ISSN (Print): 2959-4391 ISSN (Online): 2959-4405

JOURNAL OF SOCIAL SCIENCES DEVELOPMENT

www.jssd.org.pk editor@jssd.org.pk

A CORPUS BASED STUDY OF ENGLISH VOCABULARY IN DAWN NEWSPAPER SOCIAL EDITORIALS (2023-2025)

Amna Javed[1] & Dr. Muhammad Akram[2]

[1]BS English Linguistics Rahim Yar Khan Campus, Email: amnajaved721@gmail.com

[2]Assistant Professor of English Linguistics, The Islamia University of Bahawalpur, Rahim Yar Khan Campus. Email: muhammadakramw@gmail.com

KEYWORDS

Corpus Linguistics, Dawn Newspaper, Social Editorials, Vocabulary Analysis,

ARTICLE HISTORY

Date of Submission: 03-06-2026

Date of Acceptance: 05-07-2026

Date of Publication: 12-07-2026

ABSTRACT

The present study aims mainly at investigating the vocabulary used in the social editorials of the Dawn newspaper from 2023 to 2025. For this purpose, editorials representing social issues published in Dawn newspaper between 2023 and 2025 were taken as the sample of research. A corpus-based quantitative research design supported by limited qualitative interpretation was adopted. Selected social editorials were compiled into a corpus of 92,114 tokens and analyzed using AntConc and AntWordProfiler. The corpus analysis tool AntConc is used to find the most frequent lexical words and to analyze the collocational patterns of the seven most frequent lexical words using the Collocate Tool with Log-Likelihood statistics. The vocabulary profile of the corpus is analyzed using AntWordProfiler to examine lexical complexity by comparing it with the General Service List (GSL) and the Academic Word List (AWL). The findings show that women, children, health, people, government, and state are among the most common lexical words in the corpus. The collocation analysis also shows that these words are used in meaningful patterns in relation to social concerns. The results show that the editorials are characterized by a moderately complex vocabulary with a considerable proportion of high-frequency words and academic and specialized lexical items. The study contributes to corpus based research on media discourse by giving insights into the vocabulary patterns of Dawn newspaper editorials and enhancing the understanding of the linguistic representation of social issues in Dawn newspaper.

2026 Journal of Social Sciences Development

Corresponding Author Amina Javed

Email: aminajaved721@gmail.com

DOI https://doi.org/10.59075/jssd.v5i5.311

Volume/Issue 5(5), 2026, 181-191

Introduction

Vocabulary is very important in the process of communication and meaning making; especially in newspaper editorials where writers attempt to inform, persuade and influence public opinion. The selection of the vocabulary in editorials not only reflects the issues discussed in the editorial, but also the linguistic strategies used to present social realities. Dawn is one of the top English-language newspapers in Pakistan that releases editorials on various social issues like education, health, poverty, gender inequality and human rights, among others. These editorials offer a good source of language information to analyze the vocabulary patterns in media discourse.

Corpus linguistics has emerged as a powerful methodological framework for investigating language use in authentic contexts, offering empirical, computer-assisted approaches to analyzing large collections of natural texts (Biber et al., 1998). Through frequency analysis, researchers can identify the most useful and commonly used words in a language, which is a crucial aspect of vocabulary research (Szudarski, 2017). Similarly, collocation studies, grounded in the principle that words do not appear randomly but in regular patterns of co-occurrence, allow for the examination of how meaning is created and interpreted in discourse (Sinclair, 1991). Furthermore, vocabulary profiling, which involves comparing a corpus against established lists like the General Service List (GSL) and the Academic Word List (AWL), provides a means to measure lexical complexity and coverage (Nation, 2001). Through frequency analysis, collocation studies, and vocabulary profiling, corpus-based methods enable researchers to uncover systematic patterns of language use that might otherwise remain invisible through traditional qualitative analysis alone.

Despite the growing application of corpus linguistics to media studies, a significant gap exists in the examination of vocabulary patterns in newspaper editorials, particularly those addressing social issues. While previous studies have explored environmental discourse (Fatima & Arslan, 2023), academic vocabulary (Sajjad et al., 2023), headline features (Bibi et al., 2026), and interactive markers (Qasim et al., 2024) in Pakistani newspapers, no comprehensive investigation has systematically examined the overall vocabulary patterns, collocation networks, and lexical complexity of social editorials in Dawn, Pakistan's most influential English-language newspaper.

