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

CORPUS LINGUISTICS AS A TOOL FOR IMPROVING LANGUAGE TEACHING STRATEGIES

Ubaid Ur Rehman
M.Phil Scholar Department of English (Linguistics) University of Okara
Email: *syedubaidurrehman6857@gmail.com*
Ashfaq Mahmood
M.Phil Scholar Department of English (Linguistics) University of Okara
Email: *ashfaqmahmood31@gmail.com*
Muhammad Khuram
Lecturer, Department of English University of Okara
Email: *m.khuram@uo.edu.pk*

Abstract
Corpus Linguistics (CL) has emerged as a powerful approach in reshaping language-teaching practices by offering authentic data drawn from real-world language use. Within the Pakistani context, this study reviews the pedagogical applications of CL in enhancing vocabulary development, grammatical proficiency, and pragmatic awareness through data-driven learning (DDL). Evidence shows that resources such as AntConc, COCA, and learner corpora assist teachers and students in examining lexical patterns, collocations, and discourse structures in meaningful contexts, thereby promoting both linguistic competence and learner independence. Integrating corpus-based instruction addresses the limitations of traditional textbook-centered teaching by aligning more closely with Communicative Language Teaching (CLT) and constructivist perspectives on learning. Nevertheless, challenges remain, including limited teacher expertise, insufficient access to corpus tools, and the technical demands of the approach. Addressing these issues requires professional training, the design of simplified software, and institutional commitment to support classroom integration. The review underscores the importance of further investigation into the long-term outcomes of CL-oriented pedagogy in Pakistani classrooms and its adaptability.

Keywords: *Corpus linguistics, data-driven learning, communicative language teaching, vocabulary development, grammatical proficiency, pragmatic awareness, concordance tools, learner independence, authentic input.*

Introduction

Corpus Linguistics (CL) has become a significant instrument of applied linguistics and it is defining the modern teaching and learning of language. In other words, CL examines language through actual texts that are compiled in huge collections known as corpora (McEnery and Hardie, 2012). The discipline began during the late 1900s and has expanded rapidly, due primarily to the development of technology and the availability of large corpora like the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). The trends have led to the integration of evidence-based, data intensive approaches in research and classrooms (Hunston, 2002).

Teachers in Pakistan have used teacher led-Rule based lessons and elsewhere long before the advent of teacher centered lessons. They present students with fixed grammar rules and vocabulary that typically are not in line with the way people talk or write in real life (Thornbury, 2001). These approaches, even though structured, are criticized as not having any real-world applicability and being not representative of the reality of spoken and written language. The alternative learner-centered approach is provided by corpus linguistics. Through collaboration with real language data, instructors can build tasks, which are based on real language use and, therefore, assist students in identifying patterns in words, grammar, and social language use.

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

This is referred to as Data Driven Learning (DDL), which facilitates discovery learning and provides greater autonomy to students (Johns, 1991).

It has also been found out that real language data teaching has the ability to give positive outcomes in most language learning fields. Indicatively, O'Keeffe, McCarthy and Carter (2007) discovered that students who are exposed to corpus data are able to enhance their vocabulary knowledge particularly when they learn common lexical patterns and collocations. A change in grammar teaching based on corpora changes the emphasis of students who see instructions in the form of strict and prescriptive rules as opposed to the actual use of those rules. This assists learners in differentiating between spoken and written convention and between formal as well as informal styles (Biber et al., 1999). Corpus-based methods are also useful to teach discourse and pragmatic skills- aspects, which are usually overlooked in the traditional, textbook-based classrooms (Flowerdew, 2015).

Advantages

Regardless of all these advantages, the application of corpus linguistics in the Pakistani classrooms is yet to be widespread. The greatest challenge is that corpus tools are technical. A large number of teachers have lacked the specialized training required to utilize them (Romer, 2009). Moreover, the processes of getting access to sound corpora and creating the teaching resources with corpus influence are time-consuming (Boulton, 2012). These issues can only be addressed through professional growth opportunities, development of easier to use tools, and institutional guidance both at the policy and curriculum levels.

