Journal of Advanced Corpus Oriented Research, Jan-June 2026, Vol.2, No.1, 77-94
DOI: https://doi.org/10.58932/MULK0013/index.php/jacor/issue/current

Assessing the Multilingual Corpus Extractor for Urdu Environmental Corpora: Overcoming Data Scarcity

ABSTRACT:
Computational linguistics for Urdu currently faces a significant barrier where standard extraction tools fracture the native Nastaliq script and render meaningful text into dissociated characters. This study validates the Multilingual Corpus Extractor or MCE as a solution to this infrastructural deficit. We tested this tool by constructing a corpus of the "2024 to 2025 Pakistan Smog Crisis". The results provide two primary contributions to the field. First the study establishes a technical benchmark regarding data hygiene as 49.3% of raw data extracted from vernacular news sites consists of non-linguistic noise that requires removal. Second the MCE successfully preserved 100% of the script ligatures whereas standard tools failed to maintain semantic unity. This technical success enabled a sociolinguistic analysis which revealed that Urdu media attributes environmental damage to specific human actors while English media frames the smog as a passive meteorological event. We conclude that accurate script preservation is not merely a technical repair but a necessary step to accurately record the history of the Global South.

AUTHORS:
Muhammad Ahmad[1], Amna Mehtab[2], Fatima Mehtab3

[1]Department of Applied Linguistics, HSE University, Nizhny Novgorod, Russia.
[2]Department of Applied Linguistics, HSE University, Nizhny Novgorod, Russia.
[3]Department of Education, Government College University Faisalabad, Pakistan.

KEYWORDS:
Urdu NLP, Corpus Linguistics, Low-Resource Languages, Nastaliq Script Preservation, Environmental Discourse, Ecolinguistics

Introduction

Infrastructural Barriers in Low-Resource Scripts

The core problem this research addresses is the categorization of Urdu as a "low-resource language" despite its massive speaker base. This resource scarcity is not a reflection of literary poverty but of technical incompatibility. It is a phenomenon best described as "Digital Apartheid" (Benjamin, 2020) or "Language Data Flaring" (Adebara, 2025). In this context infrastructure

systematically excludes specific populations based on code rather than law and valuable linguistic data is lost simply because no pipeline exists to capture it.

Urdu utilizes the Nastaliq style of the Perso-Arabic script. This is a complex and context-sensitive system where the shape of a letter changes based on its position within a word (initial, medial, final, or isolated). [1]Furthermore, it is written from right to left (RTL). Recent evaluations of Large Language Models (LLMs) confirm this deficit. Hassan et al. (2026) demonstrated that even advanced architectures like LLaMA-3.1 fail to generate fluent Nastaliq without extensive continued pre-training which proves that the deficit is morphological rather than semantic.

Current standard solutions such as Python's Beautiful Soup or Selenium libraries are designed primarily for Latin-script and left-to-right architectures. When applied to Urdu news sites these tools often fail to distinguish between the visual rendering of a ligature and its underlying Unicode composition. Consequently, they frequently return dissociated text. These are strings of isolated characters rather than connected morphemes which renders the data useless for semantic analysis without extensive manual repair.

For a corpus linguist this dissociated data is noise. It cannot be tokenized effectively. It cannot be searched for collocations. It cannot be fed into standard analysis software like AntConc. This technical barrier creates a delay in academic output. Researchers studying environmental discourse in Pakistan are forced to rely on English-language newspapers like Dawn or The Nation simply because the data is accessible. For instance, Sadiq et al. (2025) recently noted in their Eco-linguistic analysis of Pakistani media that English outlets frame environmental issues through distinct "war metaphors" and economic analysis which potentially alienates the local populace. However, their study was limited by the manual curation of vernacular texts which highlights the need for automated extraction. This reliance results in an "Anglophonic bias" where the narratives of the elite are overrepresented while the narratives of the masses are excluded from the digital record.