This gap is filled by the present study which analyzes the social editorials of Dawn published in the period of 2023-2025 from a corpus perspective. With the help of AntConc and AntWordProfiler, the study explores the most common lexical words and looks at the collocational patterns of the seven most common lexical items and also examines the vocabulary profile of the corpus. The findings contribute to corpus-based research on newspaper editorials and provide a deeper understanding of vocabulary use in Dawn social editorials.

Research Questions

1. What are the most frequent lexical words and the collocation patterns of the top seven lexical words used in Dawn's social editorials?

2. Do the words in Dawn's social editorials are simple or complex?

Literature Review/Background of the Study

Corpus linguistics is different from most other things you might learn about in linguistics, since it doesn't directly deal with any one part of language. Instead, it is a field that concentrates on a collection of procedures or methods for examining language. (McEnery & Hardie, 2012). The essential features of corpus-based analysis are: It is empirical: it studies what happens in real texts.It takes a large and principled set of natural texts, called a "corpus," as the basis for analysis.It makes extensive use of computers for analysis, both automatically and interactively.It is based on quantitative and qualitative analytical techniques.Together, these features provide a scope and reliability of analysis that is otherwise unavailable. Much of the success of the corpus-based approach comes from the use of computers. Computer can recognize and analyze complex patterns of language use, storing, and analyzing a larger database of natural language than can be done by hand. Computers also provide consistent and reliable analyses; they don't change their mind or get tired during an analysis.Computers can also be used interactively, allowing the human analyst to make difficult linguistic judgments while the computer maintains the records (Biber et al., 1998).

According to Szudarski (2017), frequency is an important aspect of vocabulary research because it helps identify the most useful and commonly used words in a language. Frequency analysis also helps researchers understand which words learners should learn first and helps in making vocabulary lists for language learners.

Lexical words have meaning and point to things we can think of, like actions, objects, places, and so on. Function words, on the other hand, tell us about grammatical relationships and don't have much meaning. (Fromkin et al., 2014) This difference is significant for the present study as it supports the finding of lexical items during the word frequency analysis in the corpus.

Collocation is the phenomenon of words occurring together more often than would be expected by chance. According to John Sinclair (1991), words do not appear randomly in texts but rather in regular patterns of co-occurrence. He claims that language users tend to use semi-preconstructed phrases instead of single words – this is referred to as the idiom principle. From this perspective collocation is an important feature of language as it helps to show how meaning is created and interpreted in discourse. In corpus linguistics, collocation analysis is used to find words that occur together within a given distance of a target word, which allows researchers to examine the common ways language is used in real texts. Therefore, the study of collocations can provide interesting information about the semantic relations and the recurrent lexical associations of a corpus. This is especially pertinent to the present study that explores the collocational pattern of the most frequent lexical words in Dawn social editorials.

Nation (2001) classifies vocabulary into main four types based on the frequency,coverage, and usefulness in texts. These are the following types:

1. Nation (2001) highlights the importance of high frequency vocabulary in language use. He explains that high frequency words consist of both function words and content words. Nation states that the most widely recognized list of high frequency vocabulary is West's General Service List (GSL) which contains about 2,000 word families. These words are especially important since they make up a high percentage of words in texts, with almost 80% coverage of running words. High-frequency vocabulary is an indispensable part of language comprehension and communication.
2. Academic vocabulary are the vocabulary items that are high frequency across a wide range of academic texts but low frequency in everyday language. According to Nation (2001), these words play an important role in academic discourse. Usually, they make up about 9% of the running words of academic texts. He also refers to the importance of the Academic Word List (AWL) compiled by Coxhead (1998), which contains 570 word families that occur frequently in a range of academic disciplines. Academic words are those words that are frequent in the academic writing and thus it is believed that they are important for understanding and producing academic texts.
3. Technical vocabulary refers to words that are closely related to a specific subject area or discipline and are not commonly used outside that field. According to Nation (2001), such words help to signal the topic of a text and are more frequent within a particular domain, but less common in general language use. Examples may differ depending on subject areas, e.g. economics, geography, environmental studies. Nation also points out that technical vocabulary is usually around 5% of the running words in a text.
4. Low-frequency vocabulary is made up of words that are not included in a high-frequency word list, an academic word list, or the technical vocabulary of a particular subject area. According to Nation (2001), this is the largest group of vocabulary in English. It is made up of rare words, proper nouns, and words that are not often used in everyday or academic texts. These words occur infrequently individually, but together they form a noticeable proportion of any text and represent vocabulary that learners encounter less often.

