

PAKLOCCORPUS AS A DATA-DRIVEN LEARNING TOOL IN ESL CLASSROOMS

Fatima Tuz Zahra

PhD Scholar, Air University Islamabad

Lecturer, Minhaj University, Lahore

fatematuzzehra@gmail.com

Kalsoom Afzaal

PhD Scholar Department of English, Lecturer at Minhaj University Lahore

afzaalkalsoom@hotmail.com

Abstract

PakLocCorpus represents a locally grounded linguistic resource reflecting authentic usage of English in Pakistan across academic, media, and literary registers. Through classroom experiments involving corpus searches, collocation analysis, and context-based vocabulary learning, students engaged with data reflecting their own sociolinguistic reality. This study explores the use of PakLocCorpus, a purpose-built corpus of Pakistani English, as a Data-Driven Learning (DDL) tool in English as a Second Language (ESL) classrooms. Drawing inspiration from the Corpus of Contemporary American English (COCA) classroom study by Lopes (2013), social language tool from Gee (2011) and guidelines from Smith and Robbins (2023), this research investigates how Pakistani ESL learners can use corpus consultation to improve vocabulary, grammar, and cultural awareness. Results reveal significant gains in lexical competence, accuracy in collocation use, grammatical understanding, and positive attitudes toward Pakistani English. Findings show that Urduized words have a high acceptance rate in ESL students. The study argues that data-driven learning with PakLocCorpus bridges the gap between global linguistic competence and local identity. It empowers learners as researchers of their own language, fosters autonomy, and integrates local culture in the ESL curriculum. The paper concludes that PakLocCorpus can play a pivotal role in legitimizing Pakistani English and transforming English language pedagogy in Pakistan.

Keywords: Data driven learning, PakLocCorpus, global linguistic competence, pedagogy, Urduized words

Introduction

Language learning in the 21st century has shifted toward inquiry-based, evidence-driven models that encourage learners to discover language patterns through authentic data rather than memorizing rules. Data-Driven Learning (DDL) lies at the heart of this pedagogical shift, utilizing corpora, large electronic collections of authentic texts—to expose learners to real-life language usage. Corpora such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) have been instrumental in advancing corpus-informed teaching (Boulton, 2011; Bennett, 2010). However, such resources reflect primarily native English usage, leaving a pedagogical gap for learners of postcolonial or localized English varieties.

PakLocCorpus addresses this gap as a specialized corpus documenting Pakistani English across multiple genres including media, academic writing, fiction, and political discourse (Zahra, Zahra, & Abbas, 2023a; 2024). It contains millions of words of text that authentically capture how English operates within Pakistani sociocultural contexts. More importantly, it includes Urduized words and hybrid expressions, linguistic innovations resulting from the contact between Urdu and English. Words such as 'biryani', 'madrassa', 'dupatta', and 'lota' represent cultural referents embedded in Pakistani identity, and their usage in English discourse highlights the dynamic bilingual reality of Pakistan.

As an educational tool, PakLocCorpus (Zahra et al., 2022) provides a unique opportunity for ESL students to analyze English as used by their own communities. Instead of seeing local English as a deviation from British or American norms, learners can view it as a legitimate and evolving variety. This perspective aligns with the principles of World Englishes (Kachru, 1985), where local varieties are seen as distinct yet systematic linguistic systems. Zahra et al. (in press) emphasize that developing localized corpora not only supports descriptive linguistic research but also assists in revising curricula, textbooks, and grammar manuals to reflect authentic regional use.

This study explores how PakLocCorpus can be integrated into ESL classrooms as a DDL resource. The paper seeks to address the following questions:

1. How can ESL learners utilize PakLocCorpus to improve their vocabulary, collocation, and grammatical accuracy?
2. In what ways does corpus consultation enhance learners' awareness of cultural and contextual aspects of language?
3. What pedagogical outcomes emerge from integrating a local corpus into ESL instruction?

By addressing these questions, this research aims to demonstrate the pedagogical potential of Pakistani English through PakLocCorpus not only as a linguistic resource but as a transformative tool that validates the linguistic identity of Pakistani learners while enhancing their global communicative competence. Additionally, it is a step forward to legitimize Pakistani English.

