ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

GENDERED LEXICAL CHOICES IN ENGLISH NEWSPAPERS: A CORPUS-BASED ANALYSIS OF REPRESENTATION OF FEMALE POLITICIANS IN PAKISTAN

Farwa Khaliq
MPhil Scholar, Department of English, University of Okara
Email naeemfarwa8@gmail.com
Muhammad Kamran Abbas Ismail
Lecturer, Department of English, University of Okara (Corresponding Author)
Email Kamran.ch@uo.edu.pk
Javeria Mateen
MPhil Scholar, Department of English, University of Okara
Email: ljaveriamateen5@gmail.com
https://doi.org/10.5281/zenodo.20787023

Abstract

This study investigates the discursive positioning of female politicians in two of the most prominent English-language newspapers in Pakistan, Dawn and The News and how this positioning is achieved through the repetition of lexical and collocational patterns. While feminist studies on political language have documented its gendering in the West and South Asia (Lazar, 2005; Cameron, 2021), corpora-oriented research focusing specifically on the Pakistani English-language journalism corpus is scarce. This study was conducted through a purpose-built corpus consisting of 41 news articles from two newspapers, namely Dawn (20 articles) and The News (21 articles), covering top female political leaders such as Maryam Nawaz, Benazir Bhutto and Hina Rabbani Khar. These articles totaled 27083 running words. By means of word-frequency profiling and collocational analysis of the 4 target lemmas women, right(s), empower, and feminist, corpus-assisted discourse analysis was applied. The collocate networks and the frequency list were produced following standard procedures in corpus software, similar to those used for logDice- and likelihood-based extraction. The findings revealed that the strongest association was with representation, with women associating these constructions with constructions of empowerment, seats, tickets, legislators, and reserved, pointing to a discourse more about quota politics and procedural inclusion than substantive agency and policy-making. The collocate she used yielded a negative effect size, indicating underuse of the word compared to expectations, and rights were mainly associated with human, inheritance and property, placing women's rights at a legal rather than a political level. Empower, in turn, collocated with protect, digital, financial, and laws, foregrounding protectiveness and instrumentalisation over agency in the case of the first, and bringing back a narrow, infrequent collection of academic collocates (zones, pedagogies, intersectional) rather than collocates from political journalism. The results reveal that press discourse in English in Pakistan presents female politicians as visible objects, mainly in the language of quotas and protection, rather than in an agentive political register, which has traditionally been the domain of male politicians. The study contributes to the larger body of literature on language, gender and political legitimacy from a Pakistan's perspective supported by a corpus-based approach.

Keywords *The present study aims to investigate the presence of gendered discourse in the newspapers of the Pakistani media namely Dawn and The News based on the method of corpus linguistics. The corpus was drawn from two Pakistani newspapers namely Dawn and The News for a period of six months.The*

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

present study uses an approach of corpus linguistics to explore the gendered discourse in the Pakistani newspapers, Dawn and The News, for a corpus of six months.

1. Introduction

Language is not just an indicator of political reality: it is a way of producing the sense of legitimacy, agency and visibility that political actors seem to have (Fairclough, 1989; Van Dijk, 2001). For a country such as Pakistan, where women hold some of the most visible political roles in South Asia, and where they continue to be expected to act in a manner that aligns with traditional gender norms, the press is one such site where gendered political identity is negotiated (Jabeen & Shah, 2022). This media landscape is characterised by the predominance of the vernacular newspapers, the language of which is often rooted in the moral and honour-based rhetoric of women's political engagement, while the language of the English-language newspapers – which address the audience of the urban, educated – has been historically the rhetoric of rights, reform, and professional legitimacy (Rashid & Yousaf, 2023). This framing may represent real discursive equality, or simply an "alternate" way of reproducing the same inequities found in other gendered political speech.It is an empirical question whether this framing is a true equality in discourse, or simply a more refined version of the same inequities identified in other gendered political speech.

The literature on discourse-based corpus analysis is well-documented in terms of the fact that distinctive lexicosemantic structures can be identified as ideological stance, a stance not perceptible from impressionistic reading alone (Baker et al., 2008; Stubbs, 2001). In academic, journalistic and political registers, feminine markers are consistently found with classification terms, relational terms, and vulnerability terms; masculine markers are found with function or role, and institutional authority terms (Mosqueda & Sanchez, 2022; Pearce, 2008). In Pakistan, recent studies on political speeches (Tahir & Ahmad, 2025) reveal that men use assertive and authoritative language, whereas women use relational and accountability-based language. This study takes the next step in this field of research, examining the discourse of the press, which speaks on behalf of the politician, to investigate how the terms women, right(s), empower and feminist are collocated in the speech of the political actors.

It is argued here that the lexical environment for the portrayal of women in political reporting in Pakistani English-language press has been defined mostly by the terms of quota representation, legal entitlement, and protective empowerment as opposed to those of political agency, achievement, and institutional command which have traditionally dominated the portrayal of male officeholders. The argument is based on the collocational statistics of the target lemmas, which tends to be strongly collocated with representation, empowerment, seats, tickets, legislators, and reserved: a group of terms that overwhelmingly refer to the mechanics of reserved-seat politics rather than to substantive governance.

There's a good counter-argument. Some argue that it is true that these are salient quotas because the women who enter legislatures via reserved seats are a large proportion of the total number of women in these bodies, not because they are more likely to be created through discursive choices. On this interpretation, the press is merely reporting accurately on the institutional avenues that women typically use to enter the political arena, and the collocational pattern reflects political reality, rather than ideologically positioning the arena.