In vocabulary studies, lexical coverage is the percentage of known words in a text required for a reader to achieve adequate comprehension. Hu and Nation (2000) identified that the threshold of vocabulary coverage of 98% is the ideal level for a reader to achieve the unassisted reading comprehension of authentic texts without effort. If coverage drops below this threshold, reading fluency is greatly affected, because the density of unknown vocabulary is so high that the reader must rely heavily on external resources, such as dictionaries. Importantly, the amount of vocabulary required to reach this benchmark is genre-specific. Authentic journalistic prose requires a much larger vocabulary than simplified or narrative fiction texts, which may require a reader to know 2,000 to 3,000 high-frequency word families.

Previous studies on Dawn news

Fatima and Arslan (2023) investigated a corpus-based eco-linguistic analysis of Dawn newspaper and demonstrated the usefulness of corpus methods in analyzing media discourse. But

their study was on environmental discourse, not vocabulary patterns. Therefore, the Dawn's social editorials vocabulary is unexplored.

A corpus-based study on academic words in Pakistani English newspapers was performed by Sajjad, Hassan, and Gul (2023). Their findings showed the presence and significance of academic vocabulary in newspaper discourse. However, their study focused on academic words only, while the present study looks at overall vocabulary patterns, collocations and lexical complexity in Dawn's social editorials.

Bibi et al. (2026) analyzed a comparative study of syntactic and lexical features of the headlines of Dawn and The Express Tribune. Their findings showed differences in the language used in newspaper headlines. However, the study dealt with the language of headlines whereas the present study examines the vocabulary frequency, collocation patterns and lexical complexity in the social editorials of Dawn.

Qasim, Feroza, and Nawaz (2024) conducted a corpus-based analysis of interactive markers in the editorials of The News and Dawn. heir work was focused on the use of linguistic markers to structure and guide the reader in editorial discourse. However, it did not investigate vocabulary frequency, collocation patterns, or lexical complexity which are the focus of the present study.

Research Methodology

Sample of the Research

The sample data for research is taken from Dawn Newspaper. The selected data consists of 92,114 tokens taken for analysis from January 2023 to December 2025.

Research Design

This research is a corpus based study with mostly quantitative approach.The study has a limited qualitative interpretation in order to explain the quantitative results. Only editorials that primarily focus on social issues were selected. Dawn's political, economic, sports, international and letters to the editor sections were not included. Editorials addressing social issues were chosen from the official Dawn News archive for the time span 2023-2025. AntConc and AntWordProfiler were used for analysis of the corpus. AntConc was used to create frequency lists and identify the most frequent lexical words. Additionally, the software was used to analyse the collocation patterns of the most frequent seven lexical items employing the Collocate Tool along with log-likelihood statistics. The lexical complexity was measured using AntWordProfiler, which was used to compare the corpus with the General Service List (GSL) and the Academic Word List (AWL) and identify the vocabulary at various frequency levels. The results were quantitatively interpreted and only some qualitative discussion was presented to explain the lexical patterns found in the corpus

Data Analysis

Frequency Analysis of Social Editorials

The data is analyzed by the corpus tool Antconc which provides frequently used lexical words. The Antconc tool found 92114 tokens in total in 266 files.This frequency count includes grammatical and lexical words. A stop list was used to remove function words in order to focus on meaningful lexical content. Also, proper nouns (e.g., Pakistan, Polio) and number words (e.g.,one two,three) were excluded from the analysis.