This review help how corpus linguistics can be used to enhance language-teaching practices in Pakistan. It answers three key questions as follows: (1) what strategies and instruments exists to incorporate corpus linguistics into language classrooms? (2) How well do corpus-based methods enhance learning in the research? (3) What are the difficulties to their proper application in practice? Through the review of literature, the article will provide a better comprehension to teachers, researchers, and policymakers on how the use of corpus linguistics can be used to reinforce the teaching and learning of languages in Pakistan.

This paper is systematic literature review to investigate the ways in which Corpus Linguistics (CL) can be used effectively to enhance language-teaching methods in the Pakistani environment. To balance the evaluation process, reduce researcher bias, and offer a systematic synthesis of the existing scholarship, a systematic approach was selected. The review process would take three steps, which include (1) the definition of inclusion and exclusion criteria, (2) the identification and review of relevant studies, and (3) the categorization of the findings into final important thematic areas.

Data sources and search tactics

The sources of research material were obtained in the well-known academic databases, Scopus, Web of Science, ERIC, JSTOR, and Google Scholar. They were chosen due to their wide scope of coverage of the applied linguistics, education and English language teaching literature.

The keywords and Boolean combinations were searched using keywords, e.g.

- Corpus linguistics and language teaching
- "data-driven learning" OR "DDL"
- In addition, conjunction used to combine concordance tools with pedagogy.
- Learner corpora and language acquisition.
- Teaching strategies and "authentic language input.

It was restricted to the studies published in 2010-2023, as it was ensure that the review was relevant to the latest development of corpus tools and classroom use.

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

Criteria for Inclusion and Exclusion.

Included in studies were those studies that:

- devoted to the use of CL in teaching and learning
- documented empirical data in the use of learning results
- studied the application of corpus tools, e.g., concordances, learner corpora or frequency based method
- Addressed teacher training of the corpus based pedagogy.

Articles were filtered out in case they:

1. confined himself to the discussion of theory, but nothing practical was done in the classroom;
2. dealt with general linguistics which had no relation to teaching;
3. were dissertations or theses which had not been published;
4. Were published before 2010.

Table 1 gives a summary of the inclusion and exclusion criteria.

Corpus Linguistics in teaching language.

Vocabulary Teaching

Corpus-based methods are very efficient in the development of vocabulary. Students using frequency lists, collocations and real contexts memorize words more efficiently than those who use regular textbooks to learn. According to O'Keeffe et al. (2007), Pakistani students who exercised on collocation data on the BNC and COCA had a better accuracy in using word combinations. Chambers (2019) emphasized that discovery-based learning can assist students in the process of revealing lexical chunks with the assistance of concordance programs like AntConc.

Grammar Instruction

The grammar instruction based on corpus is concerned with actual practice and not prescriptive rules. Biber et al. (1999) discovered that both academic and conversational corpora exposure increased awareness of the difference between registers in students. As an example, students were taught when spoken contractions such as gamma can be used and when they have to be used in writing, which is going to. This induction method works well in Pakistan where grammar teaching is usually prescriptive.

Discourse and Pragmatics

Corpus analysis also contributes to teaching discourse markers, politeness strategies as well as genre specificities which local textbooks tend to ignore. The corpus-based tasks demonstrated by Flowerdew (2015) enhanced the students to use hedging in academic writing. Perez-Paredes (2019) has shown that pragmatic awareness has improved with exposure to conversational markers like well and actually.

Tools and Strategies for Corpus-Based Instruction

Key corpus tools identified in the studies include:

Tool	Description	Example Study
AntConc	Free concordance for word patterns and collocations	Anthony (2012)
COCA	Large corpus covering multiple genres	Davies (2010)
BNC	100-million-word British English corpus	O'Keeffe et al. (2007)
Sketch Engine	Advanced corpus query tool	Kilgarriff et al. (2014)
ICLE	Learner corpus analyzing second-language errors	Granger et al. (2015)

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

The consistent strategy across these tools was Data-Driven Learning (DDL), where learners actively explored corpus data to discover patterns. Johns (1991) found that such discovery-based activities fostered longer-term retention and analytical skills.