Literature Review

Corpus-based pedagogy, often referred to as Data-Driven Learning (DDL), has evolved as a cornerstone of modern language instruction (Bennett, 2010). The approach promotes learner autonomy by encouraging students to observe, hypothesize, and confirm linguistic patterns using empirical data from corpora. Callies (2013) famously described learner corpora as a 'discovery learning' model where students become language detectives. By examining how words and structures occur in authentic contexts, learners develop both grammatical accuracy and pragmatic intuition. In Lopes's (2013) classroom-based study with COCA, undergraduate students used corpus tasks to explore cultural and lexical dimensions of English. They analyzed real data to understand how cultural referents, such as names, institutions, and events, were represented linguistically. Lopes observed that this hands-on engagement improved learners' awareness of the relationship between language and culture. His findings support the idea that corpus consultation bridges cognitive and cultural aspects of learning by exposing students to authentic discourse. Smith and Robbins (2023) further illustrated how learners can use corpora for vocabulary development. Using COCA, they demonstrated how corpora clarify collocational norms. For example, while a dictionary may equate 'profound' with 'deep', corpus evidence shows 'deep water' but 'profound effect', revealing natural pairings that dictionaries cannot. Such activities make learners more sensitive to collocational appropriacy, a key aspect of fluent and idiomatic usage. Boulton (2011) and Callies (2013) expanded this perspective, showing that corpus-based tasks improve long-term language awareness. By exploring real-world usage, learners move beyond memorization toward hypothesis testing and critical engagement with language. Corbett (2010) and Kramsch (1998) added that integrating cultural context into language teaching is vital, as language cannot be detached from the worldview it encodes. In Pakistan, research by Zahra, Zahra, and Abbas (2023a, 2023b, 2024) and Zahra et al. (in press) has contributed to the understanding of Pakistani English as a legitimate variety. Their corpus-assisted studies analyzed Urduized words and hybrid constructions, confirming that Pakistani English exhibits stable lexical and syntactic patterns shaped by bilingualism and

culture. Zahra et al. (2023a) found that Urduized vocabulary in Pakistani magazines often reflects local socio-political realities, while Zahra et al. (2023b) demonstrated that collocational tendencies in literary texts reveal consistent linguistic behavior across genres. These findings align with global corpus research that recognizes postcolonial Englishes as systematic rather than deviant (Schneider, 2019). It indicate a distinct increase of awareness in all assessed aspects for the majority of participants. Suggestions for other teacher education programs are provided to explore similar field-based training modules.

PakLocCorpus, as described by Zahra et al. (2024), was designed not merely as a research database but as an educational tool. It provides online access to authentic Pakistani English texts and includes features such as keyword searches, frequency lists, and collocation analysis. Its interface resembles major corpus platforms like Sketch Engine and AntConc, but with localized data. This makes it accessible for ESL learners and teachers in Pakistan who wish to explore authentic English use within their own linguistic environment. Building on these foundations, the present study operationalizes DDL principles through classroom-based tasks using PakLocCorpus. The objective is to examine how learners' vocabulary, grammatical awareness, and cultural literacy can be enhanced through guided corpus exploration. Recent research by Zahra and colleagues (2023a, 2023b, 2024, in press, 2025) provides an essential localized foundation for understanding the pedagogical and sociolinguistic dimensions of corpus-assisted learning in Pakistan. Zahra et al. (2023a) examined Urduized vocabulary in Pakistani magazine discourse, showing how lexical borrowing, transliteration, and semantic shift express localized ideologies and socio-economic conditions. These findings underscore the need to include such hybridized forms in ESL instruction to reflect authentic linguistic behavior rather than enforcing imported norms. Zahra et al. (2023b) analyzed collocational patterns in literary discourse, revealing how Urdu-English code-mixing contributes to stylistic nuance and identity construction. Integrating such data into ESL classrooms allows students to appreciate the creativity embedded in localized English usage. In a forthcoming study, Zahra et al. (in press, 2026) detail the technical and theoretical framework for developing PakLocCorpus itself as an online specialized corpus of Pakistani English. This initiative integrates principles from corpus linguistics, data ethics, and linguistic diversity to ensure accessibility for educators and students. PakLocCorpus thus functions not merely as a research resource but as a pedagogical infrastructure supporting reflective, evidence-based language teaching.