Table 1: Top Frequent lexical words in the corpus of social editorials Of Dawn news (2023-2025)

Sr	Words	Freq	Sr	Words	Freq	Sr	Words	Freq
1	women	362	16	violence	110	31	workers	82
2	children	279	17	society	107	32	areas	80
3	health	267	18	social	106	33	gender	79
4	people	197	19	years	102	34	global	79
5	country	195	20	report	99	35	long	78
6	government	179	21	time	98	36	day	77
7	state	176	22	high	97	37	media	76
8	child	149	23	mental	93	38	action	75
9	public	142	24	international	87	39	due	74
10	cases	138	25	life	87	40	justice	74
11	rights	137	26	authorities	86	41	men	74
12	education	133	27	ensure	85	42	recent	74
13	world	133	28	law	85	43	safety	73
14	need	111	29	human	83	44	lives	72
15	police	110	30	abuse	82	45	water	72

As seen in the table 01, the word women appears most frequently (362) followed by children (279) and health (267). Other words that occur frequently are people (197), country (195), government (179), and state (176). All of the other lexical words have frequencies ranging from 149 to 72. The results also show that there are social-related lexical items that are repeated across the corpus including rights (137), education (133), violence (110), mental (93) and gender (79). The table provides a frequency distribution of the most common lexical items found in the corpus.

Collocations Patterns of Seven most Frequent lexical Words in Social Editorials

The collocation analysis showed the recurrent lexical patterns around the top seven frequent content words in Dawn's social editorials. Collocates were extracted using the Collocate Tool in AntConc with Log-Likelihood as the statistical measure. Low-information grammatical items (such as articles and prepositions with no semantic weight) were removed from the final tables, so that the analysis could focus on semantically important lexical relations.

Table 2: Collocations Patterns of Seven most Frequent lexical Words

Sr	Lexical words	Top collocates with frequency	Context intrepretation
1	women	against(32),men(19),violence(18) crimes(11),rights(18)	gender relation,violence against women,legal rights
2	children	school(19),missing(8),abuse(12), parents(8)	education,family,social issues
3	health	mental(60),public(26),infrastructure (12),reproductive(8),services(11)	public health system and healthcare services
4	people	died(8),killed(8),living(7),worldwide(5),suicide(6)	life, death and world social problems
5	government	must(36),provincial(6),response(6)	government, action
6	state	needs(11),must(24),inability(5), affairs(4),citizens(9)	government duties, social failures and public service.
7	child	abuse(24),labour(16),marriage(11), abduction(5),exploitation(6)	child welfare,social and legal issues

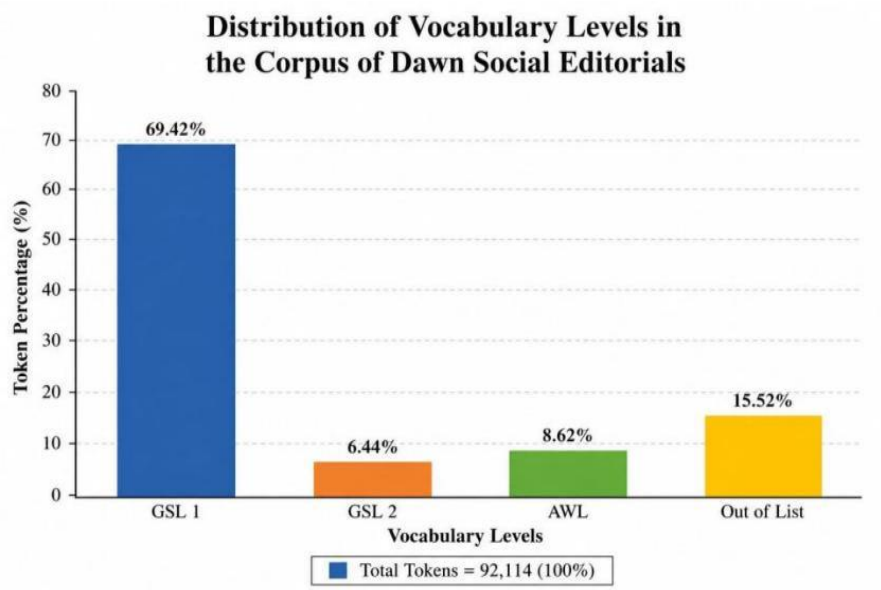
For the lexical word *women*, the strongest collocates are *against* (32), *men* (19), *violence* (18), *crimes* (11), and rights (18). The word *children* is most strongly associated with *school* (19), followed by *missing* (8), abuse (12), and *parents* (8). For *health*, the strongest collocates are *mental* (60), *public* (26), *infrastructure* (12), *reproductive* (8), and services (11). The lexical word *people* co-occurs with *died* (8), *killed* (8), *living* (7), *worldwide* (5), and suicide (6). The word *government* is most strongly associated with *must* (36), followed by *provincial* (6) and *response* (6). Similarly, *state* co-occurs with *needs* (11), must (24), *inability* (5), and *affairs* (4), citizens (9). The lexical word *child* is most strongly associated with *abuse* (24), *labour* (16), *marriage* (11), *abduction* (5), and exploitation (6). The word "country" is one of the most frequent lexical words, but no significant collocates were found in the selected range of analysis.Therefore, the researcher replaced the word with the next frequent word "child" for meaningful collocation analysis.