The high frequency of words related to women and quota, as well as protection-related words around women, together suggests that the lexical record is indeed a double-edged one: the salience of such words is very likely an artefact of the actual institutional arrangements in Pakistan, but the

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

absence of agentive, achievement-oriented collocates for women, and the comparatively low frequency of feminist as a journalistic lexical choice for all, cannot be accounted for by institutional constructs alone. The signs of the times suggest that the mass media in the country are reproducing, not critically questioning, the political structures in which female participation is institutionalised. Although there is a significant body of work on gendered political discourse in Pakistan, the following three areas are still lacking. First, there is the methodological gap as existing Pakistani studies of gender and politics have focused on the speech of politicians (Tahir & Ahmad, 2025; Mohsan et al., 2024) or on general news framing, where the analysis is qualitative FCDA and no collocational quantification is done (Jabeen & Shah, 2022; Rashid & Yousaf, 2023). Second, there has been no direct comparison of the lexical environment of the two most widely circulated English language Pakistani dailies, Dawn and The News, in so far as the representation of the female politicians as a discourse object is concerned. Thirdly, as lemmas like empower and feminist are ideologically loaded, the collocational behaviour of such ideologically loaded lemmas in Pakistani Journalistic English has not been documented yet in the domain of global feminist media discourse research (Gill & Orgad, 2023; Lazar, 2020).

This study fills these gaps by trying to answer the following research questions:

RQ1: What do the most frequently and statistically significant collocates of women, right(s), empower and feminist tell us in a corpus of political reporting of female politicians by Dawn and The News?

RQ2: What do these collocational patterns show in relation to the discursive construction of female politicians' agency, legitimacy and visibility in the press discourse in Pakistan in English?

RQ3: How does the lexical profile of women in this Corpus compare to and/or contrast with the gendered representation found in previous corpus-based discourse studies on Pakistani and international political and academic discourse?

The results of this study should be useful to a number of audiences. The findings could be used by journalists and editors of English language media in Pakistan to reflect critically on routine lexicalisation practices that, though not necessarily deliberate, affect the political legitimacy of women in power. Corpus-assisted discourse analysis and feminist critical discourse analysis scholars benefit from an empirical case study of Pakistan that can be compared with other studies of academic discourse in the Philippines (Legaspi-Torres, 2026), Pakistani political discourse (Talish, 2026), and Pakistani broadcast and print news (Jamshaid & Saeed, 2026). Finally, civil society groups and gender-policy practitioners who are interested in the media's role in influencing perceptions of women's political competence can use the study results to inform their work on media literacy and/or newsroom trainings.

2. Literature Review

The literature pertinent to this study can be categorised around three related themes: theorisation of language and gendered power in critical and feminist discourse analysis, corpus-based studies of the lexical construction of feminine and masculine social actors, and the specific discourse treatment of women's political participation in Pakistani media and political discourse. These themes are then taken up one by one, concluding with a statement of the gap filled by this study.

2.1 Theme One – Feminist Critical Discourse Analysis and the Linguistic Construction of Gendered Power

Feminist Critical Discourse Analysis (FCDA), a framework proposed by Lazar (2005, 2007), is an extension of mainstream Critical Discourse Analysis which puts patriarchy and gender ideology

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

on its analytical agenda rather than on the sidelines. Lazar's framework views discourse not as a medium that merely conveys social facts but as a place where a gender hierarchy is created, legitimated and sometimes challenged. This theoretical shift is relevant to the analysis of press, as news discourse often performs as being objective and factual yet at the same time makes assumptions regarding the constitution of the speaker as authoritative and the other as spoken about. A feminist discourse analysis of the gendered power relations which was applied to the news media in English and Urdu of Pakistan revealed that women are often depicted as victims, subject to moral regulation, as domestic actors or as mediated voices, and men as institutional actors, experts, interpreters or as legitimate public actors. That same study also noted that English-language media tend to portray women in the discourses of rights, reform, education and professionalism, while Urdu-language media portray women in the discourses of family honour, morality, modesty and communal control (Jamshaid & Saeed, 2026).

Theoretical inputs are also derived from research related to language and sexism in general. The gendered nature of language has long been a subject of research and debate among feminist scholars, who have suggested that the language used for masculinity is the linguistically unmarked 'default' while that used for femininity is the marked, exceptional category that needs to be explicitly signalled through the use of appropriate lexical markers (Cameron, 2021; Mills, 2008). This imbalance is supported by empirical studies in the corpora of gender markers in academic and journalistic discourse of non-Western countries. A collocational analysis of Filipino academic journal articles revealed that it is the constant need to qualify women's gender through use of modifiers like "female," "women," or "feminine" that potentially validates that women are visible but not the assumed norm, while men, on the other hand, are functionalised through expressions such as "men entrepreneurs," "men farmers" or "male," etc., which show that they are the unmarked gender (Legaspi-Torres, 2026). Theoretically it is important as it shows that the high number of feminine markers does not automatically mean empowerment; the same study also concluded that a high number of feminine markers does not automatically mean that women are visible – statistics do not in themselves indicate empowerment.

Another strand within this theme is the overlap between visibility and what Gill and Orgad have labeled 'confidence culture' – the use of feminist language (empowerment, confidence, resilience) in neoliberal media discourse without a significant challenge to structural inequality (Gill & Orgad, 2017; 2023). This is relevant to the present research's focus on the lemma empower because it suggests that in using empowerment related words, journalistic actors could also include protective or instrumental framings in addition to words that index autonomous political agency.