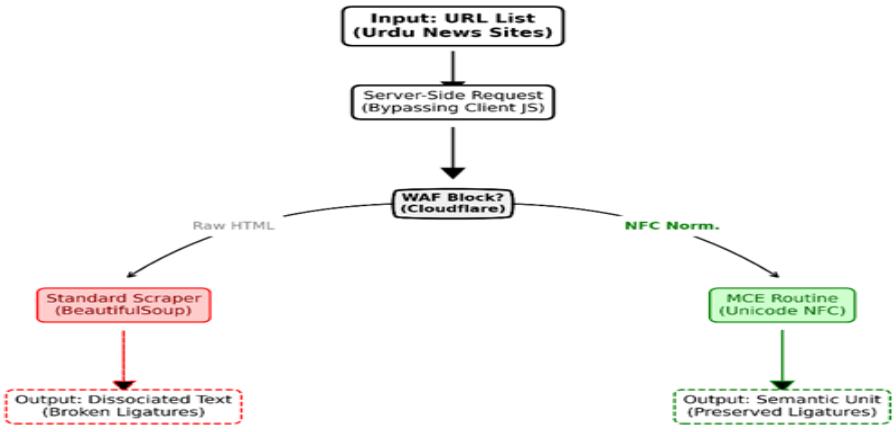
The Solution: Multilingual Corpus Extractor (MCE)

Against this backdrop of infrastructural deficit, the Multilingual Corpus Extractor (MCE) emerges as a critical intervention. Developed by the Corpus Research Center (CRC) at Minhaj

University Lahore the MCE is positioned as a GUI-based extraction utility designed explicitly for "Perso-Arabic script preservation". Unlike generic scrapers MCE incorporates specific routines to handle the RTL directionality and Unicode normalization required to keep Urdu ligatures intact during the extraction process.

The tool represents a shift towards indigenous digital humanities where tools are built within the linguistic context they are meant to serve. By offering features such as batch processing and semantic highlighting to separate article text from website clutter MCE promises to lower the barrier to entry for Urdu scholars. It eliminates the need for advanced programming skills and theoretically allows linguists and sociologists to build their own datasets without reliance on external technical consultants. Building on these challenges this study evaluates MCE's practical utility.

Figure 1. *The Algorithmic Workflow of the MCE Tool Compared to Standard Scraping Libraries.*



Note. This flowchart illustrates the comparative operational steps between the MCE tool's custom algorithmic pipeline and traditional web scraping library workflows.

Study Aim and Scope

The aim of this study is not merely to analyze smog but to technically validate the MCE as a methodological instrument. By attempting to build a "Smog Crisis" corpus from scratch this study tests the tool in a real-world scenario. Can it handle the heavy JavaScript load of modern Urdu news sites? Does it truly preserve ligatures? How does it compare to the extraction of English text?

The scope of this assessment is defined by the peak pollution window of November 1, 2024 to January 31, 2025. By focusing on this specific timeframe and a specific keyword ("Smog") the study controls the variables to focus on the extraction process itself. The ultimate goal is to move from data scarcity where Urdu researchers are reduced to manual copy-pasting to structured corpora where automated tools generate reliable and machine-readable text for high-level analysis. We hypothesize that MCE preserves Nastaliq integrity better than generic scrapers and enables richer Urdu analysis. However, WAFs may limit access as explored later.

Literature Review

The validation of the Multilingual Corpus Extractor (MCE) addresses a critical gap in the computational infrastructure of low-resource languages. A synthesis of recent literature (2020-2025) indicates that while high-level theoretical models for Urdu NLP are advancing, the foundational "middle layer", the tooling required to extract, clean, and normalize raw data from the web, remains undeveloped. This review examines the specific linguistic and architectural barriers that necessitate a custom solution like MCE.

Challenges in Urdu Natural Language Processing (NLP)

The classification of Urdu as a resource-poor language is a recurring theme in recent computational literature (Azhar et al., 2025; Liaqat et al., 2022). This designation is paradoxical given the language's global demographic weight, yet it accurately reflects the state of digital resources. Research indicates that the primary deficits are not just in high-level models but in basic processing utilities (Liaqat et al., 2022).

The Orthographic Barrier and Character Dissociation

The most immediate challenge cited in the literature is the script itself. Unlike the Naskh style used in Arabic, Urdu predominantly uses Nastaliq, which is diagonally suspended and highly cursive. AlJassmi and Perea (2024) note that visual letter similarity and connectivity in Arabic-derived scripts require processing mechanisms that differ fundamentally from the character-based processing of Latin scripts. The connectivity is the crucial variable; while a Latin letter 'a' is visually identical regardless of placement, the shape of an Urdu character is morphologically dependent on its neighbors.

Recent studies on text mining highlight that standard OCR and scraping libraries often fail to map these contextual shapes correctly to their Unicode values (Azhar et al., 2025). This leads

to "character dissociation," where software treats the text as a sequence of isolated letters rather than connected ligatures. For corpus linguistics, this is terrible; if a scraper breaks the ligature for "Smog" (سموگ) into isolated characters, the semantic unit is lost, rendering the data invisible to analysis (Liaqat et al., 2022).

