SEMANTIC ANNOTATION OF SELECTED URDUIZED WORDS IN PAKISTANI ENGLISH: EVIDENCE FROM PAKLOCCORPUS

Fatima Tuz Zahra
PhD Scholar, Air University Islamabad
Lecturer, Minhaj University Lahore
Dr. Tehseen Zahra
Associate professor
Bahria University Islamabad

Abstract

Urduized words embedded in English discourse are a defining feature of Pakistani English. Despite their frequency and sociocultural salience, these words remain under-represented in corpus annotation frameworks and natural language processing (NLP) systems. This paper presents a corpus-driven semantic annotation framework for Urduized words, aligned with data from PakLocCorpus (2022). Using concordance evidence from the PakLocCorpus Urduized word list, the study conduct a discourse analysis of selected words using Gee (2011) and develops a context-sensitive semantic tag-set that captures culturally grounded meaning domains of descriptive labels. The analysis demonstrates that Urduized words function as semantically dense cultural carriers rather than peripheral borrowings. The paper argues that systematic semantic annotation of Urduized words is essential for inclusive corpus linguistics and for reducing structural bias in English-focused NLP technologies and descriptive labels have unique morpho-syntactic features that can be used for designing tagging tools and frameworks. Findings of the current study can be used in building pedagogical strategies for English language teaching (ELT) and second language acquisition (SLA).

Keywords: Urduized words, Pakistani English, semantic annotation, PakLocCorpus, corpus linguistics

1. Introduction

Pakistani English has been widely recognized as an institutionalized variety shaped by sustained contact with Urdu and other local languages. One of its most salient lexical features is the presence of Urduized words occur naturally in English texts without translation or typographic marking. Words such as *chachi, beta, dadi, chacha*, and *sahib* are routinely used across genres including journalism, academic writing, fiction, and digital media. Seminal works of Kachru (1983), Baumgardener et al. (1993) have provided characteristic features of South Asian English (SAE) and Pakistani English. Contextual areas and semantic features of single or hybrid items have provided a foundation for the current investigation. Such traces are also found in Indian English due to the similarities between both varieties and their socio-cultural backgrounds. While Kachru focuses on the hybrid formations, Baumgardner et al. analysed the transfer of single items or Urduized words.

The findings align with a growing body of corpus-based research demonstrating that Urduized words are systematic, socially motivated features of Pakistani English rather than incidental instances of code-mixing (Zahra, Zahra, & Abbas, 2023a; Zahra, Zahra, & Abbas, 2023b).While previous research has documented lexical borrowing and code-mixing in Pakistani English, limited attention has been paid to the semantic annotation of Urduized words using corpus evidence. Existing semantic frameworks, largely developed for Inner-Circle Englishes, often fail to capture the culturally embedded meanings of these lexical items. As a result, Urduized words are frequently misclassified or excluded in corpus annotation and NLP pipelines.

This study addresses this gap by proposing a semantic annotation framework grounded in PakLocCorpus data. By aligning annotation categories with authentic corpus usage, the paper

contributes to both the linguistic description of Pakistani English and the development of more equitable language technologies.

This study seeks to answer following research questions:

1.What semantic domains do Urduized words in Pakistani English texts represent, as evidenced by PakLocCorpus data?

2. How are semantic relations constructed through Urduized words across different discourse contexts in Pakistani English?

3. To what extent do existing semantic annotation frameworks adequately capture the meanings of Urduized words, and how can a localized tag-set improve their representation in corpus linguistics and NLP?

