lexical attributes of the news corpora, such as lexical density, word length, word class, and word frequency, were analyzed using corpus methods.

## 1.2. Research Questions and Objectives

This paper draws on Leech and Short's influential theory on language style, which posits that language “style” can be reduced to “frequency” when viewed from the perspective of the reader, and can thus be measured [6]. Specifically, the theory identifies five categories, namely general nouns, verbs, adjectives, and adverbs. For this purposes, the study focus mainly on the general lexical category, which is concerned about is the vocabulary “simple or complex? formal or colloquial? descriptive or evaluative? General or specific?” [6] and the selected items, including lexical density, word length, word class and word frequency (content word and function word respectively).

Under the theoretical framework of stylistics, this study aims to analyze the lexical features of news reports related to the Russian-Ukrainian conflict using a corpus-based approach and answer the long-standing questions proposed by linguistic Roger Fowler in 1991 how journalists can “present idea with the language which is designed to be unambiguous, undistorting and agreeable to readers [7]”.

Specifically, the study aims to achieve the following objectives: First, it aims to construct two distinct corpora, namely CD (China Daily Russian-Ukrainian Conflict) and CN (CNN Russian-Ukrainian Conflict), and conduct an in-depth analysis of their lexical features including word density, word length, word class, and word frequency, individually. Secondly, it seeks to compare the lexical features of the two corpora with the reference Corpus of Contemporary American English (COCA) to identify the common lexical features shared by news reports. Thirdly, it endeavors to identify the factors contributing to the discrepancies in the lexical features of the two corpora. Lastly, it aims to provide instructive recommendations to non-native English journalists, writers, and students, based on the study's findings, on how to enhance their writing skills.

Questions that will be discussed in the study are as follows:

1. What are the lexical features, including word density, word length, word class, and word frequency, in both the CD and CN corpora separately?

2. When compared with the COCA, what are the common lexical features shared by news reports in the CD and CN corpora?

3. What factors contribute to the differences in the lexical features of the two corpora?

The study is structured into five distinct sections. The introduction section commences by offering definitions of language style and media stylistics, accentuating the significance of vocabulary within the realm of style analysis while briefly elucidating the research background and objectives. The subsequent section, the literature review presents an overview of relevant studies conducted both domestically and internationally, organized based on their targeted subjects and objectives. The third section outlines the methodology employed, elucidating the process of data collection and analysis, along with the research tools utilized. The fourth chapter, the main body of the thesis, presents the results and comparisons, as well as their practical implications. Finally, the conclusion section discusses the major findings, implications, and limitations, and proposes directions for future studies.

The purpose of this study is to analyze the lexical features of news writing using a corpus-based approach. Specifically, the study focuses on comparing the lexical features of news reports from China Daily and CNN. By using a quantitative method, the study aims to extend existing research on lexical features in news writing and contribute to our understanding of the unique features of news writing in English.

From a practical perspective, this study holds significant implications for journalism in our increasingly globalized world. Given the growing importance of English as a lingua franca, the ability to report news in English has become indispensable for journalists. By providing a guideline for the use of words in English news, especially in global affairs like the Russian-Ukrainian conflict, this study can inform and improve journalistic practices. Furthermore, language learners and students seeking to improve their use of lexical terms in English writing can use this study as a guide to select appropriate words and improve their vocabulary mastery.

## 2. Literature Review

This section reviews previous studies and research abroad and at home. Both are categorized based on their research subjects.

### 2.1. Related Studies in China

Previous quantitative studies on the lexical features of English news can be classified into three categories based on their subject matter:

First, the lexical features of the vice text in news discourse, including news headlines and leads, are one of the commonly seen subjects in previous studies. For example, Weng [8] conducted a statistical study on 2009 English news headlines of the New York Times and Yahoo News in five aspects: word length, word frequency, lexical density, word class, and word formation, and used Brown's database as the standard to conclude that the headline vocabulary tends to be shorter and more concise, with high diversity and a high proportion of nouns. In the 2018 study, she again added the conclusion that function words were omitted and "say" words were used frequently [9]. The study by Huang and Liu [10] focused on the news leads of Global Times and The New York Times and counted three indicators of TTR, word length, and lexical density at the lexical level, concluding that the lexical terms in the current English news writing (based on information in Chinese) was not rich enough, and nouns and adjectives were underused. At the same time, in some studies, although the news headlines and leads are not separated from the news body during the analysis, they are also particularly mentioned in the studies. For example, Zang [11] mentions the phenomenon that more proper nouns have been discovered in news headlines and leads in his overall discussion of stylistic features. The study of the lexical features of news headlines and introductions has implications for improving Chinese-English compilation, readability, and journalists' writing. The four studies cited above all aim to conclude the perspectives of readers' comprehension, English teaching, and the practical value of news writing.