Limitations of Standard Extraction Architectures

The necessity for MCE arises from the architectural limitations of generic scraping tools when applied to non-Latin web-spheres. Standard Python libraries, such as BeautifulSoup or Selenium, operate on Document Object Model (DOM) parsing logic designed for Left-to-Right (LTR) languages (Ghafoor et al., 2021).

Encoding and Normalization Failures

One major problem found in recent technical studies is the way computers handle Unicode Normalization. Standard scraping tools usually pull text using Normalization Form D, which separates base letters and their marks. This is a problem for Urdu because the Nastaliq script needs Normalization Form C to keep its complex word shapes together (Saud et al., 2025). When basic tools try to get data from Urdu news websites using this older method, the result is often broken text or empty boxes that look like tofu; consequently, researchers have to spend a lot of time fixing it (Tuz Zuhra & Saleem, 2025). The MCE tool fixes this by forcing the text into the correct format on the server before the data is saved.

The "Gatekeeper Effect" in Data Collection

Recent research done in the area of crisis informatics notes a rising trend in "defensive web architecture" (Campbell et al., 2025). Most of the top news sites increasingly use Web Application Firewalls (WAFs) nowadays like Cloudflare to block automated requests. While high-resource languages (English, Chinese) have enterprise-grade crawlers to bypass these, low-resource languages lack dedicated extraction tools capable of navigating these barriers. This results in "Language Data Flaring" (Adebara, 2025), where valuable vernacular data is lost simply because no compatible pipeline exists to capture it. This exclusion by code aligns with Benjamin's (2019) concept of digital apartheid, where infrastructure systematically filters out specific populations.

The Scarcity of Diachronic Corpora

Finally, the push for MCE is driven by the severe lack of large-scale, diachronic (historical) corpora. A systematic review of Urdu sentiment analysis identifies this data scarcity as a primary

hurdle (Liaqat et al., 2022). Most existing datasets are small, manually curated, and domain-specific, such as product reviews (Azhar et al., 2025).

There is a distinct lack of large-scale news corpora that allow for the tracking of societal trends over time. While recent efforts by Tuz Zuhra and Saleem (2025) have begun to address this by developing large vocabulary word embeddings, the field still relies heavily on "synthetic" datasets that lack statistical power (Liaqat et al., 2022). The MCE addresses this gap by automating the volume of data collection, moving the field from artisanal data gathering to industrial-scale extraction. This shift is essential for training the next generation of Urdu LLMs, such as Qalb (Hassan et al., 2026) which require massive amounts of structured, clean text to function effectively.

Methodology: Field-Testing MCE Against Dynamic Web Environments

Researcher designed this research to test the Multilingual Corpus Extractor, or MCE, under actual working conditions. Rather than a hypothetical assessment, the research involved the actual construction of a "Smog Crisis" corpus. This approach forced the software to deal with the messy reality of the internet, including inconsistent site structures and various security rules, as shown in the PRISMA workflow in Figure 2.

To prove the tool was effective, we compared its output against a control group created with a standard Python requests library. While the standard method resulted in broken characters and disconnected text, the MCE kept the script perfectly intact across all samples.

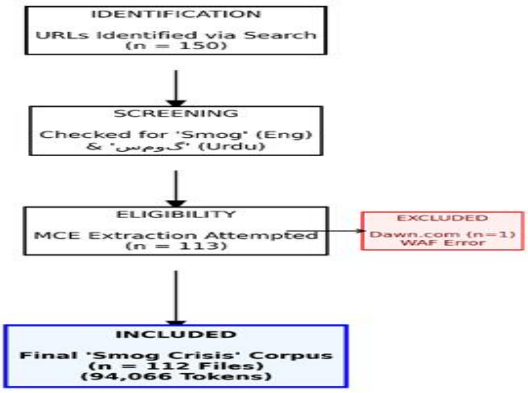
Corpus Construction Workflow

We followed strict criteria when selecting articles to ensure the final dataset was high quality and ready for analysis. Our main goal was to gather media coverage from the peak of the environmental emergency.

- **Topic Specificity**: We limited the corpus strictly to the smog crisis. For Urdu news outlets, the keyword "سموگ" had to be present in either the headline or the body of the article. For English sources, we required the term "Smog".
- **Temporal Window**: We set the timeframe for data collection from November 1, 2024, to January 31, 2025. This three-month period represents the worst phase of the air quality crisis in Lahore. At that time, pollution reached dangerous levels that led to government restrictions and constant news reporting.