Complementary research by Zahra et al. (2025) in their content analysis of five years of corpus-based studies in CORPORA journal reveals that academic discourse in Pakistan is increasingly data-driven, reflecting a paradigm shift in language research. Their meta-analysis found that most studies now adopt corpus-assisted discourse analysis (CADA) approaches to examine political, media, and academic registers. This trend indicates a growing awareness among Pakistani educators and researchers about how corpus tools democratize access to linguistic data and encourage transparent analytical reasoning in ESL pedagogy. These cumulative contributions emphasize how localized corpus research—especially through PakLocCorpus, supports linguistic decolonization. By valuing Urduized lexis and regional patterns, teachers can bridge the gap between standardized English instruction and students lived linguistic experiences. PakLocCorpus represents a strategic shift toward contextualized language pedagogy, aligning with Kramsch's (1998) argument that language teaching must reflect the symbolic dimensions of culture. It provides both a research foundation and a pedagogical roadmap for integrating cultural and linguistic hybridity into classroom discourse.

Methodology

This study employed a qualitative action research design combined with quantitative evaluation to investigate how PakLocCorpus can be effectively used as a Data-Driven Learning (DDL) resource in ESL classrooms. Following the pedagogical frameworks outlined by Lopes (2013) and informed by Gee's (2011) theory of Social Languages, the study aimed to explore how corpus consultation promotes vocabulary growth, grammatical awareness, and cultural identity integration among Pakistani learners of English.

Participants

The participants were 30 undergraduate ESL students enrolled in a BS English program at a university in Lahore. They represented diverse linguistic backgrounds (Urdu, Punjabi, Pashto, and Sindhi) and had intermediate English proficiency (CEFR B1–B2). Their exposure to Pakistani English vocabulary and Urduized lexical items was moderate but largely informal. The group provided a suitable context for testing corpus-based pedagogy because it reflected Pakistan's multilingual reality.

Instruments and Corpus Interface

PakLocCorpus served as the primary learning instrument, consisting of subcorpora from academic, media, and literary registers. The corpus interface allowed students to conduct keyword searches, collocation analysis, and concordance line inspection. The website also offered visualization tools (frequency lists, KWIC view, and collocate clusters) that helped students understand contextual meaning and frequency-based usage patterns.

Students were introduced to corpus linguistics concepts, including:

- **Keyword-in-Context (KWIC)** searches
- **Collocation networks** for discovering natural word partnerships
- **Frequency lists** for identifying dominant lexical patterns

Procedures

The research was conducted over an sixteen-week period with structured weekly activities:

1. Orientation Session:

Students were introduced to DDL principles, corpus navigation, and sample searches using familiar Pakistani words such as *neem*, *mehndi*, and *biryani*. The instructor demonstrated concordance lines to highlight cultural expressions embedded in English.

2. Collaborative Exploration:

Working in pairs, students conducted guided searches using the *KWIC* function to identify collocations (e.g., *neem leaf*, *neem tree*, *neem oil*). They examined surrounding lexical patterns to infer meaning, tone, and cultural significance.

3. Analytical Writing Task:

Each student wrote a short reflection describing how a corpus-derived expression could be integrated into personal writing. They were encouraged to explore native Pakistani lexical creativity, analysing *Urduized English* words documented by Zahra et al. (2023a, 2023b, 2024).

4. Discussion and Reflection:

The class discussed the findings in light of Gee's Social Language Tool, considering how lexical choices (like *neem* or *lota*) express not only linguistic meaning but also cultural and ideological stance.

Gee's (2011) Social Language Framework

Gee (2011) posits that every language variety represents a social language—a way of using words, grammar, and discourse features to enact identity and convey belonging to a particular community. Language, in his view, is inseparable from social practices and cultural meanings.