2. Literature Review

2.1 Urduized Words and Pakistani English

The incorporation of local-language lexicon into English has long been identified as a core feature of South Asian Englishes. Kachru(1983) conceptualized such features as markers of nativization, whereby English adapts to local sociocultural contexts. In the Pakistani context, Baumgardner(1993) documented systematic lexical and semantic patterns distinguishing Pakistani English from other varieties. Findings of previous investigations were interpreted to discover the nature and extent of the cultural and ideological stance of the local variety. These interpretations produce the categorization of Urduized words for tagging tools and frameworks. Developments of such frameworks require thematic analysis of these words as all features need to be accommodated in tagging systems. Kachru (1983) and Baumgardner et al. (1993) observed grammatical and semantic aspects and believed that these various borrowings from different contextual areas fill the gap of lexis and convey local socio-cultural context with different shades of meaning, researchers followed them, consider these insertions as Pakistanization of English. Recent corpus-based scholarship has increasingly foregrounded the methodological and analytical value of corpus approaches for examining meaning, discourse, and representation in applied linguistics. Zahra, Zahra, and Saleh (2025a), through a five-year content analysis of articles published in *CORPORUM* (2020–2024), demonstrate a clear shift in corpus research towards qualitative and interpretive orientations, particularly corpus-assisted discourse analysis (CADS). Their study highlights how contemporary corpus research moves beyond frequency-driven analysis to explore social meaning, ideology, and representation, especially in postcolonial and multilingual contexts. Complementing this methodological insight, Zahra, Zahra, and Abbas (2025b) provide an empirical illustration of such qualitative corpus work by examining the representation of women through the Urduized term *bibi* in postcolonial fiction. Their corpus-assisted discourse analysis reveals that *bibi* functions as a culturally loaded lexical item encoding authority, respect, and moral agency, challenging Western feminist assumptions that equate female representation with domestic confinement or subordination. Taken together, these studies establish both the methodological legitimacy of qualitative corpus approaches and their analytical potential for uncovering culturally situated meanings, thereby providing a strong foundation for corpus-based investigations of gender, power, and localized semantic practices in Pakistani English.

Urduized words in Pakistani English are not random insertions but culturally motivated choices that often resist substitution by English equivalents. Their meanings are socially shared and contextually negotiated, making them particularly relevant for semantic analysis.

2.2 Semantic Annotation and Cultural Meaning

Semantic annotation aims to classify lexical items according to meaning-based categories. However, most established systems prioritize universal or Western semantic domains, marginalizing culturally specific concepts. Corpus linguistics has increasingly emphasized

data-driven semantic categorization, where meaning emerges from usage patterns rather than pre-defined taxonomies (McEnery & Hardie, 2012).

2.3 Corpus Resources and PakLocCorpus

PakLocCorpus is a purpose-built corpus designed to document localized features of Pakistani English. A key component of the project is its Urduized word list, which provides researchers with curated lexical items supported by source references and concordance access. This resource enables systematic analysis of Urduized words in authentic contexts and provides an empirical foundation for semantic annotation. Once new data is added for the expansion of the corpus or more categories are inserted in the data sets, newly found Urduized words can be added manually. Polysemous words can be described through corpus-based evidence and one single entry can have multiple meanings in both literal and metaphorical sense (Zahra et al., 2025). Corpus analysis offers an in-depth analysis of the unique features possessed in the morpho-syntactic patterns of Urduized words. Urdu contains a wide range of inflected forms in nouns and verbs.

3. Data and Methodology

3.1 Corpus Data

The primary dataset for this study is drawn from PakLocCorpus, comprising written Pakistani English texts across genres. After the collection of words, pronouns were excluded as the study is limited to lexical words only. Greetings and supplications such as bismillah and astaghfirullah in Arabic or Persian (shah) were also excluded. Words from other local languages, such as Punjabi and Sindhi, were also excluded. References, keywords, headings, tables, and figures were also removed from academic discourse. The aim of the study was to extract Urduized words for syntactic and semantic tagging and the Urduized words used as names of people, places and things needed to be excluded and it was only possible through collocation and concordance analysis. The Urduized word list served as the initial inventory of candidate lexical items for annotation.