The second category of studies is based on the media discourse generated by different news agencies. One of the frequent methods in past studies of lexical features in English news is to quote or collect from the news articles of one or two news agencies. For example, Shi [12] studied the stylistic features of China Daily by building a corpus based on the text of China Daily and comparing it with COCA (Corpus of Contemporary American English). In addition, building two separate corpora using texts from Chinese and foreign news agencies is also a way: by comparing the stylistic features between Xinhua and BBC News, research concluded that Xinhua English news had less lexical diversity compared to BBC news. However, the average length of words is longer, and seven to twenty-letter words appeared more frequently [13]; by comparing the English version of People's Daily News with CNN News, it is concluded that the former has lower lexical diversity and complexity, but high lexical density which implies that certain words have been used frequently [14] ; Lv [15] collected 338 English news reports broadcast by China Radio International from March 2020 to November 2020 and 303 English news reports broadcast by BBC during the same period. By comparing the average word length, lexical frequency, and STTR, she concluded that the language proficiency of news reporters and listeners between China and the UK is different, and thus the strategy to improve news writing and broadcasting should be adjusted accordingly. The study of stylistic and lexical characteristics of a particular news organization may be influenced by the topics of the news, e.g., news reports on science and technology usually have more specialized terms than those about people's livelihood. However, based on the previous practice, it is more convenient and efficient for researchers to collect corpus by targeting one or two specific news organizations.

Third, the analysis is based on a particular topic or a particular news event. Such an approach aims to use a well-established lexical feature analysis method to analyze the texts about current or specific topics, mostly social issues. For example, economic news and business English. According to Hu [16], at a macro level, the lexical features of English news in the economic field include significant word accuracy and a large number of specialized or innovative concepts. At the micro level, verbs are used most frequently and the diversity of adjectives has increased. In the study of business English, the frequency of proper nouns in the Financial Times and Fortune from 2006 to 2015 was analyzed by Range, which empirically concluded that the minimum vocabulary threshold for a business English study is 4,000 [17]. Yi [18] studied the news reports and commentaries on the 10th anniversary of "911", tried to evaluate whether the high-frequency words could reflect the theme or not, and measured the lexical density of the material.

Based on the above discussion, it is a common and feasible method to construct a corpus and conduct lexical feature analysis by targeting one or two news agencies or certain topics and events.

## 2.2. Related Studies in Global Scale

Studies on lexical features of English news abroad have generally focused on the interaction between lexical features and content dissemination, language differences, and public perception. Based on the subject matter, studies abroad can be divided into three categories:

Firstly, studies on the correlation between lexical features and transmission efficiency in media. Such studies are commonly conducted by summarizing and modeling the lexical features of existing textual materials in order to identify and enhance communication purposes and effects. It is worth noting that the study of news vocabulary has been extended to the "media" as a whole, the latter includes social media or digital media i.e., the study of the vocabulary of English news information appearing on Twitter. Take the analysis and identification of Fake News as an example. One study proposed to "construct the union of LSACoNet model and FCNN by linking lexical features with conceptual features to identify the purveyors of news other than true/fake news in Twitter [19]".

Secondly, lexical studies were either by comparing examples with a standard corpus or with the news corpus of native English speakers. This approach is in line with the methods adopted in Chinese studies that focus on 1-2 news organizations' corpora. Nauman and Islam [20] use a similar approach for their study of the lexical features of English-language news in Pakistan, but with a special focus on the changes in lexical features before and after the COVID-19 outbreak. The second approach, for example, compares the lexical use of the same events and people in China Daily and Washington Daily to conclude that both news organizations use vocabulary to influence readers' perceptions and attitudes toward news content [21]. A study in 2022 constructed a corpus based on the materials from Xinhua and The Times, focused on the use of lexical bundles and concluded that the former tended to use more VP-based bundles than average, and both have shown a significant tendency to use more NP/PP-based bundles [22].

Third, the linkages between specific events and English news lexical features are explored. In addition to the transformation of Pakistani English news vocabulary mentioned in the previous section, in a more recent study, a small self-built corpus of China Daily is used to study the economic and legal news that appeared during the epidemic, with word length and word distribution, word clusters, and high-frequency words as indicators. This study explores the differences and commonalities between Chinese and legal text and English news [23]. In another study, Ling [24] explored the lexical features of non-native English Speaker's use of English when their writing concerns with the Belt and Road Initiative. His corpus was not limited to news corpus but also included a variety of texts, for instance, dialogues and news-translated text. Based on the idea of lexical classification, the study concluded that pronouns, conjunctions, and prepositions "of" and "and" are overused.

Based on both the studies at home and abroad, it can be concluded that corpus-based lexical feature research is a feasible means to study the lexical perspective of media stylistics. A corpus collection targeting one or two news organizations with a certain topic or event is a suitable approach for collecting research texts. The word density, word length, word class, and word frequency expected to be covered in this study are reasonable research indicators. Thus, it is feasible and meaningful to conduct a comparative study of the lexical features in the news about the Russian-Ukrainian conflict.