In this study, Gee's framework guided:

- **Data interpretation:** The way *neem* and other localized terms create *situated meanings* within Pakistani culture.
- **Classroom discourse analysis:** Students' reflections were examined for shifts in their perception of Pakistani English as legitimate and expressive.
- **Pedagogical focus:** Encouraging learners to adopt corpus inquiry as a social act, connecting linguistic form to social function and identity representation.

PakLocCorpus thus became both a tool and a social space, where learners engaged with authentic, localized English that reflected their lived linguistic environment.

Data Collection and Analysis

Data were collected from:

- Corpus search logs (frequency and collocation results)
- Student reflections and written tasks
- Pre- and post-intervention vocabulary quizzes
- Focus group interviews assessing learner attitudes

Quantitative data (vocabulary gains and accuracy rates) were analyzed descriptively. Qualitative data (reflections and interviews) were coded thematically under categories aligned with Gee's (2011) framework:

1. **Identity Construction** (how learners perceive themselves as bilingual English users)
2. **Cultural Legitimization** (how corpus data validate local English use)
3. **Linguistic Autonomy** (how learners independently employ data for writing improvement)

The research was conducted in a Pakistani university ESL classroom with 25 undergraduate students enrolled in an Applied Linguistics course. Most participants were B1–B2 level according to the CEFR framework. The study followed an action-research design, consisting of training sessions, guided corpus tasks, and reflection.

1. Training Phase: Students were first introduced to the concept of corpora and DDL. The instructor demonstrated the PakLocCorpus interface, including keyword searches, collocation lists, and concordance lines. Using examples such as the words 'neem', 'Eid', and 'kebab', the class explored the Key Word in Context (KWIC) function. Students learned to interpret patterns.

2. Vocabulary Collocation Task: Each student selected five target words from their course readings. They used PakLocCorpus to find common collocates and example sentences. Students then composed short paragraphs using these collocations in context. The results were compared with pre-corpus writing samples. Improvements were evaluated in terms of collocational accuracy and lexical variety.

3. Grammar Exploration Task: Learners examined verb-preposition patterns. Using corpus searches, they identified the most frequent combinations and corrected previously common errors. They presented their findings in small groups, explaining usage trends.

4. Cultural Term Analysis: Groups were assigned Pakistani cultural keywords (e.g., 'neem'). They searched PakLocCorpus for contexts of local terms, noted surrounding adjectives and verbs, and analyzed concordance. They discovered, for example, that 'neem' often collocates with 'shade', 'shadows', and 'tooth cleaning', indicating religious and cultural significance of the neem tree.

5. Feedback and Reflection: Students completed questionnaires on the usefulness of corpus-based learning. Reflections were qualitatively coded to extract themes about learner autonomy and attitudes.

Findings and Discussion

Zahra et al. (2024) explored intercultural communication and clothing-related discourse, identifying how English in Pakistan reflects global consumerism while embedding traditional symbols. Their corpus-assisted discourse analysis demonstrated that digital corpora can illuminate socio-cultural semiotics, enriching students' understanding of how English interacts with identity and globalization. By using PakLocCorpus, similar insights can be generated in class to help learners explore lexical fields like trees, family relations, food, and politics, domains central to Pakistani socio-discourse.

Figure 1: Concordance of selected vocabulary item 'neem'.

up so they'd catch in the limbs of the	neem tree and call Hiroko out to sit under the
entry into the courtyard. The shadows of the	neem tree fell exactly where she knew they
earthenware seed-filled plate that hung from the	neem t in the courtyard. She looked out of the
cross it from one end to the other), the sturdy	neem and peepal trees under whose shade
had an immense courtyard and in it had stood	neem trees, and such lovely guava trees
towards the window and said above the din, '	Neem leaves serve as disinfectant. It's advisable
the strictures of the Sunnat. 'I hear you've added	neem leaves to the water,' Asgri said; and
and marchers emerge from the leafy gloam of	neem trees fermenting behind the walls. It is
how he had landed in this profession: Jat ke the	neem julae Us pe ban gae darzi Lote pote ke
is right behind me. "Sana," he says. "Give her	neem leaf and get her out of here." He turns
in death by idiocy—"—bring hot water and	neem leaf, quickly—"—"will be here soon. But
Have Raj and Navid reported yet?" "What's	neem leaf?" I ask Sana when I'm sure Keenan
looking at Gurpal. That day the breeze, playing	neem leaves, would sing anthems and the whole
go to the pump and clean up. There are some	neem sticks for your teeth lying on that sill.
wheels, singing songs in the shade of the	neem trees, I keep quiet. How much life there w

Figure 1 shows concordance of selected vocabulary item from PakLocCorpus. It highlights cultural significance of selected vocabulary item.