2.2 Theme Two: Corpus-Assisted Methods for Detecting Gendered Lexical and Collocational Patterns

A second set of literature relates to the methodological arguments for employing corpus-based approaches, and collocational analysis in particular, to identify gender bias which might not be obvious through qualitative reading. Corpus-assisted discourse studies (CADS) are a way to combine systematic and replicable quantitative methods with interpretive depth, enabling researchers to identify frequency effects, collocational preferences, and semantic associations that would otherwise not be seen (Baker et al., 2008; Gillings et al., 2023). In this tradition, the collocational analysis based on logDice or likelihood-ratio is a common method for identifying which words co-occur with significant regularity and not by chance (Legaspi-Torres, 2026).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The results of several studies of the application of this method to gender-marked lemmas are remarkably similar in different national and linguistic contexts. In a study of the Spanish-speaking press, patterned sexist stereotyping was identified in the collocational behaviour of the collocators man and woman, and the study of the British National Corpus (BNC) also revealed systematic differences in the collocational behaviour of man and woman, with the latter being associated more often with collocators that referred to relational and appearance-based descriptors (Mosqueda & Sanchez, 2022; Pearce, 2008). A similar pattern has been identified in the study of literary and journalistic English where it was observed that the word "medical men" has a collocational behavior that is in contrast to the "mad women" in light of the long-standing association of femininity with irrationality or vulnerability in English-language print discourse (Jevric, 2017). A gender-comparative corpus analysis of Pakistani political speeches revealed that while male politicians tended to have a greater number of assertive or authoritative registers, the female speakers tended to have a higher number of relational and polite type of lexical choices, which was further confirmed when focused on the key words and collocations as male speakers' key words and collocations were clustered around corruption, mandate, enemies and accountability, while the female speakers' key words and collocations were clustered around development, education, health and families (Talish, 2026).

From a methodological point of view, this literature has determined that a word count is not enough by itself: what matters is the collocational context in which the target lemma is used, as the context illuminates the implicit propositional content that is usually associated with the common use of a lemma. Furthermore, it provides a clear standard to which other parameters of collocational extraction (in this case, minimum frequency thresholds, a defined collocational span, and likelihood- or logDice-based significance testing) can be benchmarked as the field standard for collocational extraction, with which the present study's procedure can be compared (Legaspi-Torres, 2026).

2.3 Theme Three: Gendered Political Discourse in Pakistani political sphere and press

The third theme takes a more specific look at Pakistan, where gender and language, class and postcolonial linguistic hierarchy overlap in unique ways. It has been established through existing scholarship that the language of the media (English versus Urdu) is not a stylistic but also an ideological matter. The study of Pakistani news reports on the phenomenon of women's selective visibility through feminist discourse analysis revealed that English-language news media frames women more frequently in the discourses of rights, reform, education, and professional participation, which are all within the framework of the maintenance and normalisation of patriarchy (Jamshaid & Saeed, 2026). The same study also revealed that in news reports of violence, women are often described in passive voice, thereby emphasising their vulnerability to violence and obscuring the responsibility of the man, such as "a woman was assaulted" and "the victim was subjected to violence.

This is backed up by the research on political speech in Pakistan from the other side. A corpus-assisted comparative analysis of the public discourse of male and female politicians in Pakistan revealed that female politicians' discourse was constructed based on inclusion, service, and joint progress, a discourse that legitimises leadership through accountability and care and is not as confrontational as could be found in the public discourse of the male politicians, which tended to be driven by urgency (Talish, 2026). While that study focused on the speech of politicians rather than newspaper reporting on politicians, that does not mean the findings were not relevant here; it

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

is clear that there is a discursive pattern of gendered rhetorical asymmetry in Pakistani political life, and journalism can either reinforce or challenge it.

The expectation that quota and procedural vocabulary will take center stage in press coverage of female politicians specifically is further bolstered by comparative international evidence. Studies on institutional discourse, such as UN documents or institutions like the EU, conducted in the cross-national perspective, have shown that women tend to be classified in formal institutional discourse as the empty object of the verb, and studies on collocations have produced consistent results in that feminine-referring lemmas collocate disproportionately with classification and relational identification, but not with independent agency and functional identification.

This collection of works, taken together, demonstrates that the Pakistani press discourse is clearly gendered on the level of transitivity, modality and source quotation (Jamshaid & Saeed, 2026); that Pakistani political speech itself is gendered in a rhetorical parallel manner (Talish, 2026); and that collocational analysis is a validated and replicable method of detecting gendered asymmetries, at the level of the lexicon, in the academic, journalistic and political registers internationally (Legaspi-Torres, 2026; Mosqueda & Sanchez, 2022; Pearce, 2008).

2.4 Research Gap

This literature lacks a study based on systematic collocational analysis applied specifically to the lexical environment of female politicians in the discourse of English language newspapers of Pakistani context, studying two most prominent English-language dailies of Pakistan (Dawn and The News) as a combined corpus object. Existing FCDA work on Pakistani news (Jamshaid & Saeed, 2026) focusses on close qualitative reading of transitivity and modality rather than on quantified collocate extraction; existing corpus work on Pakistani political discourse (Talish, 2026) uses data from the speech of politicians rather than a politically-focused South Asian news corpus; and existing collocational work on gender markers (Legaspi-Torres, 2026; Pearce, 2008) is derived from academic or general corpora. This study seeks to fill this void: it is conducted on a special corpus of news reporting in Dawn and The News on female politicians, by using two methods of analysis, namely frequency and collocational analysis, and it examines four lemmas that are theoretically motivated, namely women, right(s), empower, and feminist, to find out how these items lexically construct the legitimacy of the female politicians they report.