## 3. Research Design

### 3.1. Data Collection

The foundation of this study is the construction of two corpora, CN.txt and CD.txt, which represent Chinese and American English news coverage of the Russia-Ukraine Conflict. To ensure the representativeness of the corpora, the study focuses on media content generated by China Daily and CNN, two of the most representative news agencies for Chinese and American English, respectively. In order to ensure the universality of the material, the study used the keyword "Russia-Ukraine Conflict" to search CNN News (https://edition.cnn.com) and China Daily (http://www.chinadaily.com.cn) and randomly collected 118 pieces of news from the search results. The collected news articles were published in February, May, and June, respectively, to ensure the universality and timeliness of the material.

Table 1: Collected materials from CNN and China Daily.

| Corpus | February | May | June | Total （token） |
|---|---|---|---|---|
| CNN | 15 | 26 | 17 | 37070 |
| China Daily | 17 | 26 | 17 | 22390 |

The table above shows the number of articles published each month for each news agency, as well as the total number of tokens collected from each corpus.

### 3.2. Corpus-based Methodology

After collecting the news content, the .docx files were converted to .txt files using the AntFileConverter. These files were then used to form two self-built corpora, CD.txt and CNN.txt.

The corpora were imported into CorpusWordParser for word separation in preparation for the later calculation of type-token ratio (TTR) and standardized type-token ratio (STTR). The software AntConc and the online lexical analysis website VersaText were used for further analysis. The feasibility of using VersaText in text analysis has been previously demonstrated by He [25]. All the results were collected and processed in Microsoft Excel. The word frequency data were counted in the form of a word list, and the lexical features of CD and CNN were compared to draw the study's conclusions.

## 4. Results and Discussion

In the main body of this study, the focus is directed towards a comprehensive comparison of the lexical features among the CN and CD corpora, as well as the reference Corpus of Contemporary American English (COCA). The analysis primarily centers on key lexical aspects, including lexical density, word length, word class, and word frequency. Additionally, an examination of the factors that contribute to the observed distinctions across these three corpora is undertaken.

### 4.1. Lexical Density

Lexical Density, first suggested by Ure [26], is a measure of the proportion of lexical words to the total number of words in a text, which is commonly used to indicate the complexity of discourse. Type-token ratio (TTR) is a widely used measure of lexical richness, which is calculated by dividing the number of different word types (tokens) by the total number of words in the text. However, as the length of a text can affect the TTR, the index STTR (Standardized Type-Token Ratio) has been developed to normalize the TTR across texts of different lengths [27].

This study uses TTR and STTR to measure the lexical richness of news reports about the Russia-Ukraine conflict in China Daily and CNN. To account for differences in text length, this study adopt STTR, which is calculated by taking the mean TTR of each text segment. This approach has been used in previous studies to reveal the lexical complexity of various types of discourse. For example, Veng [8] showed that news headlines had higher lexical diversity than average. Shi [12] compared the TTR of COCA (Corpus of Contemporary American English) with a self-built corpus named CDPNC (China Daily Political News Corpus) and concluded that general English covers a wider range of topics and aspects of life.

#### 4.1.1. Result

Table 2: TTR and STTR for CN and CD.

| Corpus | Files Name | Type | Token | TTR(%) | STTR(%) |
|---|---|---|---|---|---|
| CN | 0-50000 | 1376 | 5046 | 27.26912406 | 29.31 |
|  | 5000-10000 | 1417 | 5018 | 28.23834197 |  |
|  | 10000-15000 | 1448 | 5028 | 28.79872713 |  |
|  | 15000-20000 | 1367 | 5047 | 27.08539727 |  |
|  | 20000-25000 | 1474 | 5044 | 29.22283902 |  |
|  | 25000-30000 | 1428 | 5071 | 28.16012621 |  |
|  | 30000-32020 | 762 | 2092 | 36.42447419 |  |
| CD | 0-50000 | 1183 | 5036 | 23.49086577 | 25.70 |
|  | 5000-10000 | 1139 | 5058 | 22.51878213 |  |
|  | 10000-15000 | 1130 | 5067 | 22.3011644 |  |
|  | 15000-20000 | 1534 | 5052 | 30.36421219 |  |
|  | 20000-22390 | 719 | 2411 | 29.82165077 |  |

The results of the lexical density analysis show that CNN has a higher overall STTR (29.31%) compared to China Daily (25.70%), as displayed in Table 2.

### 4.1.2. Comparison and Analysis

These findings suggest that journalists at CNN employ a greater diversity of vocabulary in their reports, despite differences in length. This result is concordant with the search result in concordance. For example, the search in concordance reveals that CNN uses a broader range of expressions in reference to the "Russia-Ukraine" event, like "conflict", "crisis", "war" and "invasion" as opposed to China Daily's limited use of "Russia-Ukraine conflict".

It is possible that these differences in lexical density may be attributed to variations in the writing and editing processes between China Daily and CNN. This could be linked to the nature of these news agencies. As a "nation's leading English language newspaper", China Daily's writing style, as detailed as the use of the words, needs to follow strict guidelines, which means every reporter, when writing the news, needs to use the same word to refer to the events, so as the Chinese government and its news agency can posit themselves in a relatively neutral position. However, as reporters on CNN are most likely to be native English speakers, also as CNN is private in nature, reporters there may have more flexibility when writing their news coverage.