Analysis of student outputs and feedback indicated several key outcomes: Vocabulary Gains: Learners improved their awareness of natural collocations. Before using the corpus, many produced phrases like 'big impact' or 'do a crime'. After corpus consultation, they used 'significant impact' and 'commit a crime', mirroring authentic patterns (Smith & Robbins, 2023). Students also used domain-specific collocations—'policy reform', 'economic stability'—reflecting corpus-derived language.

Grammar Accuracy: Error frequency decreased in verb-preposition combinations. Students replaced incorrect forms such as 'discuss about' with 'discuss the issue'. Corpus evidence empowered them to verify grammatical intuition independently, echoing findings from Lopes (2013). Learners also noted register variation: for instance, 'protest against' was common in Pakistani news contexts, while 'protest the law' appeared in formal writing. Cultural Awareness: Cultural term analysis was the most engaging task. Students reported pride in seeing local words—'biryani', 'Eid', 'madrassa'—used in credible English sources. This confirmed that local linguistic elements coexist naturally in English usage (Zahra et al., 2023a). Learners reflected that PakLocCorpus validated their bilingual identity and encouraged them to use culturally specific vocabulary appropriately.

Learner Autonomy: Questionnaire responses revealed increased confidence in using corpus tools. 80% of participants agreed that corpus searches helped them 'think like linguists'. They appreciated discovering answers independently, fostering analytical learning habits. Challenges: Initial difficulties included technical issues, interpreting KWIC lines, and limited digital literacy. However, repeated practice reduced these barriers. Students requested more classroom time for corpus exploration and a user-friendly interface with visual collocation maps.

Findings affirm the pedagogical value of PakLocCorpus in data-driven ESL instruction. Learners not only improved language accuracy but developed critical insight into language variation and identity. These outcomes align with Kramsch's (1998) view that language learning involves both linguistic and cultural competence. The integration of a local corpus redefines how Pakistani English is perceived in education. Traditionally, Pakistani curricula privileged British norms, implicitly marginalizing local expression. By using PakLocCorpus, students see empirical evidence that Pakistani English has stable lexical and syntactic conventions (Zahra et al., 2024). This supports endonormative standardization, the process by which local norms gain legitimacy. From a pedagogical standpoint, corpus use cultivates learner autonomy and analytical thinking. As learners navigate concordance lines, they engage in inductive reasoning, a process that strengthens metalinguistic awareness. This resonates with Boulton's (2011) assertion that corpus learning transforms students into researchers. Moreover, by connecting corpus findings to classroom writing tasks, the approach maintains communicative relevance. In terms of technology integration, PakLocCorpus bridges the digital literacy gap by training learners in research-oriented digital practices. This aligns with global trends in AI-assisted and corpus-informed pedagogy (Zahra et al., 2025). Teachers can extend this framework by designing blended-learning environments where corpus tasks complement writing workshops, grammar instruction, and vocabulary building. Finally, corpus-based teaching enhances inclusivity. By representing regional language, PakLocCorpus offers learners linguistic mirrors instead of distant models. This inclusion strengthens cultural identity and ownership of English—outcomes crucial for sustainable language learning in multilingual societies like Pakistan.