3.2 Selection Criteria

The study focuses on:

- Single-word Urduized items
- Compound hybrids (e.g., Urdu–English combinations)

Excluded from analysis were:

- Proper nouns
- Full Urdu phrases
- Code-switched clauses

This restriction ensures analytical consistency and semantic tractability.

3.3 Analytical Procedure

A cyclic corpus-assisted procedure was adopted:

1. Selection of Urduized words from the list
2. Extraction of concordance lines
3. Contextual semantic interpretation
4. Assignment of semantic tags
5. Iterative refinement of categories

Semantic decisions were grounded in usage patterns, not dictionary definitions alone. List of Urduized words was developed after keyword analysis and then refined after collocation and concordance analysis. All words were manually extracted and then organized in the form of list for further analysis of syntactic and semantic patterns.

4. Findings and Discussion

Drawing on Gee's theory of social language(2011), the discussion moves beyond a purely lexical or structural interpretation of Urduized words and instead conceptualizes them as

meaning-in-use resources**.** Gee argues that language is always socially situated and that speakers and writers use particular linguistic forms to enact recognizable social practices, identities, and relationships. Within this framework, Urduized words in Pakistani English cannot be understood as simple lexical borrowings or stylistic embellishments. Rather, they operate as components of a localized social language of English, one that is intelligible and meaningful within the Pakistani sociocultural context.

The corpus evidence from PakLocCorpus demonstrates that Urduized words recur systematically across genres and registers, indicating that they are not exceptional or marginal. Their frequency and patterned usage suggest that they are integral to how Pakistani English users make meaning, negotiate identity, and position themselves socially.

A key insight from the analysis is that the meanings of Urduized words are situated rather than fixed. Gee (2011)emphasizes that words take on particular meanings in specific contexts, shaped by social practices and shared cultural knowledge. This principle is clearly reflected in the corpus data. For example, words such as *chacha*, and *chachi* consistently occur in contexts that activate culturally specific frames of reference. *Chachi* indexes not merely the concept of a person, but a culturally sanctioned contract embedded within family and social norms. *Chacha* functions as more than a relation; it is associated with closeness, authority, and communal identity. Similarly, relational terms such as *chacha* and *chachi* evokes a socially dense space characterized by familiarity, surveillance, and communal belonging, which is not captured by its closest English equivalent, *uncle* or *aunty*. Figure 1 shows concordance results of 'chachi'from PakLocCorpus.

Figure 1: Concordance result of '*chacha*' from PakLocCorpus

Concordance

Sr.	Left Text	Center Word	Right Text
1	'but when did this happen?' fatty	chacha	asks
2	fatty	chacha	remains standing
3	' i can see that fatty	chacha	is doubtful, and i'm about to
4	' 'he was far away,' fatty	chacha	points out, sitting down
5	'you don't look fine, champ,' fatty	chacha	says
6	hand as i speak and fatty	chacha	keeps shaking his head, whether in
7	'he's doing well,' fatty	chacha	says
8	leave, visibly reluctant to go, fatty	chacha	insists on giving me five hundred
9	sabeen's	chacha	zahid, ashiq's younger brother, had rushed
10		chacha	zahid was next to me and

Figure 1 shows concordance result of the selected word' *chacha*' from PakLocCorpus. Urduized kinship term *chacha* operates as a social language marker rather than a neutral lexical substitute for *uncle*. The concordance lines show *chacha* repeatedly embedded within speech, stance-taking, and embodied actions (e.g., *asks*, *says*, *points out*, *remains standing*, *keeps shaking his head*). These patterns demonstrate that *chacha* is used to enact a recognizable social role associated with paternal authority, moral oversight, and relational power, rather than merely to identify a familial relation. Applying Gee's situated meaning principle, *chacha* acquires its meaning through use in interactional contexts where authority

is exercised discursively. In the concordances, *chacha* frequently initiates or controls conversational turns (*asks*, *says*, *insists*), positioning him as a legitimate evaluator of events and people. This reflects a culturally grounded understanding of paternal kin as figures of guidance and judgment, a meaning that cannot be conveyed by the English term *uncle*. The social language enacted here presupposes shared cultural knowledge in which *chacha* embodies seniority, responsibility, and moral legitimacy.