Notably, our results showcase that both self-built corpora exhibit significantly lower lexical density when compared with the standard corpus. In Malberg's [28] research, he points out that the STTR for the Corpus of Contemporary American English (COCA) is 41%, which was significantly higher than the STTR for CD and CNN. The assumption of such a difference is that the news reporting about the conflict is more formulaic and predictable than the language used in general spoken or written English.

Two factors may contribute to these findings. Firstly, journalists from China Daily and CNN may be trained to write in a specific style. For example, in order to add credit to the report, the use of quotations is frequently introduced in news writing. Quotation, in most cases, started or ended with the phrase "according to". As the table 3 shows, the frequency of the cluster "according to" in the two self-built corpora is significantly higher than its counterpart in COCA. The frequent use of certain writing strategies in writing and the inclination to use the same phrases could help to explain the lower lexical density in news reports.

Table 3: Frequency of "according to" in corpora CN and CD.

| Corpus | Frequency | Frequency(‰) |
|---|---|---|
| CN | 42/32346 | 1.5‰ |
| CD | 35/22624 | 1.2‰ |
| COCA | 208507/ ≈ 560 million | ≈ 0.3‰ |

Secondly, the differences in data collection may have contributed to these findings, as the news reports focused exclusively on the Russia-Ukraine conflict. As a result, reporters may be inclined to use similar expressions and select similar topics and perspectives in their coverage, such as "conflict", "military" and "government" among others. This point will be further illustrated in the subsequent analysis of word frequency.

## 4.2. Word Length

Word length is an essential factor in assessing the lexical complexity of a text. Butler [29] suggested that the longer the average word length, the more complex the text is. The frequency of words of a specific length can also be used to identify various writing styles.

Previous studies in this field have extensively utilized word length as a measure. Tian Yuanyuan [30], for example, employed word length to analyze the stylistic features of Australian news and found that Australian news is longer than LOB and Brown Copurs, implying that Australian news is relatively more difficult to read and comprehend.

Furthermore, in order to further examine and compare the readability and comprehensibility of the corpora, this study introduces the Automated Readability Index. The formula for the Automated Readability Index is 4.71 (characters/words) + 0.5 (words/sentences) - 21.42 [31].

### 4.2.1. Result

Table 4: Word length in corpus CN and CD.

| Corpus | Average no. of Characters per word | Average no. Of syllables per word | Average no. of words per sentence |
|---|---|---|---|
| CN | 6.15 | 1.83 | 29.6 |
| CD | 6.37 | 1.87 | 24.45 |

Table 4 shows the average word length in characters and syllables, it appears that the language used in China Daily has a slightly longer average word length compared to CNN. Specifically, the average number of characters per word is 6.41 for China Daily and 6.15 for CNN. Similarly, the average number of syllables per word is 1.87 for China Daily and 1.83 for CNN. The difference in the average number of words per sentence is also notable, with CNN having a higher average of 29.6 words per sentence compared to China Daily's average of 24.76 words per sentence.

Table 5: Automated readability index in corpus CN and CD.

| Corpus | Automated readability index |
|---|---|
| CN | 22.33 |
| CD | 20.81 |

Table 5 shows the result of the automated readability index in each corpus. CNN's coverage seems to be more difficult to read than China Daily's, as the automated readability index score for CNN is 22.33 and the score for China Daily is 20.81.

### 4.2.2. Comparison and Analysis

As shown in Table 4 and Table 5, although China Daily uses more complex vocabulary or longer words that tend to be more formal or academic, CNN's reports are still more difficult to read based on the Automated Readability Index. One of the contributing factors may be sentence length. Just and Carpenter [32] once confirmed that longer sentences are more difficult to read, and sentence length had a greater impact on reading difficulty than word length. As shown in Table 4, CNN has 5 more words on average in each sentence, which is a relatively significant difference when compared with the mild differences in word length (0.22). Thus, based on this study, a conclusion can be drawn - the news on CNN regarding the Russia-Ukraine conflict is more difficult to read and comprehend and shows a greater degree of lexical complexity than China Daily's news.

In conclusion, our findings demonstrate that CNN's reports regarding the Russia-Ukraine conflict are more complicated to read from the perspective of word length and sentence length. Meanwhile, our findings have reached a similar conclusion as Just and Carpenter that sentence length has a more significant impact on readability than word length which means word length can not be used as a single index in the judgment of complexity and readability of the news report.

## 4.3. Word Class

As is commonly recognized, word class is a linguistic terminology that distinguishes words based on their features. According to Quirk [33], there are content words and function words. Content words include nouns, verbs, adjectives, and adverbs, while function words include determiners, prepositions, etc..