Teacher training was another recurring theme. Römer (2009) emphasized that teachers who had training in AntConc and Sketch Engine were more confident in designing classroom activities. However, in Pakistan, limited exposure to corpus tools remains a significant obstacle.

Benefits and Challenges of Corpus-Based Teaching

Authentic Input: Learners gain access to real language, reducing reliance on artificial textbook examples (Boulton, 2012).

Learner Autonomy: Engagement with corpora encourages independent learning and critical thinking (Gilquin & Granger, 2010).

Improved Competence: Studies show improved vocabulary retention, collocation use, and genre awareness (Cobb, 2013).

Challenges

Technical Skills: Many teachers lack the training to use corpus tools effectively (Römer, 2009).

Resource Access: Advanced software and reliable internet connections are not always available in Pakistani institutions (Boulton, 2012).

Preparation Time: Designing and integrating corpus-informed tasks can be demanding, especially in large classrooms (Flowerdew, 2015).

A summary of benefits and challenges is given in Table 2.

Summary of Findings

The review indicates that corpus linguistics can be of great help in the teaching of English in Pakistan. It is highly effective in developing vocabulary of students, raising grammar awareness, and developing discourse skills. However, there are always difficulties: not many teachers are trained, there is not enough resources and time slows things down. The research suggests professional growth, easy to use corpus applications, and greater institutional support in order to maintain integration.

Discussion

The Corpus Linguistics and its implication on teaching strategies in language.

This review identifies a promising future of corpus linguistics (CL) as far as transforming the teaching of the English language in Pakistan is concerned. Since English is typically a second or foreign language there, CL provides a databased learning model that encourages learners to discover actual language data, rather than learning by prescriptive principles. Identified by Johns in 1991, this inductive approach is used to allow students to observe practical patterns of usage and remain actively involved.

Unlike the conventional classroom activities in Pakistan, which are often dominated by textbook-based learning and exam-oriented approaches, corpus-based teaching introduces the learners to high-frequency vocabulary and grammar patterns that, are often not accessible to the local texts (O'Keeffe, McCarthy, and Carter, 2007). As an illustration, the repeated practice with collocations with the help of such tools as AntConc and COCA has been proved to enhance the capacity of learners to use natural phraseology (Boulton and Cobb, 2017). This relates directly to the classroom setting in Pakistan whereby students often create grammatically correct sentences, which are tense or unnatural.

Corpus-informed approaches, too, are supplementary to the constructivist orientation to learning, in which learners are directly involved in building up knowledge, through analyzing contexts of examples. Corpora give real-life examples of grammar as noted by Biber et al.

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

(1999), and since the corpora contain both the formal structure of academic writing (e.g., passive voice, nominalization) and informal speech (e.g., contractions, ellipsis), corpora can assist learners in Pakistan to differentiate between the two. This discovery-based approach not only improves grammatical accuracy but also reduces the overgeneralization and fossilization of errors common in local learners. In line with Communicative Language Teaching (CLT) principles (Thornbury, 2001), corpus-based practices bridge the divide between prescriptive teaching and authentic communication.

Equally important is the role of corpora in enhancing pragmatic competence, a crucial but underdeveloped area in Pakistani ELT. Studies such as Flowerdew (2015) and Pérez-Paredes (2019) demonstrate that exposure to discourse markers, hedges, and speech acts through corpora raises learners' awareness of context-sensitive meaning. This resonates with Hymes' (1972) concept of communicative competence, ensuring that learners' language is not only accurate but also socially and culturally appropriate.

Overcoming the Domains of Corpus-Based Pedagogy.

The implementation of corpus-based language teaching (CL) in Pakistani classrooms is not an easy task. One of the issues is the technical difficulty of corpus tools. Most educators, particularly those in state-run schools, do not have a lot of education in digital pedagogy and are flustered by tools like AntConc or Sketch Engine (Rroemer, 2009). They can be deprived of the learning advantages that these tools can bring them unless they are advised appropriately. Thus, professional development workshops should be organized in a structured way in order to educate the teachers on how to interpret corpora and use the results in their daily lessons.