In terms of Gee's identity building, *chacha* indexes an identity that is simultaneously familial and authoritative. The repeated association of *chacha* with directive speech and evaluative commentary constructs him as a figure whose voice carries weight. Notably, this authority is enacted through relational proximity rather than institutional power, aligning with Eastern sociocultural models where kinship-based authority is normalized and socially sanctioned. The concordance evidence also highlights Gee's relationship building task. *Chacha* is positioned in close relational interaction with the narrator and other characters, reinforcing bonds of obligation and care. For example, actions such as insisting on giving money or expressing concern ("you don't look fine") illustrate authority intertwined with responsibility and affect. This demonstrates that power here is not coercive but relational, operating within a framework of familial duty. Based on these discourse patterns, *chacha* is appropriately assigned the semantic tag PB (People Label).

Importantly, PB does not imply dominance in a Western patriarchal sense but reflects relational authority embedded in family structure, consistent with Eastern epistemologies of power. Overall, the concordance analysis demonstrates that *chacha* functions as a socially saturated kinship label that performs cultural, relational, and authoritative work in Pakistani English discourse. Through Gee's framework, it becomes clear that writers use *chacha* to activate a localized social language in which paternal kinship legitimizes voice, stance, and moral authority. The semantic tag PB effectively captures this function, offering a culturally sensitive annotation that avoids reductive English glosses such as '*uncle*'.

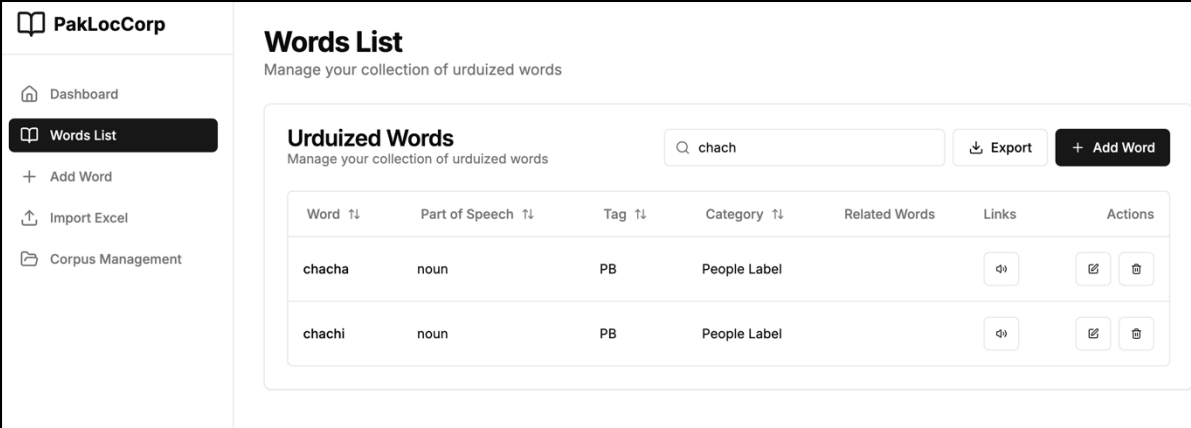Figure 2: Concordance result of '*chachi*' from PakLocCorpus

Concordance

Sr.	Left Text	Center Word	Right Text
1	" her hands trembling, his	chachi	washed the wound on his head
2	to the other side of the	chachi	he stood before the barbed wire
3	"and your	chachi	married makhan singh," he said, indicating
4	but chance—reunited him with his noni	chachi	and iqbal chacha
5	be with his uncle and noni	chachi	and his cousins
6	'i'll ask rukhsana	chachi	to arrange more clothes for you
7	was, or what happened, before my	chachi	started coughing wildly
8	'i'll ask rukhsana	chachi	to arrange more clothes for you
9	your	chachi	called today from karachi
10	her spoiled daughters, thought that my	chachi	had done some ilm on me