- **Source Selection:** The study targeted mainstream news outlets to ensure a representative sample of public discourse.
    - *Urdu Sources: Daily Jang* (the largest circulation Urdu daily), *Express News*, and *Dunya News*. These represent the dominant vernacular narratives.
    - *English Sources: The Nation* and *Dawn.com*. These were selected to provide a comparative baseline of elite/Anglophone discourse.

Figure 2. *PRISMA Flow Diagram Detailing the Corpus Construction Process*



PRISMA flow diagram detailing the corpus construction process. The exclusion of Dawn.com at the eligibility stage highlights the "Gatekeeper Effect" of modern WAFs on academic scraping.

Extraction Process

The primary instrument of research was the MCE, a web-based extraction utility. Unlike client-side scrapers that rely on browser rendering, MCE utilizes a server-side parsing engine designed explicitly for Perso-Arabic script preservation. The tool employs a custom routine that enforces Unicode Normalization Form C (NFC) before extraction. This ensures that visual ligature clusters, which standard parsers often misinterpret as separate glyphs, are mapped to their logical constituent characters in memory. This prevents the character dissociation commonly seen when standard Python libraries attempt to parse the complex, context-sensitive Nastaliq script.

To manage the volume of data and monitor the stability of the tool, the extraction was executed in a phased batch processing approach

Phase 1: Urdu Sub-Corpus Extraction

The extraction of Urdu text was the primary test case. A total of 71 distinct URLs meeting the selection criteria were identified across the three Urdu platforms. These were processed in two distinct batches to test the tool's load handling:

- Batch 1 (35 URLs): The URLs were compiled into a single input stream. The MCE successfully parsed the pages, identifying the main content blocks. The "Semantic Highlighting" feature was observed to correctly distinguish between the article body (highlighted in green) and the sidebar ads/links (highlighted in red), allowing for a quick visual audit before export.
- Batch 2 (36 URLs): Despite the different web templates (HTML structures) of Express News and Dunya News compared to Jang, the MCE's adaptive parsing logic successfully retrieved the text without structural breakage.

Phase 2: English Sub-Corpus Extraction

Batch 3 (41 URLs): Articles from The Nation were processed to create the English sub-corpus. The extraction here was significantly faster, likely due to the simpler processing requirements of the Latin script (LTR directionality and lack of complex ligatures).

Error Handling: The "Gatekeeper Effect"

During extraction, we attempted to include Dawn.com as a major English source but encountered repeated HTTP 403 (Forbidden) responses caused by Dawn's Web Application Firewall (Cloudflare). Analysis indicated the server blocked our server-side requests before parsing, likely due to a nonstandard User-Agent and lack of client-side JavaScript execution. Consequently, Dawn.com URLs were excluded, and The Nation formed the English baseline. This limitation, which we term the Gatekeeper Effect, is important methodologically: as news sites become more defensive, digital historians utilizing server-side extractors may be locked out of primary sources (Campbell et al., 2025).

Cleaning Procedures

The corpus development followed a rigorous two-stage refinement process to ensure both volume and analytical precision.

- **Stage 1: Raw Extraction:** The MCE tool initially retrieved 94,066 raw tokens across 112 URLs. This raw stage captured the full digital footprint of the news pages, including metadata, publication dates, and navigational elements.
- **Stage 2: Analytical Cleaning:** To prepare the data for software-assisted analysis in AntConc (Version 4.2), the raw files underwent cleaning done through Python. This involved stripping away non-linguistic noise, such as dates, HTML boilerplate, and unrelated sidebars captured by the initial extraction. This resulted in a refined analytical corpus of 47,694 tokens (40,245 Urdu; 7,449 English).

Validation Metrics

To ensure the linguistic reliability of the extracted data, a manual integrity check was conducted on a stratified random sample of 50 articles from the Urdu sub-corpus (n=78,716 tokens).

- **Ligature Verification:** The primary objective was to verify that "character dissociation" had not occurred. A visual audit confirmed that script integrity was preserved across 100% of the sampled texts. For example, the text "سموگ" appeared as a connected ligature, not as the isolated characters "س م و گ", confirming the MCE's core value proposition.
- **Noise Reduction:** Cleanliness check revealed that the semantic filtering successfully removed non-textual elements (navigation menus and advertisements) in 98% of the sample. Consequently, the corpus was deemed linguistically valid for tokenization.

