INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)
VOLUME 3, ISSUE 4, 2025 p-ISSN: 3007-2336 e-ISSN: 3007-2344

NAMING THE NEIGHBOUR: NATION, LEADER AND LEXICAL EMPHASIS IN INDIAN AND PAKISTANI NEWS ANCHORS' TWEETS

Sidra Shezadi sidrarif205@gmail.com	Lecturer in English, Department of English Punjab Group of Colleges, Daska, Pakistan University of Management & Technology, Sialkot Campus, Pakistan.
Sidra Ashraf	Senior School Teacher, Education Department, Government of Punjab University of Management & Technology, Sialkot Campus, Pakistan.
Rimsha Arif	University of Management & Technology, Sialkot Campus, Pakistan..

Abstract

Journalists use Twitter (now X) to foreground topics, actors and nations, yet comparative corpus studies of South Asian news anchors' platform language are scarce, and small purposive corpora are routinely compared using raw frequencies. This article re-analyses the aggregate lexical data of an existing four-part corpus of Indian and Pakistani news anchors' tweets (January 2020–December 2021; 1,865 tweets, 71,547 running words; Pakistani and Indian, male and female anchors) to compare seven pre-selected lexical items—Media, Army, Corona, Pakistan, India, Modi and Imran Khan—on a size-adjusted and statistically explicit basis. Frequencies were normalised per 10,000 words and compared using exact Poisson confidence intervals, log-likelihood (G^2), the Log Ratio effect size with 95% confidence intervals, and Holm correction for multiple comparisons. Five of the seven items differ clearly between national groups (Holm-adjusted $p < .001$): Pakistan (Log Ratio = 1.44) and Imran Khan (3.93) are more frequent among Pakistani anchors, while India (−2.25), Modi (−2.57) and Corona (−1.80) are more frequent among Indian anchors; Media does not differ. Indian anchors refer to the other country about twice as often as Pakistani anchors do (47.77 vs 23.71 per 10,000), whereas rates of naming the own-country leader are almost identical (39.50 vs 41.99). The pooled rate of the seven items is statistically indistinguishable across nations (256.28 vs 242.70 per 10,000; $p = .289$), so the larger Pakistani raw total (1,208 vs 558) reflects corpus size, not lexical density. Gender contrasts are smaller and mostly unreliable once multiplicity is controlled. The study offers a transparent, replicable template for small-corpus comparison and sets out explicitly which inferences aggregate counts can and cannot support.

Keywords: *Corpus Linguistics, Digital Journalism, Effect Size, Log Ratio, News Anchors, Normalised Frequency, South Asia, Twitter/X.*

Corresponding Author: Sidra Shezadi **(**University of Management & Technology, Sialkot Campus, Pakistan.)
Email: sidrarauf232@gmail.com

1. Introduction

Social media platforms have become important spaces in which journalists circulate news, comment on public affairs and address networked audiences. Twitter, now known as X, is particularly relevant to discourse research because its short, publicly visible posts allow journalists to publish commentary rapidly and to interact with audiences beyond conventional broadcast formats (Zappavigna, 2012). Research on journalism and Twitter has shown that practitioners adapt established professional norms to the platform while also developing new routines of commentary and self-presentation (Lasorsa, Lewis and Holton, 2012; Hermida, 2013). The words journalists choose to repeat on such platforms—the nations, leaders, institutions and events they name—provide observable evidence of the topics they make salient.

This is of particular interest in South Asia, where the relationship between India and Pakistan is a persistent object of media attention and where the period 2020–2021 combined that bilateral relationship with a global public-health emergency. Yet comparative, corpus-based evidence on how Indian and Pakistani news anchors differ in their platform language is limited, and the small purposive corpora that are feasible for such work are often compared on the basis of raw counts, which are uninterpretable when the groups differ in size (Baker, 2006; Oakes, 1998).

The present article revisits the quantitative component of an existing study of Indian and Pakistani news anchors' Twitter posts. Rather than claiming a new corpus, it takes the four sub-corpus sizes and the seven lexical-item frequency totals reported in that study and subjects them to a size-adjusted, effect-size-based and multiplicity-controlled comparison. The article makes three contributions. First, it provides a normalised descriptive profile of the selected items for national and gender groupings, with interval estimates. Second, it shows empirically how the conclusions drawn from raw totals change once corpus size is accounted for. Third, it specifies, and applies, a defensible inferential procedure for aggregate frequency data and states plainly what such data cannot establish.