3. Methodology

The research design used in this study was Corpus-assisted discourse analysis, both quantitative counting the frequency of the words and collocational analysis and qualitative interpretation based on feminist critical discourse analysis (Lazar, 2005). This design was chosen because it enables systematic and replicable identification of lexical patterning across a corpus of politically-oriented news texts while retaining the ability to interpret patterns within their socio-political context (Baker et al., 2008; Gillings et al., 2023).

3.1 Research Design

The design used was descriptive-quantitative that is based on the use of a corpus. The study was not designed to test a pre-registered hypothesis using inferential statistical modelling, but rather, in keeping with the tradition of corpus-assisted discourse studies (CADS) it sought to identify statistically salient lexical and collocational patterns and interpret them qualitatively (Baker et al., 2008).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.2 Corpus and Sample

The texts in the corpus were 41 articles from two English language newspapers of Pakistan, Dawn (20 articles) and The News (21 articles), with a total of about 27,083 running words. The articles were selected within the period from January 2021 to December 2025. All articles were purposively selected on the basis that the main subject of the article was a female political figure from Pakistan in a formal political role or challenging it, such as Maryam Nawaz, Benazir Bhutto, Hina Rabbani Khar and Sherry Rehman.

The articles were manually collected from Google and from the official website of Dawn (dawn.com) and The News (thenews.com.pk) using the keywords "female politicians", "gender discourse", and "corpus-based" and then identifying the articles pertaining to this topic in the time period of 2021 to 2025. Using Notepad, articles that met the selection criterion were copied directly from the source webpage into a plain-text file, and each article was labelled with source outlet, sequential article number and publication date, providing traceability to the original web source. The manual, key word guided retrieval procedure is typical of the type of sampling procedures found in corpus-assisted discourse studies of a delimited topical area (Baker et al., 2008).

3.3 Instrument: AntConc

Corpus processing and analysis were done using a freeware concordance and text-analysis tool kit, AntConc (Anthony, 2019). The 41 articles were put in a single plain-text corpus file, each article is marked with the source outlet and its number in the sequence, and the publication date; and loaded into AntConc as the target corpus. The two built-in tools which were used in AntConc were Word List, which generates and outputs a ranked frequency list of all the word types in the corpus, and Collocates, which lists the words that co-occur with a search word (the node word) more than would be expected by chance within a specified window span (Anthony, 2019). No manual tweaking was done to the window span or minimum collocate frequency prior to extraction; default AntConc settings were used.

3.4 Procedure

The Word List tool was then used on the entire 41-article corpus to provide an overall frequency profile, in which all word types were ranked by raw frequency, range (dispersion) and normalised frequency per million words. Based on the theoretical justification from the literature review, four target lemmas were chosen for the collocational analysis, namely: women, right(s), empower, and feminist. The ideological weight each of these lemmas carries in the field of feminist and gender-discourse scholarship is different: women as the primary gendered social-actor term, right(s) as an index of entitlement discourse, empower as an index of agency discourse, feminist as an index of explicit ideological self-identification.

The search word for each target lemma was entered in the Collocates tool of AntConc, which first needed the generation of a Word List for the loaded corpus (built-in dependency of the collocates tool) before the calculation can take place. The Collocates tool was configured to do the calculations for Log-Likelihood and Effect Size, as the statistical measures of association displayed in the output tables for this study (Anthony, 2019). All four lemmas have been extracted using the default window span and default minimum collocate frequency in AntConc's Collocates tool, which means that the extraction procedure was consistent for the four lemmas. The rank, scaled frequency, the frequency of the co-occurrence on the left side of the collocate, the frequency of the co-occurrence on the right side of the collocate (Freq L and Freq R), Range (dispersion

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

across the corpus), Log-Likelihood value, and Effect Size value are reported by AntConc for each collocate returned.

3.5 Data Analysis

The frequency data generated using the Word List tool was descriptively analysed to gain an overall picture of the lexical profile of the corpus, focusing on content words related to political institutions, named politicians and gender. Each of the four target lemmas' collocational data was exported from the Collocates tool, ranked by Log-Likelihood, the measure used by AntConc to show the strength of a co-occurrence relationship, and complemented by Effect Size, which shows strength of the lexical attraction, independent of raw frequency (Anthony, 2019). Then, collocates were qualitatively interpreted based on the semantic or the ideological categories they represent (classification, relational identification, institutional or quota mechanism, protective framing, or agentive framing), using the interpretive logic developed in previous collocation-based gender discourse studies (Legaspi-Torres, 2026; Mosqueda & Sánchez, 2022). Because the study was designed and carried out descriptively, with a corpus-assisted approach, no inferential statistical tests apart from the Log-Likelihood and Effect Size measures that are produced naturally by the collocates tool of AntConc were employed.

3.6 Ethical Considerations

No private, unpublished, or personal data were collected, and there was no direct recruitment, surveying or interviewing of human participants; all source texts were from publicly available, professionally published news articles. Formal ethical clearance was not required for this design, in keeping with previous research on discourse in publicly published journalistic and academic texts (Legaspi-Torres, 2026) that deals with the publicly circulated linguistic record of professional news reporting instead of private individuals' data.

4. Results

This section presents the descriptive frequency and collocational results of the 41-article Dawn and The News corpus. The results are organised as follows: In the first part, an overall profile of the frequency of the target words in the corpus is provided, followed by the collocational profile of each of the four target lemmas: women, right(s), empower, and feminist. No interpretations of these findings are made in this section - interpretation is left for the Discussion.