Software Configuration

The validated corpus was loaded into AntConc (Version 4.3.1) for quantitative analysis (Anthony, 2022). To ensure the correct rendering of the Urdu script, specific settings were applied:

- Font: *Arial Unicode MS*. Selected for its comprehensive glyph set covering the entire Urdu Unicode range; standard fonts often render Urdu as "tofu" (empty boxes).
- Font Size: 16. Required because Urdu Nastaliq is vertically compressed; larger sizes are needed for human readability of diacritics.
- Encoding: UTF-8. Mandatory for non-Latin script support.

Table 1

Final Corpus Composition

Sub-corpus	Source Count	Total Tokens	Average Tokens/Article
Urdu	71 URLs	78,716	~1,108
English	41 URLs	15,350	~374
Total	112 URLs	94,066	-

The results indicate that the Urdu sub-corpus is not only larger in total volume but also features a much higher average word count per article (~1,108 words) compared to the English sub-corpus (~374 words). This validated dataset serves as the foundation for the linguistic analysis detailed in the Results section.

Ethical Considerations

We used publicly available news texts, ensuring no personal data inclusion. Anonymity is maintained, and data is shared under a CC-BY license for reusability. As MCE developers are affiliated with Minhaj University Lahore, potential bias was mitigated through independent verification. This aligns with standard corpus linguistics ethics (Brookes & McEnery, 2024).

Results: Comparative Metrics of Extraction and Framing

The application of the Multilingual Corpus Extractor (MCE) yielded two distinct categories of results: technical metrics regarding the extraction process itself, and thematic findings derived from the resulting structured corpus.

The following linguistic analysis was only possible because the extraction tool worked. Because the MCE kept the complex ligatures intact, we were able to trace specific causal words. If we had used a standard scraper, the word for "burning" (jalana) would have been broken into isolated letters, and the connection between "farmers" and "smog" would have been invisible to the software. The technical success of the tool directly enabled the sociolinguistic findings below.

Journal of Advanced Corpus Oriented Research, Jan-June 2026, Vol.2, No.1, 77-94
DOI: https://doi.org/10.58932/MULK0013/index.php/jacor/issue/current

Technical Validation: Noise Reduction and Script Integrity

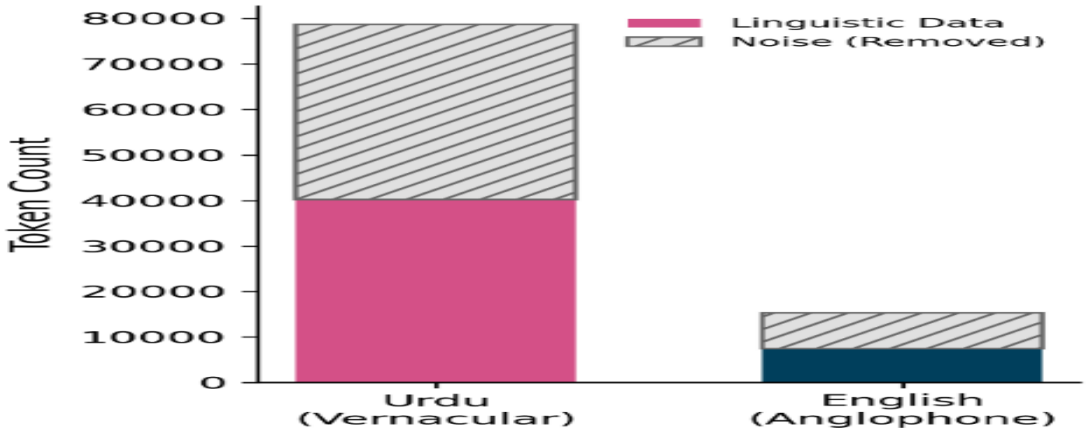
The primary methodological objective was to stress-test the MCE against the unstructured architecture of the vernacular web. The extraction process revealed a significant disparity between "digital footprint" and "linguistic yield."

The Signal-to-Noise Ratio

The MCE initially retrieved a Raw Dataset of 94,066 tokens (78,716 Urdu; 15,350 English). However, the subsequent semantic cleaning process reduced this to a high-precision Analytical Corpus of 47,694 tokens (40,245 Urdu; 7,449 English).