The study addresses three research questions:

RQ1. How do the normalised frequencies of the seven selected lexical items differ between Pakistani and Indian anchors, and how large and how reliable are those differences?

RQ2. How does size-adjustment change the conclusions that raw totals would suggest, including which item is most frequent in each national sub-corpus?

RQ3. Within each country, do male and female anchors differ in their use of the selected items?

2. Background

2.1 Frequency comparison and normalisation in corpus research

Corpus linguistics treats language data as observable patterns in collections of naturally occurring texts. Frequency information is useful for identifying recurring lexical items, but its interpretation requires attention to corpus composition, sampling and the research question (Sinclair, 1991; McEnery and Wilson, 2001). Biber, Conrad and Reppen (1998) emphasise that quantitative findings become meaningful when related to linguistic context and corpus design, and Hunston (2002) cautions that corpus evidence supports claims about language use only in light of what the corpus represents. Comparative work conventionally normalises frequencies to a fixed base so that groups of unequal size can be compared: 100 occurrences in 10,000 words is not equivalent to 100 occurrences in 50,000 words (Oakes, 1998; Baker, 2006).

2.2 Keyness, effect size and statistical caution

Keyword analysis identifies items that are unusually frequent in one corpus relative to a reference corpus (Scott, 1997; Kilgarriff, 2001; Bondi and Scott, 2010), and log-likelihood is widely used to test frequency differences (Dunning, 1993; Rayson and Garside, 2000). However, statistical significance and the size of a difference are distinct, and researchers have argued for reporting effect-size measures such as Log Ratio alongside significance tests (Gabrielatos, 2018; Hardie, 2014). Two further problems are well documented. Word tokens are not independent observations, because they cluster within texts and speakers, so token-based tests are anti-conservative when words are unevenly dispersed (Kilgarriff, 2005; Brezina and Meyerhoff, 2014; Lijffijt et al., 2016). And a word concentrated in a few texts has a different distributional profile from one spread widely (Gries, 2008; Egbert and Biber, 2019; Jeaco, 2020; Gries, 2025). Because the present data are aggregate, dispersion cannot be measured; the article therefore uses frequency-level tests as exploratory, effect-size-led evidence and does not use the technical term keyword for its seven items, which are a fixed list rather than the output of a reference-corpus comparison.

2.3 Social-media corpora, digital journalism and gender

Social-media data raise specific methodological problems, including shifting platform practices, uneven user activity, reposting, sampling bias and the difficulty of defining a representative population (Boyd and Crawford, 2012; Moreno-Ortiz and García-Gámez, 2023). Quantitative patterns can map topic prominence, but should not be read as direct evidence of intention or audience effect without further data (Zappavigna, 2012). Corpus-assisted discourse studies commonly pair quantitative pattern identification with qualitative interpretation (Partington, 2004; Baker et al., 2008; Conrad, 2002); the present article supplies the quantitative stage only. Corpus approaches to gender caution

against assuming that observed differences reflect gender as such rather than topic, genre, role or individual variation (Baker, 2014), a caution that is especially relevant here because each sub-corpus draws on only a handful of accounts.

3. Data and Method

3.1 Design

The study is a quantitative, descriptive secondary analysis of aggregate data reported in the author's MPhil thesis. No tweets were collected and no new corpus was built for this article. All calculations are reproducible from the counts in Tables 1 and 2.

3.2 Source corpus

The source corpus comprises 1,865 tweets (71,547 running words) posted between January 2020 and December 2021 by Indian and Pakistani news anchors, divided into four sub-corpora by nationality and gender (Table 1). The original study selected five male and five female anchors per country purposively. The corpus is therefore not a sample of all journalists or all Twitter users in either country, and claims are restricted to the accounts studied. Mean tweet length varies (31.0 words in the Indian male sub-corpus; 40.5–44.7 in the others), which is a further reason to compare normalised rather than per-tweet figures.

Table 1. Composition of the four sub-corpora.

Sub-corpus	Tweets	Running words	Words per tweet
Pakistani male anchors	1,041	42,179	40.5
Pakistani female anchors	180	7,595	42.2
Indian male anchors	512	15,869	31.0
Indian female anchors	132	5,904	44.7
Total	**1,865**	**71,547**	**38.4**

Note. Pakistani (male + female): 49,774 words; Indian (male + female): 21,773 words.