4.1 Overall Corpus Frequency Profile

The combined corpus comprised 41 articles (Dawn, n = 20; The News, n = 21) totalling approximately 27,083 running words. The most frequent content words, excluding grammatical function words, were she (n = 282; rank 9), women (n = 274; rank 11), said (n = 192; rank 14), Pakistan (n = 191; rank 15), and minister (n = 185; rank 16). The pronoun her occurred 144 times (rank 20), while the gender-neutral political noun government occurred 101 times (rank 29). Named political figures appeared with substantial frequency: Maryam (n = 81; rank 36), Bhutto (n = 68; rank 42), Nawaz (n = 64; rank 44), and Benazir (n = 63; rank 46). The lexical item gender itself occurred 57 times (rank 51), and rights occurred 32 times (rank 95). Table 1 presents the twenty highest-ranking content words from the frequency list, excluding closed-class function words (articles, prepositions, auxiliary verbs, and conjunctions).

Table 1. *Top Content Words in the Combined Dawn/The News Corpus (Excluding Function Words)*

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Rank	Word	Frequency	Normalised Frequency
9	she	282	10,342.93
11	women	274	10,049.51
14	said	192	7,041.99
15	pakistan	191	7,005.32
16	minister	185	6,785.26
20	her	144	5,281.50
29	government	101	3,704.38
31	punjab	92	3,374.29
36	maryam	81	2,970.84
42	bhutto	68	2,494.04
44	nawaz	64	2,347.33
46	benazir	63	2,310.65
47	elections	61	2,237.30
49	country	59	2,163.95
49	education	59	2,163.95
51	gender	57	2,090.59
56	assembly	50	1,833.85
57	chief	49	1,797.18
60	climate	48	1,760.50
95	rights	32	1,173.67

The frequency profile showed that political-institutional vocabulary (minister, government, assembly, chief, elections) and named-individual vocabulary (Maryam, Bhutto, Nawaz, Benazir) together accounted for a substantial share of the corpus's most frequent content words, alongside the core gendered terms she, women, and her.

4.2 Collocates of "Women"

Collocational analysis of the lemma women returned eleven statistically significant collocates within the extraction parameters applied. The strongest association by log-likelihood was the genitive marker s (LL = 37.774, effect size = 1.239), followed by representation (LL = 35.796, effect size = 2.702), empowerment (LL = 33.934, effect size = 2.829), and the pronoun she (LL = 29.728, effect size = −2.503). The collocate she was the only collocate in this set to return a negative effect size. Procedural and quota-related terms were prominent among the remaining

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

collocates: seats (LL = 29.718, effect size = 2.241), tickets (LL = 23.064, effect size = 3.315), legislators (LL = 23.064, effect size = 3.315), and reserved (LL = 22.251, effect size = 2.264). The collocate men returned a log-likelihood of 19.726 and an effect size of 2.993, while their (LL = 17.325, effect size = 1.446) and rights (LL = 17.091, effect size = 2.015) completed the eleven-item set. Table 2 reports the full collocate list for women.

Table 2. *Collocates of "Women" (Ranked by Log-Likelihood)*

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
1	s	2740	6	59	37.774	1.239
2	representation	260	7	10	35.796	2.702
3	empowerment	210	2	13	33.934	2.829
4	she	2820	0	5	29.728	-2.503
5	seats	400	8	11	29.718	2.241
6	tickets	80	7	1	23.064	3.315
6	legislators	80	0	8	23.064	3.315
8	reserved	290	8	6	22.251	2.264
9	men	100	6	2	19.726	2.993
10	their	840	5	18	17.325	1.446
11	rights	320	8	5	17.091	2.015

4.3 Collocates of "Right(s)"

Collocational analysis of right(s) returned six significant collocates. The strongest association was human (LL = 63.515, effect size = 5.927), the highest log-likelihood value recorded among all four target lemmas in this study. This was followed by inheritance (LL = 23.520, effect size = 5.605) and women (LL = 17.091, effect size = 2.015). Two collocates were tied at rank four: property (LL = 14.042, effect size = 6.413) and advancing (LL = 14.042, effect size = 6.413). The collocate reproductive returned a log-likelihood of 13.234 and the highest scaled frequency among the lower-ranked items (effect size = 4.538). Table 3 reports the full collocate list for right(s).

Table 3. *Collocates of "Right(s)" (Ranked by Log-Likelihood)*

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
1	human	140	10	0	63.515	5.927
2	inheritance	70	3	1	23.520	5.605
3	women	2740	5	8	17.091	2.015
4	property	20	2	0	14.042	6.413

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
4	advancing	20	1	1	14.042	6.413
6	reproductive	110	3	0	13.234	4.538

4.4 Collocates of "Empower"

Collocational analysis of empower returned five significant collocates. The strongest association was protect (LL = 27.061, effect size = 7.869), the highest effect size recorded among the women collocate set but lower than several values recorded for right(s). This was followed by digital (LL = 21.254, effect size = 6.490) and financially (LL = 20.133, effect size = 8.605), the latter returning the single highest effect-size value across all four lemmas examined in this study. The collocate laws returned a log-likelihood of 12.525 (effect size = 5.905), and women itself appeared as a collocate of empower at rank five (LL = 11.298, effect size = 2.829). Table 4 reports the full collocate list for empower.

Table 4. *Collocates of "Empower" (Ranked by Log-Likelihood)*

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
1	protect	50	3	0	27.061	7.869
2	digital	130	2	1	21.254	6.490
3	financially	20	1	1	20.133	8.605
4	laws	130	2	0	12.525	5.905
5	women	2740	1	4	11.298	2.829

4.5 Collocates of "Feminist"

Collocational analysis of feminist returned the smallest and lowest-frequency collocate set of the four target lemmas, comprising seven items at considerably lower scaled frequencies than those recorded for women, right(s), or empower. The strongest associations, tied at rank one, were zones (LL = 11.763, effect size = 9.828) and pedagogies (LL = 11.763, effect size = 9.828) the latter returning the single highest effect size recorded in the entire dataset. A further five collocates were tied at rank three, each with a log-likelihood of 10.327 and effect size of 8.828: aligns, lacks, explicit, intersectional, and perspective. Table 5 reports the full collocate list for feminist.