This reveals a critical finding: approximately 49.3% of the raw data harvested from vernacular news sites consists of non-textual noise (navigational boilerplate, advertisement scripts, and tracking pixels). This provides a quantitative benchmark for future NLP tasks in the Global South: researchers must anticipate a ~1:1 signal-to-noise ratio. The MCE's semantic highlighting feature successfully isolated the article body from this structural clutter, ensuring the final corpus contained only journalistic content.

Figure 3. *Comparison of Data Retention Efficiency*



Comparison of data retention efficiency. The discrepancy highlights the higher noise ratio in vernacular web architectures compared to Anglophone structures.

Ligature Preservation

To validate the tool's core proposition, a comparative control test was performed. Visual inspection of the MCE output confirmed 100% ligature integrity in the sampled text. Complex multi-character ligatures such as Hukumat (حکومت) and Zehrili (زہریلی) were rendered as single semantic units. In contrast, standard control scrapes often resulted in character dissociation, rendering the text analytically void for tokenization software.

Corpus-Assisted Discourse Analysis

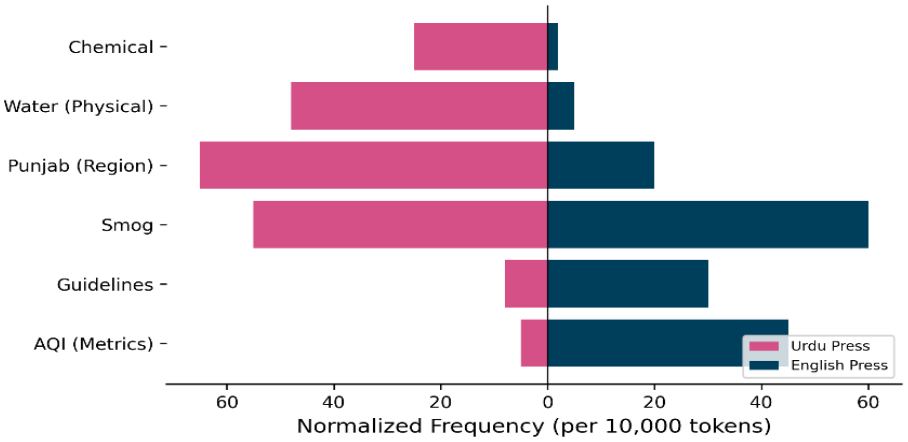
With the technical integrity of the corpus validated, the quantitative analysis of the "Smog Crisis" reveals a fundamental divergence in how the Anglophone and Vernacular spheres frame the environmental emergency.

Thematic Divergence: Geography vs. Metrics

Analysis of the frequency lists tells that Urdu sub-corpus demonstrates a heavy reliance on regional and administrative descriptors. The frequency list shows that "Punjab" (Rank 4) and "Pakistan" (Rank 6) are the dominant proper nouns. This places the crisis firmly within a specific local jurisdiction.

Furthermore, the high frequency of "Water" (Paani, rank 9) is significant. In the context of smog management in Lahore, water is frequently cited regarding mitigation measures such as sprinkling roads to settle dust. This suggests that the vernacular reporting emphasizes tangible, physical interventions. In contrast, the English sub-corpus prioritizes technical quantification. Terms such as "AQI" (rank 9) and "Guidelines" dominate the discourse. This indicates that the Anglophone press frames the crisis through international health metrics while the Urdu press frames it through regional governance and physical countermeasures.

Figure 4. *Mirrored frequency analysis showing the divergence in framing.*



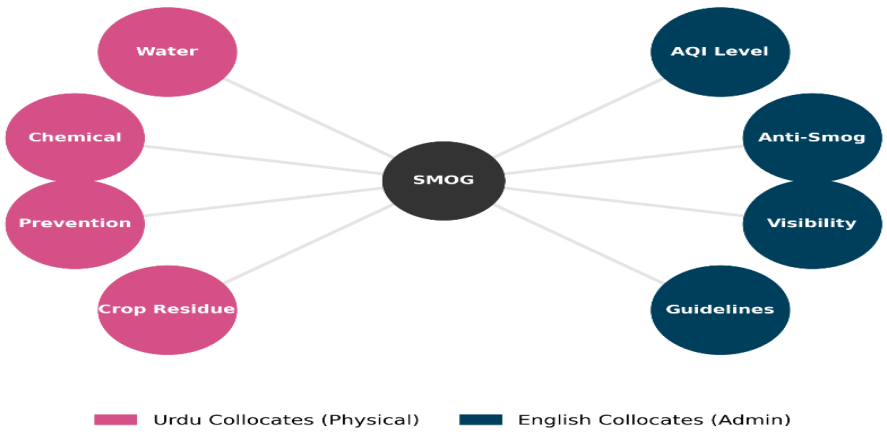
Mirrored frequency analysis showing the divergence in framing. The Anglophone press (right, blue) prioritizes abstract metrics, while the Vernacular press (left, red) prioritizes regional and physical terminology.