3.3 Lexical items

Seven items were selected in the original study: Media, Army, Corona, Pakistan, India, Modi and Imran Khan. They span four categories: the two nations (Pakistan, India), the two heads of government in office for most of the period (Imran Khan, Modi), an institution (Army), and a topical term (Corona), together with a reflexive professional term (Media). Frequencies for each item in each sub-corpus are used as reported.

3.4 Normalisation and statistical procedure

Normalised frequency was calculated as (frequency ÷ running words) × 10,000. National figures pool the male and female sub-corpora (frequencies and word totals are summed before dividing). Uncertainty in each normalised frequency is expressed as an

exact Poisson 95% confidence interval (Garwood, 1936). Two sub-corpora are compared item by item using:

(i) the log-likelihood statistic G^2 for a 2 × 2 comparison of the item against all other tokens (Dunning, 1993; Rayson and Garside, 2000), referred to a χ^2 distribution with one degree of freedom;

(ii) Log Ratio, the binary logarithm of the ratio of the two normalised frequencies, as a measure of effect size, with a 95% confidence interval from its large-sample standard error (Hardie, 2014); a value of 1 means the item is twice as frequent in the first group, 2 four times as frequent, and 0 no difference. Where an observed frequency is zero, 0.5 is substituted so that the ratio is defined, and the corresponding interval is flagged as unstable; and

(iii) Holm's sequentially rejective procedure to control the family-wise error rate (Holm, 1979), applied separately to the seven national comparisons and to the fourteen gender comparisons (seven items × two countries). Reported p values are Holm-adjusted unless marked.

A difference is described as clear when its adjusted $p < .001$ and the Log Ratio interval excludes zero, and as tentative when adjusted $p < .05$. These thresholds are deliberately stringent because token-based tests overstate certainty when dispersion is unknown. Overall distributional differences across all seven items are additionally summarised with a χ^2 test of independence and Cramér's V (Brezina, 2018). Because document-level counts are unavailable, no dispersion measure, keyness analysis against a reference corpus, or test that treats tweets or anchors as the unit of analysis can be reported (see Section 6). All values were computed in Python (SciPy) from the reported counts and are rounded to two decimals.

4. Results

4.1 Distribution and size-adjusted profile of the selected items

Table 2 gives raw and normalised frequencies for the four sub-corpora. Raw counts are dominated by the larger Pakistani male sub-corpus (e.g. Pakistan, 550 occurrences), but normalisation shows the same rank order within countries: Pakistan is the most frequent item in both Pakistani sub-corpora (130.40 per 10,000 words among men; 125.08 among women), and India is the most frequent item in both Indian sub-corpora (102.72 among men; 138.89 among women). Across all seven items the pooled rate is 250.36 per 10,000 words for Pakistani male anchors (95% CI 235.49–265.93), 200.13 for Pakistani female anchors (169.58–234.60), 265.93 for Indian male anchors (241.16–292.55) and 230.35 for Indian female anchors (193.27–272.48).

Table 2. Raw frequency (normalised frequency per 10,000 words) of the selected items by sub-corpus.

Item	Pakistani male	Pakistani female	Indian male	Indian female

Media	74 (17.54)	17 (22.38)	24 (15.12)	17 (28.79)
Army	60 (14.23)	12 (15.80)	9 (5.67)	6 (10.16)
Corona	40 (9.48)	0 (0.00)	54 (34.03)	7 (11.86)
Pakistan	550 (130.40)	95 (125.08)	95 (59.87)	9 (15.24)
India	113 (26.79)	5 (6.58)	163 (102.72)	82 (138.89)
Modi	29 (6.88)	4 (5.27)	73 (46.00)	13 (22.02)
Imran Khan	190 (45.05)	19 (25.02)	4 (2.52)	2 (3.39)
All seven items	**1,056 (250.36)**	**152 (200.13)**	**422 (265.93)**	**136 (230.35)**

Note. Sub-corpus sizes: 42,179; 7,595; 15,869 and 5,904 words.