The next challenge is the extra time leeway to prepare lessons based on corpus. Compared to commercially available textbook units, corpus tasks require teachers to replicate, interpret, and otherwise remix data to fit classroom requirements. The preparation is not always feasible in the resource-constrained Pakistani contexts where teachers commonly have more than 50 students in a single classroom and teach several different subjects (Flowerdew, 2015). To decrease the technical load and simplify corpus work, it is possible to use pre-constructed pedagogical corpora and simple platforms like Lextutor (Cobb, 2013).

The issue of accessibility is also a problem. Although big corpora such as the COCA are available at no cost to download, sophisticated interfaces and more specialized databases can be very expensive to subscribe to - beyond the means of most Pakistani schools and universities. The rural setting is one of the most pronounced digital divides since the internet connectivity is volatile. It is important to encourage the use of open-access corpora and localized corpus projects so that every learner can be provided with equal opportunities (Boulton, 2012).

Table 1: Challenges and Proposed Solutions in the Pakistani Context

Challenge	Proposed Solutions
Technical complexity	Teacher training workshops, integration of corpus methods into B.Ed./M.Ed. programs
Time-intensive preparation	Use of ready-made pedagogical corpora and lesson templates
Limited access to resources	Promotion of free/open-access corpora, government or institutional funding for licenses

Theoretical Implications

The results of this review have been consistent with the key Second Language Acquisition (SLA) models such as constructivism and Communicative Language Teaching (CLT). Corpus-based strategies promote discovery and learning in Pakistan where learning by rote is highly prevalent. They are echoed by Vygotsky (1978) concept of the Zone of Proximal Development

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

(ZPD). Corpus tools provide scaffolding, which guides a learner through the teacher-directed exploration to independent analysis.

There is also corpus-informed learning that helps to support the 1985 Input Hypothesis of Krashen. This hypothesis brings out the importance of comprehensible input in SLA. Teachers give enriched input to Pakistani learners by introducing them to real life, diverse corpus data compared to using standard textbooks. This has a direct effect of boosting communicative competence (Canale & Swain, 1980). It provides the learners with the practical flexibility that is required in academic, professional and social set ups.

Future projections and practical applications.

The review defines some of the ways in which CL can be integrated into the language education system in Pakistan. Simple concordance exercises allow teachers to demonstrate the usage of vocabulary, collocations as well as grammatical patterns, at the school level. Students can be taught to analyses actual academic corpora in universities, where academic writing is very problematic. They will detect discourse structures, citation patterns and frequent patterns of errors using learner corpora like ICLE (Granger et al., 2015).

Institutional support is needed in order to maintain integration. Corpus based pedagogy needs to be taught in teacher training colleges and universities. Partnerships between Pakistani scholars and ELT professionals may also create localized corpora that can be used to reflect South Asian varieties of the English language. This would help learners find learning more pertinent.

Further studies in Pakistan must focus on the long term effects of CL based instruction. The most important ones are the effects on the writing fluency, spoken interaction, and pragmatic competence of learners. Another potential avenue is the application of AI-based tools of corpus. These tools make analysis simpler and less technical between the teachers and the learners.

Conclusion

This review points out the increased role played by corpus linguistics (CL) in transforming the teaching of the English language in Pakistan. Through the incorporation of real world, natural language data into the classroom, CL helps fill the gap between the textbook teaching and teaching the active use of the English language in daily communication. The methodology provides room in more meaningful, context-sensitive and effective teaching activities and assists the learners to acquire English as a tool of academic and communication.

The data demonstrate that corpus-based pedagogy has a huge potential to enhance vocabulary acquisition, grammatical fluency, and pragmatic consciousness and encourage learner independence with the help of data-driven learning (DDL). The AntConc, COCA or the British National Corpus (BNC) could help Pakistani learners who are accustomed to rote learning and the use of exams. These aids can be used to investigate frequency of words, collocations, and natural use and are able to promote discovery learning as opposed to memorization. This is in line with Communicative language Teaching (CLT) concepts where it is important that students learn to write and talk in the correct and appropriate way in various social and academic contexts.