### 4.3.1. Result

Table 6: Word class in the corpora CN and CD.

| Word Class | CN | CD |
|---|---|---|
| Noun | 35% | 36% |
| Verb | 17% | 16% |
| Preposition | 16% | 16% |
| Determiner | 11% | 11% |
| Adjective | 7% | 9% |
| Adverb | 4% | 3% |
| Conjunction | 2% | 2% |
| Pronoun | 3% | 2% |

The analysis of word classes in the news corpora of CN and CD, as shown in Table 6, showcases a similar distribution of word classes in both corpus, with nouns being the most frequent, 35% and 36% respectively, followed by verbs (17% and 16%) and prepositions (16% in both corpora). The proportion of adjectives and adverbs in two corpora turn out to be different. In corpus CN, 7% of the words are adjectives and 4% of the words are adverbs. In corpus CD, 9% of the words are adjectives and 3% of the words are adverbs.

### 4.3.2. Comparison and Analysis

According to Table 6, China Daily tends to use adjectives more frequently while using relatively fewer verbs compared to CNN. The linguistic style of China Daily emphasizes descriptive adjectives to convey information, while CNN's style tends to be more action-oriented, using verbs to convey events and developments. This difference is evident in the examples of news reports from the two sources about the same event, the use of the word cluster before the word "sanctions" sheds light on this point, as shown in table 7.

Table 7: Frequency of the word "sanctions" and the adjective, verb, and clusters before

| Corpus | Frequency of "sanctions" | Frequency of adjective or adjective cluster before "sanctions" | Frequency of verb or verb cluster before "sanctions" |
|---|---|---|---|
| CN | 44 | 11 | 14 |
| CD | 89 | 27 | 11 |
| COCA | 19515 | 5505 | 2507 |

Table 7 presents the results of searching for the term "sanctions" in three corpora: CN, CD, and COCA. The findings reveal differences in the lexical patterns used by China Daily and CNN. Specifically, China Daily tends to use adjectives to modify "sanctions" with high-frequency adjectives like "new" being the most commonly used. In contrast, CNN tends to use verbs before "sanctions", such as "lift", "impose", and "announce".

However, after searching for "VERB sanctions" and "VERB * sanctions" in COCA, this study found that CNN's rhetorical strategy of using more verbs instead of adjectives is not prevalent in general English writing. The use of adjectives or adjective phrases before the noun remains a widely used way of writing. Our findings suggest that CNN deliberately uses fewer adjectives and more verbs to demonstrate its objectivity in reporting news, or tries to offer content that is "factual instead of evaluative" and "knowledge instead of opinion" in the previous study [34].

Hence, based on the findings as well as the examples provided above, a conclusion can be drawn - reporters in China Daily write in a style in which adjectives are more frequently used while CNN tends to use more verbs so as to demonstrate the objectivity of its reports.

Lastly, the frequent use of nouns as well as the least use of pronouns in both corpora drive us to the analysis that news articles tend to focus on specific individuals or events rather than referring to people or things in general.

## 4.4. Word Frequency

The index of word frequency implies the number of occurrences of certain words in a given text or corpus. Previously, word frequency could be used to detect keywords and recognize central topics in discourses. Also, word frequency, to some extent, manifests the language context. "To find out what is distinctive about the style of a text, this study just measure the frequency of the features it contains. The more we wish to substantiate what we say about style, the more we will need to point to the linguistic evidence of texts; and linguistic evidence has to be couched in terms of numerical frequency" [35]. For example, in Biber and Johansson 's study, they classify the words "sofa" as the word in the "fiction" category while the word "constant" in the academic category [36]. Such classification of the categories was realized by calculating the word frequency. In this section, we divide the words into notional words and functional words and display their results separately.

### 4.4.1. Result

Table 8: Notional Word frequency in corpora CN and CD

| Rank | CN | | CD | |
|---|---|---|---|---|
| | Word | Frequency | Word | Frequency |
| 1 | Ukraine (Ukrainian) | 542 | Ukraine (Ukrainian) | 494 |
| 2 | Russian (Russia) | 597 | Said | 285 |
| 3 | Said | 353 | Russia (Russian) | 453 |
| 4 | War | 125 | Military | 123 |
| 5 | Putin | 119 | Conflict | 102 |
| 6 | Military | 97 | President | 90 |
| 7 | President | 95 | Sanctions | 89 |
| 8 | CNN | 93 | Zelensky | 88 |
| 9 | invasion | 90 | Kyiv | 72 |
| 10 | people | 84 | Nato | 70 |
| 11 | Donbas | 73 | EU | 65 |
| 12 | forces | 71 | Countries (Country) | 62 |
| 13 | Region | 71 | Moscow | 60 |

Table 8: (continued).

| | | | | |
|---|---|---|---|---|
| 14 | Moscow | 67 | European | 57 |
| 15 | Country | 59 | Forces | 56 |
| 16 | Eastern | 53 | Putin | 52 |
| 17 | Monday | 53 | Defense | 50 |
| 18 | Zelensky | 53 | Minister | 50 |
| 19 | Kyiv | 52 | Foreign | 48 |
| 20 | Biden | 51 | People | 48 |
| 21 | Conflict | 51 | States | 48 |
| 22 | Mariupol | 49 | New | 47 |
| 23 | State | 48 | City | 46 |
| 24 | world | 28 | ministry | 45 |
| 25 | city | 47 | two | 45 |

In Table 8, the top 25 of the most frequently seen notional words in each corpus are shown.