4.2 National comparison (RQ1)

Table 3 and Figures 1–2 compare the pooled national sub-corpora. Five items differ clearly. Imran Khan (Log Ratio = 3.93) and Pakistan (1.44) are more frequent among Pakistani anchors; Modi (−2.57), India (−2.25) and Corona (−1.80) are more frequent among Indian anchors. Army is more frequent among Pakistani anchors (1.07; roughly double the Indian rate), but only tentatively (adjusted p = .010), and its interval (0.27–1.87) is wide. Media shows no difference (−0.04; p = .875). The distribution of the seven items across the two national corpora differs substantially ($\chi^2(6) = 545.82$, $p < .001$, Cramér's V = 0.56).

Table 3. National comparison of the selected items: frequency (per 10,000 words), log-likelihood, Holm-adjusted p and Log Ratio.

Item	Pakistani n (per 10k)	Indian n (per 10k)	G^2	p (Holm)	Log Ratio [95% CI]
Media	91 (18.28)	41 (18.83)	0.02	.875	−0.04 [−0.57, 0.49]
Army	72 (14.47)	15 (6.89)	7.96	.010	1.07 [0.27, 1.87]
Corona	40 (8.04)	61 (28.02)	38.55	<.001	−1.80 [−2.38, −1.23]
Pakistan	645 (129.59)	104 (47.77)	112.04	<.001	1.44 [1.14, 1.74]
India	118 (23.71)	245 (112.52)	210.74	<.001	−2.25 [−2.56, −1.93]
Modi	33 (6.63)	86 (39.50)	88.06	<.001	−2.57 [−3.15, −2.00]
Imran Khan	209 (41.99)	6 (2.76)	111.17	<.001	3.93 [2.76, 5.10]
All seven items	**1,208 (242.70)**	**558 (256.28)**	**1.12**	**.289‡**	**−0.08 [−0.22, 0.06]**

Note. Log Ratio = $\log_2$(Pakistani rate ÷ Indian rate); positive values indicate higher Pakistani use. df = 1. ‡ Pooled-set test shown separately; unadjusted and not part of the Holm family.

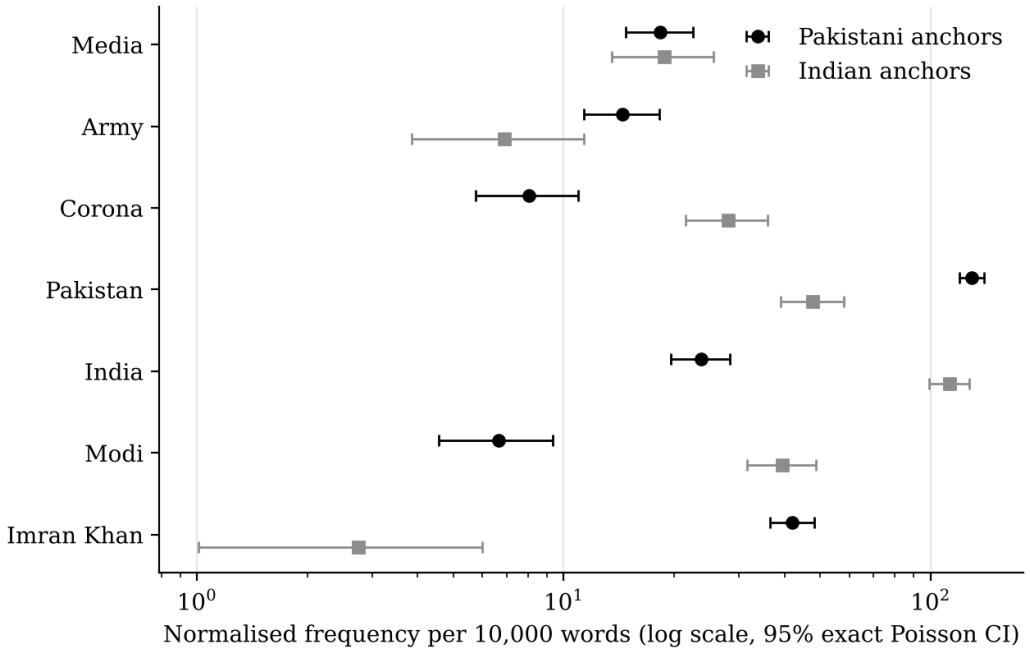

Figure 1. Normalised frequencies of the seven items among Pakistani and Indian anchors (log scale; bars show exact 95% Poisson confidence intervals).