Conclusion

This study demonstrates that the integration of PakLocCorpus into ESL classrooms not only enhances linguistic competence but also redefines how learners engage with English as a socially situated language. The findings confirm that localized corpora serve as powerful pedagogical tools that bridge linguistic inquiry with cultural authenticity. ESL learners can utilize PakLocCorpus to explore authentic patterns of word usage within Pakistani English. By engaging in collocation searches and keyword-in-context (KWIC) analyses, students move beyond rote memorization to discover how words naturally co-occur in discourse. This process reinforces lexical accuracy and helps learners internalize syntactic patterns that are contextually appropriate. Tasks such as frequency comparison and phrase extraction empower learners to correct grammatical inconsistencies autonomously, thereby transforming them into researchers of their own language, a key aim of Data-Driven Learning (DDL). Corpus consultation allows learners to uncover the cultural meanings embedded in language. Through concordances such as the one for *neem*, learners see how vocabulary reflects collective values, traditions, and ideologies. Applying Gee's (2011) Social Language tool, these linguistic traces can be interpreted as enactments of social identity. Students recognize that expressions like *neem leaf*, *biryani rice*, or *mehndi design* embody cultural histories that standard textbooks often overlook. Consequently, they gain awareness of how English, when localized, becomes a

medium of cultural representation rather than imitation. The integration of a local corpus into ESL instruction yields several pedagogical benefits. Learners report increased motivation, ownership of learning, and validation of their bilingual identities. Teachers find that corpus-driven tasks promote critical reflection on how English operates in Pakistani society within academic, media, and literary discourse. The classroom thus becomes a site of linguistic empowerment, where students experience English as both global and personal. In line with Gee's (2011) view of discourse as social practice, learners not only acquire grammatical accuracy but also develop an understanding of how language constructs identity and community. In sum, PakLocCorpus bridges linguistic precision with cultural expression, turning ESL learning into a process of both linguistic mastery and cultural authorship. By combining the analytic rigor of corpus linguistics with the contextual richness of local discourse, it reimagines English education in Pakistan as inclusive, data-driven, and culturally grounded.

References

- Andersen, H. L., Lund, K., & Risager, K. (2006). *Culture in language learning*. Aarhus University Press.
- Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.
- Boulton, A. (2011). Language awareness and medium-term benefits of corpus consultation. In A. Gimeno Sanz (Ed.), *New trends in computer-assisted language learning: Working together* (pp. 39–46). Macmillan ELT.
- Byram, M., & Grundy, P. (Eds.). (2003). *Context and Culture in Language Teaching and Learning*. Clevedon: Multilingual Matters.
- Callies, M. (2013). Advancing the research agenda of interlanguage pragmatics: The role of learner corpora. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies* (pp. 9–36). New York: Springer.
- Corbett, J. (2010). *Intercultural Language Activities*. Cambridge: Cambridge University Press.
- Gee, J. P. (2011). *An introduction to discourse analysis: Theory and method*. Routledge.
- Kachru, B. B. (1985). The bilinguals' creativity. *Annual review of applied linguistics*, 6, 20–33.
- Kramsch, C. (1998). *Language and Culture*. Oxford: Oxford University Press.
- Lopes, A. (2013). Exploring culture-related content in the COCA with task-based activities in the EFL classroom. In L. Bradley & S. Thouësny (Eds.), *20 Years of EUROCALL* (pp. 169–174). Research-publishing.net.
- Schneider, E. (2019). Impact of an English Language Learner Course on Preservice Teachers. *SRATE Journal*, 28(1), 43–51.
- Smith, A., & Robbins, J. (2023, July 22). Use a language corpus to improve your vocabulary. VOA Learning English.
- Zahra, F., Zahra, T., & Abbas, A. (2022, June 15). PakLocCorp. Retrieved July 8, 2025, from <https://pakloccorp.com/>
- Zahra, F. T., Zahra, T., & Abbas, A. (2023a). Neoliberal co-option of English: A study of Urduized terms in Pakistani magazine articles through corpus approaches. *RIFL*, 17(2), 174–188.
- Zahra, F. T., Zahra, T., & Abbas, A. (2023b). 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., & Abbas, A. (2024). Globalization of narratives for clothing and appearance: Corpus-assisted analysis of intercultural communication. *Migration Letters*, 21(13), 532–551.
- Zahra, F. T., Zahra, T., & Abbas, A. (in press; expected 2026). Developing an online specialized corpus of Pakistani English. *Corpora*, 21(3).
- Zahra, F. T., Zahra, T., & Saleh, M. (2025). 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.