Table 9: Function Word frequency in corpora CN and CD

| Rank | CN | | CD | |
|---|---|---|---|---|
| | Word | Ratio (%) | Word | Ratio (%) |
| 1 | the | 5.39% | the | 7.53% |
| 2 | to | 2.66% | of | 3.20% |
| 3 | of | 2.60% | to | 2.82% |
| 4 | in | 2.20% | and | 2.50% |
| 5 | and | 2.00% | in | 2.36% |
| 6 | a | 1.76% | on | 1.50% |
| 7 | on | 1.20% | that | 1.23% |
| 8 | that | 1.01% | for | 0.96% |

Table 9 displays the top eight most frequently occurring function words, with their frequencies expressed as proportions relative to the total number of tokens in the corpus, taking into account the differences in corpus size. The proportions are calculated by dividing the frequency of each function word by the total number of tokens in the corpus.

### 4.4.2. Comparison and Analysis

Regarding the frequency of notional words, the following findings are worth noticing.

Firstly, both corpus share similarities in using words which show the focus of the news writer from China Daily and CNN are similar too. For example, both news agencies put more emphasis on reporting events relating to Ukraine (Ukrainian). The presidents of Russia and Ukraine appear to be two of the central characters in the news reports, given that their names are frequently mentioned. These similarities point out a common feature shared by CNN and China Daily, journalists will constantly focus on certain people or events and show their relevance by mentioning the object frequently.

Moreover, several differences reflect the differences in writing and editorial styles as well as the distinctive attitudes held by CNN and China Daily.

Firstly, the differences in the use of the word “war” and “conflict” reflects the differences in the severity of the events according to both agencies. The former word is preferred by CNN while the latter by China Daily. This conclusion is supported by the previous study about the use of euphemisms in the political field which indicates that "understatement" and the "use of fuzzy words" are ways to disguise the blood battle [37]. As for CNN, it emphasizes the military conflict and actions between the two countries and shows no intention of understating Russia's military action, thus describing the event directly as "war" and using almost no euphemism. China Daily, however, considering China's neutral position in the Russia-Ukraine conflict, describes the event as "conflict", which is a widely used euphemism that refers to a blood battle.

Meanwhile, the differences in the use of the words reflect the writer's attitude towards the event. For example, the word "invasion" was frequently used in CNN. Invasion, defined by Webster as "a foreign army enters it by force", is an accusation word towards Russia in nature. As for China Daily, we can see almost no word expressing certain supportive or accusatory meanings in its reports. Also, the words "NATO", and "EU" are in China Daily's list, which suggests a focus on international relations and diplomacy as a third party. These terms do not appear in CNN's list, which suggests a greater focus on the military conflict itself.

Overall, the results for notional words suggest that both CNN and China Daily are closely following the ongoing conflict between Ukraine and Russia, but with some differences in focus and editorial priorities. The attitudes held by the two news agencies are clearly revealed by their use of the words.

Regarding the frequency of the function words, there are findings and analysis as follows:

Firstly, in both corpora, the definite article "the" is the most frequently seen function word. In CN, the ratio is 5.39%, and in CD the ratio is 7.53%. By searching "the" in COCA, the frequency in this general corpus is around 5.006% (50067877 in total). Comparing these three corpora, we can see China Daily is excessively using the word "the" in its coverage. Mentioning the "event's name" frequently is one of the reasons why China Daily uses "the" frequently. For example, "the Russia-Ukraine conflict" was mentioned 29 times in China Daily's reports while appearing only three times in CNN's reports. Moreover, "the West", which is used to refer to the Western world, is a unique expression that has been used frequently in China Daily's reports while not being seen in CNN's reports, the use of "the West" demonstrates China Daily's point of view as a news agency from an eastern and developing countries. These two reasons above are the factors that contribute to the differences in using the function word "the" in CNN and China Daily.

Secondly, according to Table 9, in corpus CNN, "a" has been frequently used but in CD it is relatively less frequently seen. The higher frequency of "a" in CNN could indicate a preference for using indefinite articles in their news articles. It may also suggest that CNN's writing style tends to be more conversational or informal, as the use of "a" is often associated with less formal language. On the other hand, the lower frequency of "a" but the high frequency of "the" in China Daily indicates a preference for more concise or formal language in their news articles.

## 5. Conclusions

### 5.1. Summary of Findings

This study focused on the discourse of the news reports about the Russia-Ukraine conflict in China Daily and CNN. Firstly, the results show CNN uses a greater variety of words compared with China Daily but still has a lower variety compared with COCA, which indicates words in news writing are more formulated and predictable. Secondly, CNN's news writing is more difficult to read considering it tends to use more words in one sentence. While China Daily's use of the words is longer and more academic. Thirdly, the preference for verb indicate a more objective tone in CNN's report and the preference for adjective in China Daily's reports shows the opposite. In the last section of this study, it is clear that euphemism is considered in choosing words about warfare, meanwhile, CNN's criticizing attitude towards Russia and China Daily's neutrality are reflected in their differences in word use.