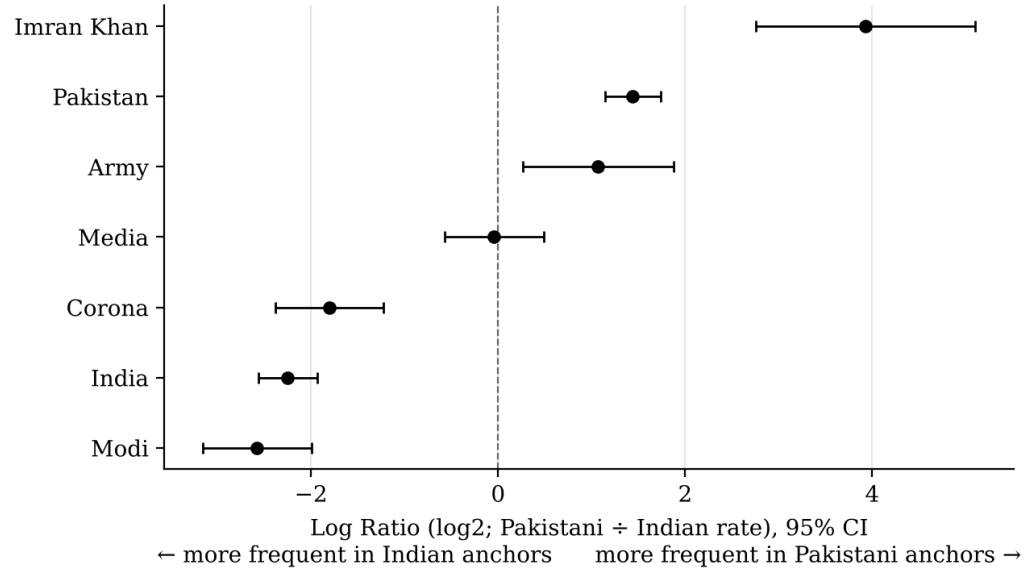

Figure 2. Log Ratio effect sizes with 95% confidence intervals. Items whose interval excludes zero differ between national groups.

4.3 Raw totals versus size-adjusted comparison (RQ2)

Raw totals suggest that Pakistani anchors use the selected items far more often (1,208 against 558 occurrences). However, the Pakistani corpus is 2.29 times larger than the Indian one (49,774 vs 21,773 words). Once adjusted, the pooled rates are 242.70 per 10,000 words (95% CI 229.20–256.78) for Pakistani and 256.28 (235.45–278.45) for Indian anchors. This difference is not statistically reliable ($G^2 = 1.12$, $p = .289$; Log Ratio = −0.08, 95% CI −0.22 to 0.06), so the appropriate conclusion is that the density of the selected vocabulary is comparable, not that either group uses it more. Normalisation therefore removes an apparent difference at the aggregate level while the item-level differences in Table 3 remain.

Normalisation also changes what can be said about the profile within each national corpus. In rank order, Pakistani anchors' most frequent items are Pakistan (129.59), Imran Khan (41.99) and India (23.71); Indian anchors' are India (112.52), Pakistan (47.77) and Modi (39.50). In the Indian corpus, therefore, the other nation is the second most frequent item, ahead of the prime minister; in the Pakistani corpus, the own-country leader is second, ahead of the other nation.

To examine this pattern more directly, four post hoc contrasts were run that pair each item with its counterpart in the other country (Table 4). Indian anchors refer to Pakistan about twice as often as Pakistani anchors refer to India (Log Ratio = −1.01, CI −1.39 to −0.63; adjusted $p < .001$). By contrast, the rates at which anchors name their own country's leader are almost identical (Imran Khan 41.99; Modi 39.50; Log Ratio = 0.09, CI −0.27 to 0.45), and the rate of naming their own country is similar (129.59 vs 112.52; adjusted $p = .114$). The own-to-other ratio is 5.5 for Pakistani anchors but 2.4 for Indian anchors, and own-country references account for 53.4% of selected-item occurrences in the Pakistani corpus against 43.9% in the Indian corpus. These contrasts were defined after inspecting the data and are exploratory.

Table 4. Exploratory own-country versus other-country contrasts (Pakistani anchors vs Indian anchors).