Analysis of Lexical Complexity

This analysis was conducted using AntWordProfiler which compares the corpus against the West's General Service Lists and the Coxhead's Academic Word List.

Table 3: Vocabulary Profile of Dawn Social Editorials

Vocabulary Level	Tokens (Count)	Token%	Cumulative %
GSL 1	63,945	69.42%	69.42%
GSL 2	5,931	6.44%	75.86%
AWL	7,943	8.62%	84.48%
Out of List	14,295	15.52%	100.00%
Total	92,114	100%	



The GSL is divided into two levels: GSL 1 (the first 1,000 most frequent word families) and GSL 2 (the second 1,000 most frequent word families). The AWL is made up of 570 academic word families, and the Out-of-List category includes words that are not in the GSL or the AWL. The analysis shows that GSL 1 covers 69.42% and GSL 2 covers 6.44% of the corpus. The total GSL coverage is 75.86%. The Academic Word List (AWL) makes up 8.62% of the total vocabulary and the Out-of-List words make up 15.52% of the corpus.

Discussion

High-Frequency Lexical Vocabulary

The frequency analysis shows that the lexical vocabulary of Dawn social editorials is highly concentrated on the major Pakistani social concerns. The high frequencies of words like women, children, health, people, government, and state indicate that the editorials have a consistent focus on issues of gender equality, child welfare, public health, governance, and social responsibility.The frequent appearance of words like rights, education, violence, mental, and gender also indicates that the newspaper pays a lot of attention to human rights, education, mental health and social justice. These results suggest that Dawn utilizes its editorial platform to draw attention to issues of major social significance and public impact.Frequency analysis is an effective method to find the most significant vocabulary in a corpus because frequent words are the dominant topic of a text, according to Szudarski (2017). Thus, the lexical items pointed out in the present study represent the main themes talked about in Dawn's social editorials.

Collocational Patterns of the Most Frequent Lexical Words

The collocation analysis shows that the most frequent lexical words in Dawn social editorials appear in meaningful and consistent lexical patterns. These words do not occur in isolation but are repeatedly associated with other semantically related words, which reflect the major social concerns discussed in the editorials. For instance, the collocates of women are connected to violence, rights and gender relations, while children and child often co-occur with words concerning education, abuse, labour, marriage and family welfare. Likewise, health is highly correlated with mental health, public health, and healthcare services, while government and state are frequently associated with responsibility, public services, and policy action. Such recurring lexical links suggest that the editorials tend to frame social issues in terms of human rights, public welfare and institutional responsibility.

These findings support Sinclair's (1991) view that words do not occur randomly but form regular collocational patterns that contribute to meaning in discourse.The recurrent collocations identified in the present study demonstrate how Dawn constructs social issues through stable lexical associations, enabling readers to recognize the central concerns emphasized in the editorials. Hence, the collocation patterns offer a more comprehensive perspective on the linguistic representation of social issues than the frequency analysis alone, which emphasizes the importance of corpus-based methods in the study of newspaper discourse.