Overall, this study provides insights into the lexical features of news reporting on a specific topic in different languages and media outlets, highlighting the importance of considering text length, word frequency, and formulaic expressions in analyzing the complexity of discourse and attitude of the writers.

### 5.2. Implications

This study, based on a Chinese perspective, aims to provide useful suggestions for non-native English journalists or students who want to improve their English writing skills. Based on the findings, the following instructive implications are suggested:

Firstly, regarding word density, non-native English journalists can improve the nativeness of their news reports by using a diverse range of words or expressions. However, considering the writing style of China Daily, it is important for reporters to adhere to a set of widely accepted and appropriate words and expressions for certain social events. Rechecking previous reports can help to ensure accuracy and consistency in news writing.

Secondly, the findings about word length suggest that using longer and more complex sentences, rather than long and academic words, can enhance the language efficiency of news writing and make the reports sound more natural.

Thirdly, based on the results about word class, it is recommended that journalists use more verbs to convey news stories in a factual manner and demonstrate the neutral stance of the news agency. This feature distinguishes news writing from general English writing. Therefore, China Daily's English reporters are encouraged to adopt a similar writing style to establish the authority of the results and maintain a third-party point of view.

Lastly, the results about word frequency indicate that non-native English journalists and writers should use expressions that attribute quotations to news sources, such as "according to" or "said". It is also important to check for differences in meaning, particularly in the context of use, since the choice of words reflects the ideology and point of view of the writer towards certain social events. Additionally, using euphemisms is an effective strategy in news writing to demonstrate controlled ideology and attitude, as demonstrated in the examples of "conflict", "invasion", and "war". From the differences in frequency of the function words, the high frequency of the word "the" in CD indicates its inclination to use "the" to modify the name of an event and its frequent use of "the west" which refers to the Western world.

### 5.3. Limitations and Future Directions

This study focuses on the lexical features and differences in news writing about the Russia-Ukraine conflict as conducted by China Daily and CNN. However, the study has limitations due to the fact that both corpora are about a single social event in the political field. Therefore, it is not possible to compare the corpus focusing on this specific social event with usual news writing that covers a wider variety of topics. As a result, our findings about lexical features of news writing may be insufficient.

Moreover, as this study focuses on language style, lexical features are not the only indicators that can show differences between native and non-native English news writers. Other indicators such as sentence length, clause structure, or tense could be explored in future research. Therefore, a more comprehensive analysis that covers multiple perspectives in language style, as well as a broader corpus, would allow for a more conclusive understanding of news writing style.

Lastly, recent research on news discourse has turned towards examining discourse seen on new media platforms and linking linguistic features with ideology and audience response. This direction of research was not explored in this study, thus comparing media discourse on new media platforms with their linguistic features could be an applicable direction for future research.

## References

[1] *Li, X. Z. (2016). The Formation and Acquisition of Style J. Contemporary Rhetoric, (6), 4-10.*

[2] *Lambrou, M., & Durant, A. (2014). Media stylistics [A]. In P. Stockwell & S. Whiteley (Eds.), The Cambridge Handbook of Stylistics (Cambridge Handbooks in Language and Linguistics) [C] (pp. 503-519). Cambridge: Cambridge University Press.*

[3] *Chen, Y. (1980). Language and Social Life: Sociolinguistics Notes M. Shanghai: Xinzhi Sanlian Bookstore Publishing.*

[4] *Lewis, M. (1993). The Lexical Approach: The state of ELT and a Way Forward [M]. Hove: Language Teaching Publications.*

[5] *Baker, P. (2006). Using Corpora in Discourse Analysis [M]. London: Continuum.*

[6] *Leech, G. N., & Short, M. (2007). Style in fiction: A linguistic introduction to English fictional prose (No. 13) [M]. London: Pearson Education.*

[7] *Fowler, R. (2013). Language in the News: Discourse and Ideology in the Press [M]. Oxfordshire: Routledge.*

[8] *Weng, Y. X. (2016). Vocabulary Feature Study of English News Headlines D. Dalian: Dalian Maritime University.*

[9] *Weng, Y. X. (2018). Corpus-Based Quantitative Study on Vocabulary Distribution of English News Headlines J. Journal of Guangdong University of Foreign Studies, (5), 26-33.*

[10] *Huang, Z. X., & Liu, Z. Y. (2020). Corpus-Based Analysis of English News Lead Writing Style J. Journal of Wuhan Metallurgical Management Institute, (3), 91-93.*

[11] *Zang, Y. Y. (2011). Stylistic Features of English News Vocabulary J. News Lovers, (10), 114-115.*