Contrast	Pakistani (per 10k)	Indian (per 10k)	G^2	p (Holm)	Log Ratio
Own country (Pakistan vs India)	129.59	112.52	3.61	.114	0.20
Other country (India vs Pakistan)	23.71	47.77	26.22	<.001	−1.01
Own leader (Imran Khan vs Modi)	41.99	39.50	0.23	.632	0.09

Other leader (Modi vs Imran Khan)	6.63	2.76	4.74	.089	1.27

Note. Holm correction over the four contrasts. Log Ratio positive = higher Pakistani rate.

4.4 Gender patterns within each country (RQ3)

Among Pakistani anchors (Table 5), Pakistan is used at near-identical rates by men and women (Log Ratio = 0.06). Men use India more often (26.79 vs 6.58; adjusted p = .002) and Corona more often (9.48 vs 0), though the latter rests on a zero count in the smaller female sub-corpus and its interval is unstable. Imran Khan is higher among men (45.05 vs 25.02), but not after correction (adjusted p = .079). Media and Army are somewhat higher among women, but these differences are small relative to their uncertainty. Pooled across all items, Pakistani men use the selected set at a higher rate than Pakistani women (250.36 vs 200.13; $G^2 = 7.04$, p = .008), and the item distributions differ ($\chi^2(6) = 22.42$, p = .001, V = 0.14).

Table 5. Gender comparison within the Pakistani sub-corpora: frequency (per 10,000 words), G^2, Holm-adjusted p and Log Ratio.

Item	Male n (per 10k)	Female n (per 10k)	G^2	p (Holm)	Log Ratio [95% CI]
Media	74 (17.54)	17 (22.38)	0.78	1.000	−0.35 [−1.11, 0.41]
Army	60 (14.23)	12 (15.80)	0.11	1.000	−0.15 [−1.04, 0.74]
Corona	40 (9.48)	0 (0.00)	13.25	.003	3.85 [−0.18, 7.87]†
Pakistan	550 (130.40)	95 (125.08)	0.14	1.000	0.06 [−0.25, 0.37]
India	113 (26.79)	5 (6.58)	14.82	.002	2.02 [0.73, 3.32]
Modi	29 (6.88)	4 (5.27)	0.27	1.000	0.38 [−1.12, 1.89]
Imran Khan	190 (45.05)	19 (25.02)	7.02	.079	0.85 [0.17, 1.53]

Note. Log Ratio = $\log_2$(male rate ÷ female rate). † Zero female count replaced by 0.5; interval unstable. Holm correction over the 14 gender comparisons in Tables 5 and 6.

Among Indian anchors (Table 6), the clearest gender difference concerns Pakistan, which is far more frequent among men (59.87 vs 15.24; Log Ratio = 1.97; adjusted p < .001). Corona is tentatively higher among men (34.03 vs 11.86; adjusted p = .030). Women show higher normalised rates for India (138.89 vs 102.72), Media (28.79 vs 15.12), Army (10.16 vs 5.67) and Imran Khan (3.39 vs 2.52), and men show a higher rate for Modi (46.00 vs 22.02), but none of these survives correction (all adjusted p ≥ .079), and the counts for Imran Khan (4 and 2) are too small to interpret. The overall rate of the selected set does not differ between Indian men and women (265.93 vs 230.35; p = .140), although the item distributions do ($\chi^2(6) = 43.19$, p < .001, V = 0.28; some expected counts fall below 5, so this test should be read cautiously).

Table 6. Gender comparison within the Indian sub-corpora: frequency (per 10,000 words), G^2, Holm-adjusted p and Log Ratio.

Item	Male n (per 10k)	Female n (per 10k)	G^2	p (Holm)	Log Ratio [95% CI]
Media	24 (15.12)	17 (28.79)	3.92	.335	−0.93 [−1.82, −0.03]
Army	9 (5.67)	6 (10.16)	1.16	1.000	−0.84 [−2.33, 0.65]
Corona	54 (34.03)	7 (11.86)	8.96	.030	1.52 [0.39, 2.66]
Pakistan	95 (59.87)	9 (15.24)	22.34	<.001	1.97 [0.99, 2.96]
India	163 (102.72)	82 (138.89)	4.79	.229	−0.44 [−0.82, −0.05]
Modi	73 (46.00)	13 (22.02)	7.06	.079	1.06 [0.21, 1.91]
Imran Khan	4 (2.52)	2 (3.39)	0.11	1.000	−0.43 [−2.87, 2.02]

Note. Log Ratio = $\log_2$(male rate ÷ female rate). Holm correction as in Table 5.