Vocabulary Profile and Lexical Complexity

The majority of the text (75.86%) consists of simple GSL 1 and 2 vocabularies, these numbers are high because simple words like 'the,' 'is,' and 'people' repeat constantly in every sentence. The Academic Word List (AWL) shows 8.62% coverage which is an indication that the text is formal and university level writing. Moreover, 'Out-of-List' words account for 15.52% which includes technical and low frequency words. According to Cumulative Percentage 75% of the vocabulary belongs to high frequency everyday words, but 25% are not. A reader relying only on general service vocabulary (GSL 1 & 2) would achieve only 75.86% coverage of this corpus. This is significantly lower than the Nation lexical threshold of 98% comprehension for unassisted reading, therefore the vocabulary of the Dawn social editorials needs to be labeled as moderately complex and advanced, requiring a significant knowledge of academic (AWL) and technical and low frequency words.

Conclusion

This study investigated the vocabulary of Dawn's social editorials published from 2023 to 2025 using a corpus-based approach. The frequency analysis showed that words like women, children, health, people, government, and state were some of the most frequent lexical items. Collocation analysis also revealed that these words are found in meaningful patterns associated with major social concerns such as violence against women, child welfare, public health and governmental responsibility. The vocabulary of the Dawn editorials is moderately complex and advanced. General service vocabulary (GSL 1 & 2) covers only 75.86% of the text, falling well short of

Nation's 98% comprehension thresholds required for easy reading. Consequently, readers require strong academic (AWL) and specialized vocabulary to effectively understand these texts. The study contributes to corpus-based research on media discourse and offers insights into the linguistic representation of social concerns in Dawn newspapers.

Recommendations

Corpus-based tools like AntConc and AntWordProfiler should be used more often in media language research as they offer a reliable and objective linguistic analysis. Lexical profiling should be included in future analyses of newspaper discourse to evaluate the level of vocabulary complexity and its suitability for different groups of readers. Universities and researchers should promote more corpus-based studies of Pakistani English newspapers to strengthen linguistic research and extend understanding of media language in Pakistan. English language teachers can use the authentic newspaper editorials as the classroom materials to enhance the students' vocabulary, reading comprehension and awareness of the current social issues. Researchers and educators are suggested to promote the use of authentic newspaper corpora for the improvement of vocabulary learning and to develop better understanding of real-world language use.

Recommendations for Further Research

Future researchers may extend this study by comparing the vocabulary of social editorials from different Pakistani English newspapers to identify similarities and differences in language use.A larger corpus covering a longer time period may provide more comprehensive findings. Future researchers may investigate the vocabulary of specific social issues, such as education, healthcare, gender equality, child rights, or poverty, instead of analyzing all social editorials together. Future studies may also compare vocabulary of social editorials in Pakistani newspapers with the vocabulary used in international newspapers using corpus-based methods.

References

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

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

Szudarski, P. (2017). *Corpus linguistics for vocabulary: A guide for research*. Routledge.

Fromkin, V., Rodman, R., & Hyams, N. (2014). *An introduction to language* (10th ed.). Cengage Learning.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-430.

Fatima, G., & Arslan, M. F. (2023). Eco-Linguistic Analysis of 'Dawn' Newspaper: A Corpus Based Study. Journal of Asian Development Studies, 12(4), 112–122. https://doi.org/10.62345/jads.2023.12.4.8

Sajjad, U., Hassan, S. S., & Gul, N. (2023). Academic words in discourse news: A corpus based study of Pakistani English newspaper. *Journal of Development and Social Sciences , 4*(2), 515-523. https://doi.org/10.47205/jdss.2023(4-II)46

Bibi, S., Razzaq, W., Hussain, W. M., & Tahir, A. (2026). A comparative study of syntactic and lexical features in English newspaper headlines: Dawn and The Express Tribune. The Critical Review of Social Sciences Studies, 4(1), 801–812. https://doi.org/10.59075/7et4e270

Qasim, H. M., Feroza, & Nawaz, M. S. (2024). A corpus-based analysis of interactive markers in English editorials of The News and Dawn News. *Corporum: Journal of Corpus Linguistics, 7*(1), 192–214. https://journals.au.edu.pk/ojscrc/index.php/crc/article/view/314