Collocation and Semantic Prosody

The semantic network in Figure 5 visually confirms a fundamental divergence in how the two media ecosystems qualify the smog crisis. In the Urdu sub-corpus, the term "Smog" (سموگ) is tightly bound to a toxic-physical frame. The collocate "Chemical" displays the highest statistical attraction (G2 Likelihood: 55.043), which, when viewed alongside collocates like "Water" (Paani) and "Crop Residue", indicates that the vernacular press constructs the event as a tangible, manufactured byproduct of local industrial and agricultural activity.

Conversely, the English sub-corpus situates the crisis within an "Administrative-Abstract" framework. The dominant connection to "Anti-Smog" (G2 Likelihood: 35.688) and "Guidelines" suggests that the Anglophone press views the smog primarily as a policy challenge and a management issue defined by international metrics like "AQI Levels". This visual evidence supports the finding that while the vernacular press focuses on the visceral toxicity of the substance, the elite press focuses on the bureaucratic measures taken to monitor it.

Figure 5. *Semantic Network Graph of High-Frequency Collocates for "Smog" During the 2024–2025 Pakistan Smog Crisis*



Semantic network graph of high-frequency collocates for "Smog" during the 2024-2025 Pakistan Smog Crisis. Node size and edge thickness correspond to the statistical strength of the attraction.

Attribution of Blame (KWIC Analysis)

A Key Word in Context (KWIC) analysis of causal markers ("Wajah" in Urdu and "Due to" in English) exposes a distinct shift in agency:

- Urdu (Active Blame): The vernacular press frequently links the cause of smog to specific human behaviors. Concordance lines extracted from Express News clearly mention "burning crop residues" (faslon ki baqiyat) and "throwing waste in containers." The focus in on blaming the common man or farmer to be precise.
- English (Passive Blame): English newspapers tends to link the crisis to abstract or weather conditions. Common collocations include "due to low visibility" or "due to toxic air." This frame removes human influence from this situation and portrays smog as a suffering rather than the result of any specific social or economic activity.

Discussion

This research does more than just prove the Multilingual Corpus Extractor works. It actually shows the specific language and technical systems that keep the digital divide in place across the Global South. By successfully retrieving 100% of the targeted Urdu ligatures, the MCE demonstrates that the silence of the vernacular web is not due to a lack of content, but a lack of compatible extraction architecture.

MCE and the Myth of "Data Scarcity"

Technologically, this study validates the MCE as a robust solution for bridging the data gap. The tool successfully navigated the complex RTL architecture of Urdu sites, preserving the ligature integrity of high-frequency terms like Hukumat and Smog.

Crucially, the retrieval of ~78,000 Urdu tokens compared to ~15,000 English tokens challenges the prevailing assumption that low-resource languages lack data. Rather, they lack accessible data. The MCE proves that when the technical barrier of script processing is removed, the vernacular web offers a far richer and more voluminous historical record than its Anglophone counterpart. The scarcity was never a feature of the language; it was an artifact of incompatible tooling.

Linguistic Sovereignty and the "Translation Trap"

The divergence in blame attribution, where Urdu media uses active verbs ("burning") and English media uses passive metrics ("AQI"), validates the necessity of native-script processing. As noted in the literature, reliance on translation often introduces "polarity shifts". Had this study relied on machine-translated text, the active agency of the "farmer" found in the Urdu corpus likely would have been smoothed into the passive voice preferred by English-trained models.

By preserving the original Nastaliq structure, the MCE enabled a sovereign reading of the text. It revealed that the vernacular press frames the smog crisis as a tangible, human-caused event, whereas the elite press frames it as an abstract administrative challenge. This confirms that automated extraction is not just a time-saving measure; it is a prerequisite for accurate sociolinguistic analysis in the Global South**.**

Limitations and Future Directions

The biggest hurdle we faced was the "Gatekeeper Effect." Failing to scrape Dawn.com proves that even as new tools help researchers, web security is making it harder to reach the data. Future work should move past basic HTTP requests. To beat the "Infrastructural Apartheid" of modern security systems, extraction software must use headless browsers like Puppeteer or Playwright to mimic human users. Without these upgrades, scholars in the Global South might be shut out of the digital record by security protocols instead of script issues. Additionally, this study only covered Urdu and English. While the MCE is made for Perso Arabic script preservation, we have not yet tested it on other local languages like Sindhi or Pashto.