5. Discussion

The findings answer the three research questions in turn. For RQ1, the selected lexical profile is clearly not uniform across the two national groups: the strongest and most reliable contrasts concern the two nations, the two leaders and Corona, whereas Media is shared. For RQ2, size-adjustment changes the headline reading. What raw totals present as a large Pakistani excess is almost entirely a function of corpus size, and the pooled rates are statistically indistinguishable. The original reading that the Indian rate is "slightly higher" should therefore be softened to "comparable". Importantly, equal overall density conceals different composition, which is what the item-level analysis reveals.

The own-country/other-country contrasts add the most interpretively interesting pattern, which we offer as a hypothesis rather than an explanation. Indian anchors devote proportionally more of the selected vocabulary to the neighbouring country than Pakistani anchors do, while anchors in both countries name their own head of government at essentially the same rate. One possible reading is that the neighbouring state is a more frequent reference point in the Indian anchors' posts than the reverse, but frequency alone cannot show whether the references are critical, neutral or incidental, nor whether they arise from reporting events, quoting others or commenting. A word count records occurrence, not stance, sentiment or ideological function. Establishing these would require concordance and collocation analysis of the underlying tweets, which is the natural qualitative second stage of a corpus-assisted discourse study (Partington, 2004; Baker et al., 2008).

The higher normalised rate of Corona in the Indian material is consistent with the pandemic context of the period, but the data cannot show why it differs between nationalities. Differences in timing of tweeting, lexical variants (e.g. COVID, coronavirus; see Section 3.2), or the activity of particular accounts could each produce it.

The gender results are modest. After correction, only a few gender contrasts are reliable (India among Pakistani anchors; Pakistan among Indian anchors), and both involve an item whose reference is the other country, so they may reflect which stories individual anchors covered rather than gender as such (Baker, 2014). With five accounts per sub-corpus and unequal group sizes, a single prolific account could drive any of these contrasts. We therefore regard the gender findings as hypotheses for tweet-level and anchor-level follow-up.

Methodologically, the study illustrates three practices that strengthen small-corpus comparison even when only aggregate data exist: normalising to a common base, reporting effect sizes with intervals rather than significance alone, and controlling the family-wise error rate. It also shows what these practices do not do. Log-likelihood and Poisson intervals assume independent tokens; because tweets and anchors are clusters, the reported p values and intervals are likely narrower than the true uncertainty, which is why stringent thresholds and effect sizes were used to guide interpretation, and why the term keyword is avoided.

6. Limitations

First, the analysis rests on aggregate counts, not tweet-level data, so dispersion, tweet-level or anchor-level variance and keyness against a reference corpus cannot be assessed, and token-based tests are anti-conservative if words cluster within accounts. Second, only seven pre-selected items are examined; the results describe this selected vocabulary, not the lexical profile of either national discourse. Third, the sample is purposive and small (ten accounts), so the findings concern these accounts, and unequal sub-corpus sizes, especially the two female sub-corpora (7,595 and 5,904 words), widen intervals and limit gender inferences. Fourth, the Indian-male sub-corpus has shorter tweets on average (31.0 words per tweet), which may affect the composition of the vocabulary. Fifth, the post hoc own/other contrasts were not specified in advance. Sixth, details of collection, tokenisation and item definitions inherited from the source study (Section 3.2) determine what each count includes, and the findings depend on them. These constraints are stated to avoid overstating the evidence.

7. Conclusion

This article has provided a size-adjusted, effect-size-based comparison of seven lexical items in the Twitter posts of Indian and Pakistani news anchors from 2020 to 2021. National lexical emphases differ clearly for Pakistan, India, Imran Khan, Modi and Corona, but not for Media, and the overall density of the selected vocabulary is comparable across the two nations despite a larger Pakistani raw total. An exploratory analysis suggests that the main asymmetry lies in how often each group refers to the other country, not in how often it names its own leader. Gender differences are smaller and mostly not robust.

The contribution is methodological and descriptive: it shows how normalisation, interval estimation and multiplicity control change the reading of a small comparative corpus, and it specifies the limits of aggregate data. Future work using the complete tweet-level dataset could extend the analysis to the full vocabulary, test keyness against a reference corpus, incorporate dispersion and mixed-effects or resampling approaches that treat anchors and tweets as units, and add concordance and collocation analysis to interpret how the items are used.