Table 5. *Collocates of "Feminist" (Ranked by Log-Likelihood)*

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
1	zones	10	1	0	11.763	9.828
1	pedagogies	10	0	1	11.763	9.828
3	aligns	20	1	0	10.327	8.828
3	lacks	20	1	0	10.327	8.828

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Rank	Collocate	Freq (Scaled)	Freq L	Freq R	Likelihood	Effect Size
3	explicit	20	1	0	10.327	8.828
3	intersectional	20	1	0	10.327	8.828
3	perspective	20	0	1	10.327	8.828

4.6 Cross-Lemma Comparison

Among the four target lemmas, only the lemma of women had a negative effect size for the collocate she (–2.503) and only the collocate set of women was dominated by the quota- and procedure-related vocabulary (representation, seats, tickets, legislators, reserved). The single highest log-likelihood in the data set was Right(s) (human, 63.515) and was the only lemma whose collocate set was anchored in legal and bodily-entitlement vocabulary (human, inheritance, property, reproductive). Empower was the highest individual effect-size value among lemmas that occurred more than the minimum corpus value (financially: 8.605) and was the only lemma whose collocate set contained protective vocabulary (protect, laws), and instrumental vocabulary (digital, financially). Feminist had the lowest absolute frequencies of all the lemmas examined, the highest effect sizes for all lemmas (zones and pedagogies, 9.828), and the only collocate set that was composed entirely of academic register words (pedagogies, intersectional, perspective) and not journalistic or political register.

5. Discussion

5.1 Summary of Key findings

The collocational analysis revealed that women in the Dawn/The News corpus are lexico-semantically located in the collocational field of quota-based political representation (representation, empowerment, seats, tickets, legislators, reserved), but not in the collocational field of independent political achievement, command within the institution, or policy formulation. The collocate she got the sole negative effect size of all in the whole data set, which means that, according to the statistics, women are less likely to occur alongside the third-person feminine pronoun co-occurrence than what would be predicted by chance in this corpus. Right(s) was connected almost exclusively to legal and bodily-entitlement vocabulary (human, inheritance, property, reproductive), and empower combined protective vocabulary (protect, laws) with instrumental vocabulary (digital, financially) instead of vocabulary of self-directed political agency. The lemmas of feminist returned the smallest collocate set, of low frequency and narrow academic vocabulary of the four lemmas, and there were no collocates from the discourse of mainstream electoral and institutional politics in Pakistan.

5.2 Comparison to the Literature

These findings are quite similar to previous corpus studies that examined gender markers in non-Western academic and journalistic discourse. The collocational profile of women in this corpus consisting mainly of classification- and procedure-oriented terms, and lacking in functionalising and agency-marking terms directly corresponds to the fact that with the high frequency of feminine markers, there is a constant need for gender identification among women, so that statistical visibility does not necessarily imply empowerment (Legaspi-Torres, 2026). The use of seats, tickets, legislators, and reserved as collocates to women also aligns with what the same study found: that terms such as “female leaders” or “women entrepreneurs” may signal women’s

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

presence in certain contexts, but suggest that their presence is somehow "procedurally" different than the norm represented by the more general term "politician".

The empower result also echoes the notion of 'confidence culture' in the media, which uses feminist terminology like empowerment and situates structural inequalities as background concepts (Gill & Orgad, 2017, 2023). In contrast, in the current corpus empower is not paired with language of political action directed by women themselves, but with protect, laws, digital, and financially, which suggest that women are being empowered by external protective or instrumental means. The majority of the participants in the study expressed a view that aligned with FCDA's findings, which argue that Pakistani media discourse not only mirrors social inequality but also plays a significant role in sustaining and normalising patriarchal power, even in apparently progressive and rights-oriented English-media reporting (Jamshaid & Saeed, 2026).

The collocational profile of right(s) with its human, inheritance, property and reproductive templates sets women's rights in a legalistic and bodily discourse, but not a directly political one. This is broadly consistent with the observation that English-language Pakistani media more frequently feature women in a rights-based narrative that focuses on their rights to reform, rights to education, and rights to professional involvement (Jamshaid & Saeed, 2026); and in this relatively progressive rights-based approach, the rights most closely associated with women in this corpus are those relating to gendered rights to property and inheritance.

Lastly, the high de-feminisation and the narrow and academic collocate profile of feminist is a bit surprising, given the high prevalence of feminist movements and feminist vocabulary in international protest and political discourse studies (Irfan et al., 2026). While slogans in the global feminist movement have been seen to utilise agentive and explicitly self-identifying language in a short, concise manner (Irfan et al., 2026), the word feminist in this Pakistani press corpus rarely collocates with words from electoral or institutional political reporting, but rather with words from the academic register, indicating that explicit self-identifying feminist language is less present, or purposely avoided in mainstream reporting specifically of female politicians.