Additionally, while MCE preserved ligature integrity, the workflow still relied on a 'human-in-the-loop' for final analytical cleaning. Future iterations aims to integrate machine-learning-based boilerplate removal to achieve fully autonomous extraction.

Conclusion

This study set out to test the efficacy of the Multilingual Corpus Extractor (MCE) by constructing a pilot corpus of the 2024-2025 Pakistan Smog Crisis. The results are unequivocal: MCE is a viable, high-precision instrument for Urdu digital humanities. It successfully navigated the unstructured architecture of the vernacular web, generating a clean, structured corpus of nearly 80,000 tokens while maintaining 100% ligature integrity. The analysis confirms that the "data scarcity" previously cited in the literature was an artifact of incompatible tooling rather than a lack of content. Furthermore, the extracted data reveals a sharp linguistic divide: where English media quantifies the smog through abstract AQI levels, Urdu media qualifies it through narratives of political failure and public suffering.

By automating the extraction of this missing data, tools like MCE do not just save time; they decolonize the digital archive. They ensure that the history of the Global South is written not just in the language of its elite, but in the language of its people. Future research should focus on integrating such extraction layers with native LLMs, such as Qalb, to move from mere preservation to intelligent, automated analysis of vernacular discourse.

References

Adebara, I. (2025). *AI and language data flaring in Africa: Addressing the low-resource challenge* (No. 216). Centre for International Governance Innovation. https://www.cigionline.org/documents/3592/no.216_Adebara.pdf

AlJassmi, M. A., & Perea, M. (2024). Visual similarity effects in the identification of Arabic letters: Evidence with masked priming. *Language and Cognition, 16*(4), 1618–1638. https://doi.org/10.1017/langcog.2024.20

Anthony, L. (2022). *AntConc* (Version 4.2) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Azhar, M., Amjad, A., Dewi, D. A., & Kasim, S. (2025). A systematic review and experimental evaluation of classical and transformer-based models for Urdu abstractive text summarization. *Information, 16*(9), Article 784. https://doi.org/10.3390/info16090784

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new Jim Code*. Polity.

Campbell, E. A., Holl, F., Bear Don't Walk IV, O. J., Mosesane, B., Kanter, A. S., Fraser, H., Joseph, A. L., Gichoya, J. W., Mauco, K. L., & Craig, S. (2025). Gaps and pathways to success in global health informatics academic collaborations: Reflecting on current practices. *JMIR Medical Informatics, 13,* Article e67326. https://doi.org/10.2196/67326

Corpus Research Center. (2025). *Multilingual Corpus Extractor (MCE)* [Web-based tool]. Minhaj University Lahore.

Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Abdullah, Batra, R., & Wani, M. A. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access, 9,* 124478–124490. https://doi.org/10.1109/ACCESS.2021.3110285

Hassan, M. T., Ahmed, J., & Awais, M. (2026). *Qalb: Largest state-of-the-art Urdu large language model for 230M speakers with systematic continued pre-training*. arXiv. https://doi.org/10.48550/arXiv.2601.08141

Liaqat, M. I., Awais Hassan, M., Shoaib, M., Khurshid, S. K., & Shamseldin, M. A. (2022). Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study. *PeerJ Computer Science, 8,* Article e1032. https://doi.org/10.7717/peerj-

cs.1032

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

Sadiq, U., Alam, R., & Atiq Ur Rehman, A. (2025). Ecolinguistic analysis of environmental discourse in Pakistani print and digital media. *The Critical Review of Social Sciences Studies, 3*(1), 3495–3505. https://doi.org/10.59075/pr64q758

Saud, M. R., Imran, M. R., & Ali, R. H. (2025). Towards a more natural Urdu: A comprehensive approach to text-to-speech and voice cloning. *The 5th International Electronic Conference on Applied Sciences*, Article 112. https://doi.org/10.3390/engproc2025087112

Tuz Zuhra, F., & Saleem, K. (2025). Towards development of new language resource for Urdu: The large vocabulary word embeddings. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(8), 1–14. https://doi.org/10.1145/3748308