Declarations

Data availability. This article uses aggregate frequency counts and corpus sizes reported in the author's MPhil thesis; all counts needed to reproduce every statistic are given in Tables 1 and 2. The tweet-level files were not available for this secondary analysis.

Ethics. The source study analysed publicly available Twitter posts. This research reproduces no tweet text and adds no new personal data.

Funding. No external funding is reported for this secondary analysis.

References

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

Baker, P. (2014) *Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T. and 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), pp. 273–306.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bondi, M. and Scott, M. (eds.) (2010) *Keyness in Texts*. Amsterdam: John Benjamins.

Boyd, D. and Crawford, K. (2012) 'Critical questions for big data', *Information, Communication & Society*, 15(5), pp. 662–679.

Brezina, V. (2018) *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, V. and Meyerhoff, M. (2014) 'Significant or random? A critical review of sociolinguistic generalisations based on large corpora', *International Journal of Corpus Linguistics*, 19(1), pp. 1–28.

Conrad, S. (2002) 'Corpus linguistic approaches for discourse analysis', *Annual Review of Applied Linguistics*, 22, pp. 75–95.

Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics*, 19(1), pp. 61–74.

Egbert, J. and Biber, D. (2019) 'Incorporating text dispersion into keyword analyses', *Corpora*, 14(1), pp. 77–104.

Gabrielatos, C. (2018) 'Keyness analysis: Nature, metrics and techniques', in Taylor, C. and Marchi, A. (eds.) *Corpus Approaches to Discourse: A Critical Review*. London: Routledge.

Garwood, F. (1936) 'Fiducial limits for the Poisson distribution', *Biometrika*, 28(3/4), pp. 437–442.

Gries, S.T. (2008) 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics*, 13(4), pp. 403–437.

Gries, S.T. (2025) 'Not just frequency: Keyness should integrate frequency, association, and dispersion', in Pawłowski, A., Embleton, S., Mačutek, J. and Xanthos, A. (eds.) *Mathematical Modelling in Linguistics and Text Analysis: Theory and Applications*. Amsterdam: John Benjamins, pp. 17–26. doi: 10.1075/cilt.370.02gri.

Hardie, A. (2014) 'Log ratio—an informal introduction'. Lancaster: ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University.

Hermida, A. (2013) '#Journalism: Reconfiguring journalism research about Twitter, one tweet at a time', *Digital Journalism*, 1(3), pp. 295–313.

Holm, S. (1979) 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics*, 6(2), pp. 65–70.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jeaco, S. (2020) 'Key words when text forms the unit of study: Sizing up the effects of different measures', *International Journal of Corpus Linguistics*, 25(2), pp. 125–155.

Kilgarriff, A. (2001) 'Comparing corpora', *International Journal of Corpus Linguistics*, 6(1), pp. 97–133.

Kilgarriff, A. (2005) 'Language is never, ever, ever random', *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 263–276.

Lasorsa, D.L., Lewis, S.C. and Holton, A.E. (2012) 'Normalizing Twitter: Journalism practice in an emerging communication space', *Journalism Studies*, 13(1), pp. 19–36.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. and Mannila, H. (2016) 'Significance testing of word frequencies in corpora', *Digital Scholarship in the Humanities*, 31(2), pp. 374–397.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics*. 2nd edn. Edinburgh: Edinburgh University Press.

Moreno-Ortiz, A. and García-Gámez, M. (2023) 'Strategies for the analysis of large social media corpora: Sampling and keyword extraction methods', *Corpus Pragmatics*, 7, pp. 241–265. doi: 10.1007/s41701-023-00143-0.

Oakes, M.P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Partington, A. (2004) 'Corpora and discourse, a most congruous beast', in Partington, A., Morley, J. and Haarman, L. (eds.) *Corpora and Discourse*. Bern: Peter Lang.

Rayson, P. and Garside, R. (2000) 'Comparing corpora using frequency profiling', in *Proceedings of the Workshop on Comparing Corpora*. Hong Kong: Association for Computational Linguistics, pp. 1–6.

Scott, M. (1997) 'PC analysis of key words—and key key words', *System*, 25(2), pp. 233–245.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Zappavigna, M. (2012) *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Continuum.