Theoretically speaking, corpus linguistics backs up major theories in the field of second language acquisition (SLA). Corpus-informed methods are based on constructivist theories of learning (Vygotsky, 1978) and the input hypothesis of Krashen (1985), which encourages active learning and analysis. In a multilingual nation such as Pakistan in which English is used in both academic, professional, and social spheres, the authenticity and variability of corpus information enhances the communicative competence (Canale and Swain, 1980) by exposing students to a variety of registers and real life situations.

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

Meanwhile, there are a number of challenges remaining. Technical needs of corpus tools, the lack of teacher preparation and the time it takes to create classroom resources are significant impediments in implementation in Pakistan. It is also important to note that many teachers (particularly those who work in the public sector schools and colleges) have no resources or confidence to implement corpus linguistics. Combating these obstacles requires specific professional training, easy-to use and user-friendly corpus tools, and institutional reinforcement. University, teacher education, and policy-making investments can make corpus-based pedagogy a sustainable and viable component of language teaching in the country.

In the future, research and innovation has a lot of room. Future research in Pakistan must focus on the long-term implications of teaching corpus-based instruction to learner proficiency, in particular, writing, listening comprehension, and spoken fluency- these aspects are under-researched in the country. The introduction of tools based on AI driven corpus might allow content analysis to be more accessible, and may aid in the reduction of technical barrier to both teachers and learners. Formulation of localized corpora on Pakistani use of English would also increase learning experiences by availing culturally and linguistically oriented learning resources.

To sum it up, corpus linguistics is an effective, evidenced based approach to language pedagogy in Pakistan. It encourages critical analysis, linguistic competence and learner autonomy by shifting the focus away from prescriptive approaches and providing the tools that learners can use to analyze real data. Although training, access, and resource issues are expected to be considered, corpus linguistics can change the English language education in Pakistan and help close the gap between classroom teaching and the real world of communication in the modern linguistically heterogeneous society.

References

1. Biber, D., Conrad, S., & Reppen, R. (1999). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.
2. Boulton, A. (2012). Beyond concordance: Multiple affordances of corpora in university language degrees. Language Learning & Technology, 16(3), 31–44.
3. Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2), 348–393.
4. Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47.
5. Chambers, A. (2019). The impact of corpus tools on learners' lexico-grammatical competence. ReCALL Journal, 31(1), 5–19.
6. Cobb, T. (2013). Internet and literacy in the second language classroom: How the corpus revolution affects the learning of L2 vocabulary. International Journal of Corpus Linguistics, 18(1), 49–67.*
7. Davies, M. (2010). The Corpus of Contemporary American English (COCA): 1990–present. Brigham Young University.
8. Flowerdew, L. (2015). Corpus-based research and pedagogy in EAP: Current applications and future directions. Journal of English for Academic Purposes, 19, 33–47.
9. Granger, S., Gilquin, G., & Meunier, F. (2015). The Cambridge Handbook of Learner Corpus Research. Cambridge University Press.
10. Hymes, D. (1972). On communicative competence. In Sociolinguistics: Selected Readings (pp. 269–293). Penguin.

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

11. Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning? In T. Johns & P. King (Eds.), Classroom Concordance (pp. 1–13). ELR.
12. Krashen, S. (1985). The Input Hypothesis: Issues and Implications. Longman.
13. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
14. O'Keeffe, A., McCarthy, M., & Carter, R. (2007). From Corpus to Classroom: Language Use and Language Teaching. Cambridge University Press.
15. Ortikov, U. (2023). Practical uses of corpus analysis in designing language-teaching materials. Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 3(7), 304-309.
16. Ortikov, U. K. U. (2024). The effectiveness of technology-enhanced language learning methods. Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 4(3), 162-179.
17. Pérez-Paredes, P. (2019). The pedagogic architecture of data-driven learning. Language Learning Journal, 47(4), 404–417.
18. Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. Annual Review of Cognitive Linguistics, 7(1), 140–162.
19. Thornbury, S. (2001). Uncovering Grammar. Macmillan.