From the perspective of social language, the Urduized kinship term *chachi* functions as a socially enacted role rather than a simple feminine counterpart of *uncle* or *aunt*. The concordance lines show *chachi* embedded in actions of care, organization, mediation, and moral interpretation (e.g., *washed the wound*, *arrange more clothes*, *called today*). These

patterns indicate that *chachi* is used to enact a maternal–managerial identity grounded in kinship authority. Applying Gee's principle of situated meaning, *chachi* acquires meaning through repeated association with caregiving and intervention. In the concordance data, *chachi* performs physically and emotionally consequential actions, washing wounds, arranging clothing, responding to illness, and managing family communication. These actions are not peripheral; they place *chachi* at the center of crisis response and everyday governance within the family. Unlike the English term *aunt*, which is semantically vague, *chachi* indexes proximal authority within the paternal household, activating expectations of responsibility, vigilance, and entitlement to intervene. Meaning here is grounded in practice, not lexical definition.

In terms of Gee's identity building task, *chachi* constructs an identity that blends nurturance with control. Importantly, this identity is authoritative rather than submissive. *Chachi* does not merely serve; she evaluates, decides, and acts. Her authority is normalized within the social language of Pakistani English, reflecting culturally sanctioned female power exercised through kinship and moral knowledge. Gee emphasizes that social languages enact relationships. The concordances show *chachi* operating as a relational hub, coordinating between family members, maintaining ties across distance ("called today from Karachi"), and managing resources. This positions *chachi* as a mediator whose role sustains family cohesion. This relational authority is distinct from the directive authority observed with *chacha*. While *chacha* often controls discourse through speech acts (asking, insisting), *chachi* controls social flow through action, care, and coordination. Together, they form complementary authority structures. Crucially, the data does not support a Western feminist reading in which caregiving equates to confinement or lack of power. Instead, *chachi*'s authority is enacted through socially indispensable labor and moral positioning. Her power over male and female family members is implicit but effective, demonstrating an Eastern feminist model of authority rooted in relational centrality. Based on the concordance evidence, *chachi* is assigned the semantic tag (PB) from the category of people label. Tags are non-exclusive; a word may receive multiple tags depending on context. Figure 3 shows entry list of Urduized words for both tags.

Figure 3: Entry List of Urduized Words for '*chacha*' and '*chachi*'

Figure 3 shows search results of both words. Add word option is available on the top right side whereas dash board and other tools are on the left side of the window. Both are nouns in grammatical category. Audio file displays production of the word and can be heard in the list of Urduized words. The corpus-assisted discourse analysis demonstrates that *chachi* functions as a semantically dense, authority-bearing kinship label in Pakistani English. Through Gee's social language framework, *chachi* emerges as a figure of maternal authority, moral

regulation, and relational governance. The semantic tag accurately reflects this role, offering a culturally grounded alternative to reductive gendered interpretations based on Western feminist assumptions. These findings support the argument that Urduized words develop semantic relations through repeated contextual use, and that their meanings are co-constructed by writers and readers who share cultural competence within the Pakistani social context. Another central function of Urduized words is their role in identity construction. According to Gee(2011), social languages allow individuals to enact particular identities that are socially recognizable. The use of Urduized words enables writers to align themselves with local cultural norms while writing in English, thereby negotiating a hybrid linguistic identity.

Corpus patterns reveal that Urduized words are often used without glossing or explanation, presupposing reader familiarity. This presupposition itself is an identity move, positioning both writer and reader as insiders within a shared cultural community. In this way, Urduized words function as markers of culture and identity within English discourse, reinforcing local identity while maintaining participation in a global language. The study thus demonstrates that Urduized words are not evidence of linguistic deficiency or incomplete competence in English, but rather strategic and meaningful choices that reflect sociocultural affiliation and linguistic ownership.