[12] *Shi, Z. Y. (2017). Corpus-Based Study on Stylistic Features of English Political News Vocabulary in China Daily D. Henan: Xinyang Normal University.*

[13] *Yang, Q. Q. (2018). Comparative Study on English News Reporting Style of Xinhuanet and BBC Based on Corpus D. Henan: Henan University of Technology.*

[14] *Han, D. J. (2021). Corpus-Based Study on the Vocabulary Richness of English News in Chinese Media J. Modern Communication, (11), 82-84.*

[15] *Lu, J. (2022). Comparative Study on the Stylistic Features of Chinese and English Broadcast English News Reporting Based on Corpus J. Journal of Henan Polytechnic University (Social Science Edition), (5), 87-92.*

[16] *Hu, L. (2014). Lexical Features of English News from the Perspective of Economic and Trade News J. News Front, (9), 119-120.*

[17] *Lin, J., Li, C. F., & Yu, J. (2018). Setting the Threshold of Business English Professional Vocabulary: Based on the Analysis of Business English News in Financial Times and Fortune from 2006 to 2015 J. Journal of Xiamen University (Philosophy and Social Sciences), (2), 165-172.*

[18] *Yi, S. (2013). Comparative Study on Language Features of English News Reporting and News Commentary D. Wuhan: China University of Geosciences.*

[19] *Giglou, H. B., Razmara, J., Rahgouy, M., & Sanaei, M. (2020, September). LSACoNet: A Combination of Lexical and Conceptual Features for Analysis of Fake News Spreaders on Twitter. In CLEF (Working Notes).*

[20] *Nauman Ahmed, H., & Islam, M. (2020). Influence of COVID-19 on the Lexical Features of English in Pakistan [J]. Linguistics and Literature Review, 6 (2), 69- 82.*

[21] *He, X. Z., & Zhou, X. Z. (2015). Contrastive analysis of lexical choice and ideologies in news reporting the same accidents between Chinese and American newspapers [J]. Theory and Practice in Language Studies, 5(11), 2356.*

[22] *Xu, M., & Sun, T. (2022). Structural Analysis of Lexical Bundles in English News on Health from China and UK Newspapers [A]. In Daren Zheng (Eds.). 2022 International Sociology, Economics, Education and Humanities Conference [C]. Canada: Clausius Scientific Press.*

[23] *Liu J (2022) Lexical Features of Economic Legal Policy and News in China Since the COVID-19 Outbreak [J]. Front. Public Health, 10:928965.*

[24] *Ling, Z. H. (2016). A Corpus-Based Study of Lexical Features of Nonnative English from the Perspective of the Belt and Road [A]. In Y. W. Ge, L. Hale, and J. Zhang (Eds.). In Proceedings of International Symposium on Policing Diplomacy and the Belt & Road Initiative [C] (pp. 246-251). Georgia: The American Scholars Press.*

[25] *He, Anping. (2022). The Advantages and Applications of Corpus Intelligent Learning Platform: A Case Study of High School English Teaching [J]. Research on Classroom Teaching in Primary and Secondary Schools, (01), 1-5.*

[26] *Ure, J. (1971). Lexical Density and Register Differentiation [J]. Contemporary Educational Psychology, 5, 96-104.*

[27] *Covington, M. A., McFall, J. D., & Millsap, R. E. (2016). Standardized assessment of computerized linguistic and psychoacoustic features of speech [J]. Journal of Speech, Language, and Hearing Research, 59(4), 767-777.*

*[28] Mahlberg, M., Smith, N., & Wallis, K. (2017). Standardized Type-Token Ratios (STTR) as a Standardized Measure of Lexical Diversity [J]. Language Resources and Evaluation, 51(4): 1009-1021.*

*[29] Butler, C. (1985). Statistics Linguistics [M]. Oxford: Basil Blackwell.*

*[30] Tian, Yuanyuan. (2009). Corpus-Based Study on Australian News English Vocabulary [D]. Dalian: Dalian Maritime University.*

*[31] Amoroso, C. G. N., & Smith, E. M. (1967). An Automated Readability Index [J]. Journal of Applied Psychology, 51(2), 119–125.*

*[32] Just, M. A., & Carpenter, P. A. (1980). The effects of sentence length, relevance, and familiarity on reading rates [J]. Journal of Verbal Learning and Verbal Behavior, 19(5), 467-478.*

*[33] Quirk, R. (1990). Language varieties and standard language. English today, 6(1), 3-10.*

*[34] Van Dijk, T. A. (1998). Opinions and ideologies in the press [J]. Approaches to media discourse, 21(63).*

*[35] Xu, Youzhi. (2005). English Stylistics Course [M]. Beijing: Higher Education Press.*

*[36] Biber, D., Conrad, S., & Johansson, S. (2009). Longman grammar of spoken and written English [M]. Beijing: Foreign Language teaching and research Press.*

*[37] Guo, Qing, Tan, Yingwen, & Dai, Weiping. (2015). English News and Vocabulary Research [M]. Guangzhou: World Book Publishing House.*