5.3 Interpretation: How These Results May Have Arisen

These patterns can be explained by a number of non-exclusive explanations. Secondly, the constitutional guarantee for women's legislative representation has made women's presence in the institutions a real institutional fact, and the frequency of the collocates seats, tickets, legislators, and reserved may not only be a reflection of the proper use of these terms but also of the actual presence of women in legislative institutions in Pakistan. Second, English-language Pakistani journalism's focus on an urban, rights-aware audience may account for its association with the legalistic register of human, inheritance, and property rather than with explicitly political register words; this is in line with the overall finding that English-language media tend to favour discourses of rights and reform, while Urdu-language media are more likely to support honour-based discourses (Jamshaid & Saeed, 2026). Thirdly, the limited and academic nature of the node 'feminist' may also be a reflection of the tendency of mainstream political reporting in Pakistani media to be wary of explicitly feminist self-labelling, which is consistent with the overall observation that feminist resistance and counter-discourse in Pakistani media are limited and contestable even when they are present (Jamshaid & Saeed 2026).

6. Limitations

The design and data of this study has a number of limitations. First, the corpus consisted of 41 articles and around 27,083 running words which is adequate for the descriptive collocational

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

analysis but too small to allow for robust generalisation to the wider population of political reporting in Dawn and The News over time. Secondly, the corpus fused the two newspapers (Dawn and The News) into a single frequency and collocate lists; the data did not allow for a collocational comparison from each newspaper individually; hence, any difference in the lexical practice of the two newspapers could not be isolated in the present study. Third, the four target lemmas were chosen theoretically before analysis: other lexemes which could equally illuminate the discursive construction of female politicians were outside the scope of this collocational extraction, for example, titles specific to leadership or evaluative adjectives. Fourth, details of the materials used to generate the underlying frequency and collocate tables were not fully described, and this makes it hard to replicate the procedure for other researchers in order to obtain the same results. Fifthly, this is a corpus-assisted study so the analysis is a description of the patterns of co-occurrence, and cannot determine whether the lexical patterns identified are due to intentional framing by the editor, an unconscious linguistic habit of the author, or whether they reflect the actual institutional structure in accurate ways; the Discussion above should be read as providing plausible accounts and interpretations of the patterns found, and not necessarily as causal explanations.

7. Implications

7.1 Theoretical Implications

The results further the work of Feminist Critical Discourse Analysis and Corpus-based Gender-marker research (Lazar, 2005; Legaspi-Torres, 2026) by showing that the corpus of a politically-oriented South Asian English language press corpus is not only characterised by the procedural/protective lexical pattern documented in the academic and general journalistic corpora, but it is also specifically so. This implies that "visibility without agency" (Legaspi-Torres, 2026), which was found in the non-political academic discourse, is not unique to that register, but is also found in the explicitly political register of news reporting, which provides cross-register support to the overall theoretical statement that women's visibility in discourse does not necessarily mean empowerment.

7.2 Practical Implications

The findings give journalists and editors a tangible, hands-on observation: When describing women's political activity, journalists and editors may routinely use words associated with the quota and procedure (seats, tickets, reserved), even if factually accurate, and this can have discursive effects that indicate that women politicians are the politically exceptional rather than the politically general. To achieve a more balanced representation of both genders, editors may want to make a special effort to change the lexical context of female political actors from one dominated by achievement terms to one that includes more terms that concentrate on the agency to implement policy, as well as terms that accurately describe the procedures involved in the action. The study can provide concrete examples of corpus-based lexical habit, which in the absence of explicit gender bias can re-create gendered asymmetries in news coverage of political phenomena in Pakistan, for newsroom trainings for media literacy teachers and civil society organisations engaged in the issues of gender and media in Pakistan.

8. Future Research Directions

There are some specific extensions of the present study that are suggested for future research. First, a much bigger corpus, separately coded by outlet (Dawn versus The News) and separately coded by language (English versus Urdu) would make it possible to compare the two outlet-specific and the two language-specific collocational profiles, in addition to the qualitative comparison already

noted in previous FCDA research (Jamshaid & Saeed, 2026). Second, further research on the set of target lemmas should be expanded to include explicit leadership titles (e.g., chief minister, chairperson) and evaluative adjectives, to find out if the procedural/protective pattern found in the present study is also true for the functionalising vocabulary of women's institutional positions. Thirdly, a diachronic extension of the collocational profile, looking at changes before and after significant political events (for instance, the swearing-in ceremony of Maryam Nawaz as Chief Minister of Punjab) would be useful to establish if the identified lexical patterns in this profile are indeed stable or change over time in Pakistani press discourse. Fourthly, in future work, a set of matched male politicians could be taken and a collocational profile of women politician could be compared to the collocational profile of male politician to test against each other; not just against general literature alone.

9. Conclusion

The aim of this study was to fill the gap between qualitative FCDA studies in the literature on Pakistani news (Jamshaid & Saeed, 2026), corpus-based studies on Pakistani political speech (Talish, 2026), and collocational gender-marker studies based on non-political corpora (Legaspi-Torres, 2026) by determining the discursive identity of female politicians as constructed in the lexicon of two leading news sites: Dawn and The News in Pakistan. Using a 41-article, 27,083-word corpus and collocational analysis of four target lemmas, the study showed that women's collocates are more likely to be quota- and procedure-related (representation, seats, tickets, legislators, reserved), than agentive; right(s) is anchored in a legalistic register of bodily entitlement (human, inheritance, property, reproductive); empower does not include any agentive collocates, but rather a protective and instrumental register (protect, laws, digital, financially); and feminist returns a narrow, low-frequency, academic-register collocate set largely absent from mainstream political vocabulary. In response to the guiding questions of the study, these findings show that the collocational profile of female politicians in this corpus is one of procedural visibility and protective entitlement, not of general political authority as expressed in an agentive and achievement-oriented register as would be expected, and, more importantly, that this profile is generally in keeping with, rather than divergent from, previous corpus-based studies of gendered discourse in Pakistani and international academic and political contexts. The point for working journalists and editors is rather simple: reporting accurately about the processes of women's political representation doesn't have to involve a necessarily clichéd reference to the political agency and success of the women involved, and careful variations in commonly used terms can make a meaningful impact on readers' perceptions of the political legitimacy of the women those journalists cover.