Gee's framework also highlights the role of language in enacting social relationships. Urduized honorifics and politeness markers such as *sahib* and *ji* illustrate how writers use localized resources to manage respect, hierarchy, and interpersonal distance. Corpus evidence shows that these words frequently occur in institutional, formal, or deferential contexts, where relational meanings are foregrounded. Their pragmatic force lies not in propositional content but in the social work they perform. For instance, the use of *sahib* signals respect and authority in ways that standard English titles may fail to capture within the Pakistani context. This finding underscores the inadequacy of semantic frameworks that ignore pragmatic and relational dimensions of meaning. Urduized words demonstrate that semantic relations in Pakistani English are deeply intertwined with social structure and cultural norms.

The findings have significant implications for semantic annotation practices. Traditional semantic annotation systems tend to treat meaning as stable and universal, an approach that is misaligned with Gee's conception of meaning as socially situated. Urduized words challenge such systems because their meanings cannot be fully captured through decontextualized labels. The semantic tagset proposed in this study, aligned with PakLocCorpus data, responds to this challenge by incorporating culturally grounded categories such as food culture, kinship, religion, honorifics, and social organization. By doing so, it captures not only what Urduized words refer to, but also what they do socially.

This approach demonstrates how corpus linguistics and discourse theory can be productively integrated, ensuring that annotation frameworks reflect the lived linguistic realities of World Englishes. This study set out to examine how writers develop semantic relations through the use of Urduized words in Pakistani English, drawing on corpus evidence and Gee's theory of social language. The findings demonstrate that Urduized words function as socially situated meaning-making resources rather than as peripheral lexical borrowings. Their meanings emerge through recurrent patterns of use that are culturally intelligible to Pakistani English users.

From a social language perspective, meaning is not inherent in linguistic forms but arises from how those forms are used to enact recognizable social practices, identities, and relationships. The corpus evidence shows that Urduized words such as *nikah*, *biryani*, and *mohalla* consistently activate culturally specific semantic frames. These words cannot be adequately glossed by their English equivalents without a substantial loss of meaning. Their semantic relations are developed through co-textual patterns that index religion, social

organization, and communal life, confirming Gee's claim that words acquire meaning through social use rather than abstract definition.

The analysis further reveals that Urduized words play a central role in identity construction. Their unmarked use in English texts presupposes shared cultural knowledge and positions both writers and readers as members of a localized interpretive community. This aligns with Gee's argument that social languages allow individuals to enact socially recognizable identities. In this sense, Urduized words serve as linguistic markers of Pakistani identity within English discourse, enabling writers to assert local belonging while operating within a global language. Urduized words also function to enact social relations and pragmatic meanings, particularly through honorifics and politeness markers. Items such as *sahib* and *ji* perform relational work by signalling respect, hierarchy, and social distance. These meanings are not merely semantic but interactional, reinforcing the argument that semantic relations in Pakistani English are deeply intertwined with social structure. Conventional semantic annotation frameworks, which prioritize denotational meaning, fail to account for this relational dimension.

Based on concordance evidence, both '*chacha*' and '*chachi*' are assigned the category of people label (PB). Taken together, these findings highlight the limitations of existing semantic models that treat meaning as static and universal. By contrast, the PakLoc-aligned semantic tagset proposed in this study captures culturally grounded domains allowing for a more accurate representation of meaning-in-use. This approach demonstrates the value of integrating corpus linguistics with social language theory in the study of World Englishes. The analysis demonstrates that Urduized words encode semantic density, carrying cultural, pragmatic, and identity-related meanings simultaneously. For example, *nikah* cannot be adequately translated as "marriage" without losing its religious and legal connotations. Conventional English semantic frameworks struggle to accommodate such culturally embedded meanings. By contrast, the PakLocCorpus-aligned tagset allows for context-sensitive annotation, capturing meaning as it is socially constructed in Pakistani English. This study contributes a replicable method for annotating localized lexical items, advancing descriptive work on World Englishes through corpus evidence. Semantic annotation of Urduized words improves:

- Sentiment and stance detection
- Information extraction
- Machine translation involving Pakistani English

Without such annotation, NLP systems risk perpetuating linguistic bias against non-standard varieties.

Conclusion

By aligning semantic annotation with the PakLocCorpus Urduized word list, this study offers a robust, corpus-driven framework for analysing culturally embedded lexical items in Pakistani English. Urduized words emerge not as marginal borrowings but as central meaning-making resources. Incorporating them into semantic annotation practices is essential for both linguistic inclusivity and technological equity.

This paper has argued that Urduized words in Pakistani English function as integral components of a localized social language. Drawing on corpus evidence and Gee's theory of social language, the study shows that writers develop semantic relations through Urduized words by embedding culturally situated meanings, constructing identities, and negotiating social relationships. Rather than viewing Urduized words as instances of linguistic interference or code-mixing, this research demonstrates that they are systematic, meaningful, and socially motivated choices. Their semantic richness challenges Anglocentric models of meaning and underscores the need for culturally sensitive semantic annotation frameworks.

The study contributes theoretically by extending Gee's concept of social language to corpus-based semantic analysis, and methodologically by proposing a replicable annotation framework grounded in authentic Pakistani English data. These contributions have practical implications for corpus linguistics and natural language processing, particularly in the development of language technologies that adequately represent non-standard varieties of English.

In conclusion, recognizing Urduized words as social language resources is essential for understanding how meaning is constructed in Pakistani English. Incorporating such resources into semantic annotation practices not only advances linguistic description but also promotes more inclusive and representative models of English in global and computational contexts.

References

Baumgardner, R. J. (1993). *The English language in Pakistan*. Oxford University Press.

Gee, J. P. (2011). *How to do discourse analysis: A toolkit* (2nd Ed.). Routledge. https://doi.org/10.4324/9780203850992

Kachru, B. B. (1983). The Indianization of English: *The English language in India*. Oxford University Press. https://doi.org/10.1017/s026607840000198x

Kachru, B. B. (1986). *The alchemy of English: The spread, functions, and models of non-native Englishes*. Pergamon.

McEnery, T. & Hardie, A (2012) "*Support website for Corpus Linguistics: Method, theory and practice*". Available online at http://corpora.lancs.ac.uk/clmtp.

Mehboob, A. (2013). Pakistani English. In E. L. Low & A. Hashim (Eds.), *English in Southeast Asia* (pp. 103–122). John Benjamins.

Zahra, F., Zahra, T., & Abbas, A. (2022, June 15). PakLocCorp. Retrieved July 8, 2022, from https://pakloccorp.com/

Zahra, F., Zahra, T., & Abbas, A. (2023 a). Neoliberal Co-option of English: A Study of Urduized Terms in Pakistani Magazine Articles through Corpus Approaches. *Rivista Italiana di Paleontologia e Stratigrafia. RIFL, 17*(2), 174-188

Zahra, F., Zahra, T., & Abbas, A. (2023 b). Corpus assisted discourse analysis of Urduized words and selected collocation patterns: a case study of trespassing by Uzma Aslam Khan. *Research Journal of Language and Literary Studies, 3(2), 125–144.*

Zahra, F. T., Zahra, T., & Saleh, M. (2025a). Content analysis of corpus research across five years: Study of journal articles (CORPORUM, 2020-2024). *Journal of Applied Linguistics and TESOL (JALT)*, *8*(4), 219-232.

Zahra, F., Zahra, T., & Abbas, A. (2025b). Women Representation in Post-Colonial Fiction: A Corpus-Assisted Discourse Analysis of The Selected Word 'Bibi'. Qualitative research Journal for social Studies, *2*(4)*,* 220-229