References

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, 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. https://doi.org/10.1177/0957926508088962

Cameron, D. (2021). Feminism and linguistic theory (3rd ed.). Palgrave Macmillan.

Coates, J. (2015). Women, men and language (3rd ed.). Routledge.

Dada, S., Ashworth, H. C., Dhatt, R., et al. (2021). Words matter: Political and gender analysis of speeches made by heads of government during the COVID-19 pandemic. BMJ Global Health, 6(1), e003910. https://doi.org/10.1136/bmjgh-2020-003910

Fairclough, N. (1989). Language and power. Longman.
Fairclough, N. (1992). Discourse and social change. Polity Press.
Fairclough, N. (1995). Critical discourse analysis: The critical study of language. Longman.
Gill, R. (2007). Gender and the media. Polity.
Gill, R., & Orgad, S. (2017). Confidence culture and the remaking of feminism. New Formations, 91, 16–34.
Gill, R., & Orgad, S. (2023). Confidence culture. Duke University Press.
Gillings, M., Mautner, G., & Baker, P. (2023). Corpus-assisted discourse studies. Cambridge University Press. https://doi.org/10.1017/9781009168144
Hassan, R., & Ahmed, N. (2023). Digital feminist activism and #MeTooPakistan. Feminist Media Studies, 23(4), 678–695.
Irfan, M. S., Iqbal, R. H., & Rakha, A. (2026). Language & power: A corpus-based discourse analysis of protest slogans. Liberal Journal of Language & Literature Review, 4(2), 925–964.
Jabeen, T., & Shah, N. (2022). Language, media, and postcolonial patriarchy: A feminist discourse analysis of Pakistani news channels. Asian Journal of Communication, 32(4), 355–372.
Jamshaid, W., & Saeed, M. I. (2026). Feminist critical discourse analysis of gendered power relations in Pakistani news media texts. Social Science Review Archives, 4(2), 731–741. https://doi.org/10.70670/sra.v4i2.2116
Jevric, T. (2017). “Medical men” and “mad women”: A study into the frequency of words through collocations. [sic] – A Journal of Literature, Culture and Literary Translation, 8(1). https://doi.org/10.15291/sic/1.8.lc.2
Kiani, F. (2023). Gender stereotypes in Pakistani television dramas. Pakistan Journal of Media Studies, 8(2), 89–104.
Lakoff, R. (2004). Language and woman’s place. Oxford University Press.
Lazar, M. M. (2005). Feminist critical discourse analysis: Gender, power and ideology in discourse. Palgrave Macmillan.
Lazar, M. M. (2007). Feminist critical discourse analysis: Articulating a feminist discourse praxis. Critical Discourse Studies, 4(2), 141–164.
Lazar, M. M. (2020). Feminist critical discourse analysis and the politics of gender. Critical Discourse Studies, 17(4), 421–437.
Legaspi-Torres, K. D. (2026). The naturalised men and over-represented women: A collocational analysis of gender markers in Filipino academic discourse. SUKISOK Journal of the Arts and Sciences, Special Issue, 33–47.
Machin, D., & Mayr, A. (2012). How to do critical discourse analysis: A multimodal introduction. Sage.
Mills, S. (2008). Language and sexism. Cambridge University Press.
Mills, S., & Mullany, L. (2011). Language, gender and feminism: Theory, methodology and practice. Routledge.
Mohsan, M., Anwar, M., & Nawaz, S. (2024). Unveiling ideology and discourse: A critical analysis of Maryam Nawaz’s oath-taking ceremony. Pakistan Languages and Humanities Review, 8(3), 420–431. https://doi.org/10.47205/plhr.2024(8-III)29
Mosqueda, H. C., & Sánchez, I. R. (2022). Sexist stereotypes in the Spanish-speaking press: A collocational analysis of the lemmas man and woman. Revista de Humanidades Digitales, 7, 57–79. https://doi.org/10.5944/rhd.vol.7.2022.34108

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Pearce, M. (2008). Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. Corpora, 3(1), 1–29. https://doi.org/10.3366/E174950320800004X

Rashid, A., & Yousaf, M. (2023). Victimisation framing in Pakistani news coverage of gender-based violence. Journalism Studies, 24(6), 745–761.

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

Sunderland, J. (2004). Gendered discourses. Palgrave Macmillan.

Tahir, M., & Ahmad, F. (2025). Gendered language and power: A stylistic analysis of Imran Khan and Maryam Nawaz's speeches. Journal of Applied Linguistics and TESOL, 8(3). https://doi.org/10.63878/jalt1238

Talish, F. (2026). Gender, persuasion, and political communication: A comparative corpus-assisted discourse analysis of public addresses by Pakistani politicians. Social Science Review Archives, 4(1), 1528–1540. https://doi.org/10.70670/sra.v4i1.1683

Van der Pas, D. J., & Aaldering, L. (2020). Gender differences in political media coverage: A meta-analysis. Journal of Communication, 70(1), 114–143. https://doi.org/10.1093/joc/jqz046

Van Dijk, T. A. (1998). Ideology: A multidisciplinary approach. Sage.

Van Dijk, T. A. (2001). Critical discourse analysis. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), The handbook of discourse analysis (pp. 352–371). Blackwell.

Xue, J. (2025). Men as offenders while women as victims: A corpus-based study of men and women in the United Nations. Frontiers in Communication, 10. https://doi.org/10.3389/fcomm.2025.1535312