ACADEMIA International Journal for Social Sciences
Volume 4, Issue 4, 2025 ISSN-L (Online): 3006-6638

Using Ai to Analyze Language Learners' Discourse: A Corpus-Based Study of Learner Language Development

Arfa Maham
arfa@gscwu.edu.pk
Department of Computer Science & Information Technology, Government Sadiq College Women, University Bahawalpur, 63100, Pakistan.

Muniba Saleem
muniba@gscwu.edu.pk
Department of Computer Science & Information Technology, Government Sadiq College Women, University Bahawalpur, 63100, Pakistan.

Muhammad Ismail Rahu
ismail.rahu@quest.edu.pk
Lecturer at the Department of English, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah.

Sohail Ahmad
ahmad.sohail664@gmail.com
MPhil. in English Linguistics. SSE English School Education Department (SED), Govt. of Punjab, Pakistan
https://orcid.org/0000-0001-8710-3237.
Corresponding Author: * Sohail Ahmad ahmad.sohail664@gmail.com

Received: 21-08-2025 **Revised:** 26-09-2025 **Accepted:** 16-10-2025 **Published:** 25-11-2025

ABSTRACT

This research examined patterns of language development of Pakistani university students using AI-driven corpus tools. The researchers gathered 150 students' written and spoken discourse samples from Punjab and Sindh provinces for six months and built a small learner corpus of around 500,000 words. The researchers used natural language processing and machine learning tools to evaluate the samples for lexical variability, grammatical precision, syntactic intricacy, and coherence. The researchers used a mixed-method design for the study, incorporating quantitative frequency analysis and qualitative thematic analysis. The analysis demonstrated advanced and less advanced learners' levels of proficiency and lexical sophistication and syntactic complexity to a higher degree. The researchers found patterns of common errors, which included articles, propositions, and subject-verb agreement. The AI managed to recognize the gaps of interlanguage and track the development level. Qualitative analysis produced five themes: L1 transfer, rule overgeneralization, lexical discourse fossilization, organization problems, and code-switching. The study demonstrated the corpus-based approach's ability to detect the language learner's level of development. This research helped understand the second language acquisition processes of South Asian learners and illustrated how AI technology can benefit learning and teaching research conducted to improve language education.

Keywords: *Patterns, language development, spoken discourse, natural language processing, machine learning tools, lexical variability, grammatical precision, syntactic intricacy, coherence.*

INTRODUCTION

Artificial intelligence technologies continue to evolve, impacting many facets of educational research, especially applied linguistics and second language acquisition (Huang et al., 2023). In Pakistan, where students have traversed multiple educational epochs in English, learning the language and understanding the various linguistic expressions graduate students use has always been a concern for educators and policymakers. Pakistani learners often perform English competence at a sub-certificate level, which could lead to a decrease in their employability and overall productivity. Traditional approaches to learner language are often too labor-intensive and overly superficial to detect valuable, systematic relationships across a wide array of data. In the field of corpus linguistics, coupled with computer-aided learning, there are numerous possibilities for scaled approaches to learner language, offering deep insights into what students at a given educational level have been able to articulate and what has been lacking (Wang et al., 2023).

The field of language studies witnessed an increase in the use of corpus-based research methodologies for its systematic and objective data processing of large volumes of language data (Ahmad et al., 2021). Other than intuition-based methodologies, corpus approaches garnered research from actual instances of language use, exhibiting true patterns of language usage in contrast to prescriptive approaches. This methodology was particularly useful in learner language studies, as it chronicled actual processes and patterns of learner language development, errors and interlanguage. The universities of Pakistan and their students from the diverse provinces of Punjab and Sindh offered a great context for the study of specific regional variations and universal developmental patterns. The advanced computational methodologies in the study of Pakistani learners' discourse in the corpus are virtually non-existent, and it is this particular gap in the literature that the present study seeks to address (Mahlobogoane, 2024).

The ability to incorporate AI within the context of language study has given researchers the capacity to uncover incredibly nuanced language features. This is something no human analyst is capable of. It is true that human analysts cannot wrap their heads around the sort of analytical complexity that new machine learning algorithms are able to achieve (Kumar et al., 2024). Modern machine learning algorithms have the ability to cluster several linguistic features and their correlations at once, considering relationships between different lexical choices, grammar and the organization of discourse. Coupled with tools of natural language processing, researchers can have enormous data sets automatically coded, and thus save a countless number of hours from coding the data manually, while also providing a higher level of consistency. It is no wonder that the advances of AI have been embraced with open arms by researchers in Pakistan. For them, AI has made possible the analysis of data sets which would have required enormous teams of trained human coders. The introduction of AI into the analysis of language corpora has been alongside a pioneering approach to methodology and the complex phenomena of language itself (Andleeb et al., 2025).

The present study sought to document Pakistani university students' developing English language proficiency, a phenomenon that, to our knowledge, remains undocumented in the literature. Collecting written discourse from 150 undergraduate students in two of the country's most populous provinces represented a first attempt to capture learner language in the context of higher education in Pakistan. With a six-month longitudinal design, the study demonstrated not only the learners' proficiency level, but the changes in their proficiency over time, thereby capturing language acquisition in real time. Collecting both written and spoken discourse enabled the researcher to report on the learners' overall level of communicative competence. In addition, the study focused on specific academic genres to address the students' real-world communicative needs, thereby enhancing the relevance of the study to pedagogy. The study provided a rich description of the learners' proficiency and their challenges in language use, particularly in relation to the quantity and quality of their written and spoken language through an in-depth analysis of their lexicon, grammar, syntax, and discourse. This study, therefore, provided all stakeholders with a means to build both knowledge and pedagogy applicable to the English language education system in Pakistani universities and contributed to second language acquisition, corpus linguistics and educational technology.

Research Objectives

1. To use corpus analysis powered by AI tools to establish the patterns, errors, and developmental features of discourse generated by university students in Pakistan.
2. To analyze the disparity as a result of biennially studies conducted on the participants of the provinces of Punjab and Sindh on the variables of lexical diversity, grammatical accuracy, syntactic complexity, and discourse coherence.
3. To evaluate the effectiveness of AI and NLP tools in analyzing learner corpora to longitudinally track language development and identify interlanguage features.

Research Questions

1. What are the primary linguistic, error, and developmental patterns that are present within the written and spoken discourse of Pakistani university students?
2. What is the relationship between lexical diversity, grammatical accuracy, and syntactic complexity and discourse coherence for students of differing proficiency levels in the provinces of Punjab and Sindh?
3. What is the usefulness of tools augmented with artificial intelligence, and natural language processing, for the temporal analysis of language development for learners within corpus-based studies?

Significance of The Study

This research has significant implications for various stakeholders within the higher educational institutions in Pakistan. This study provided learners' language attributes which assist in the development of educational curriculums and the design of teaching pedagogies at the university levels. The findings encapsulated the recognition of learners' error patterns and developmental levels which would enable educators to formulate instructional plans that address the learners' needs rather than the educators' perceived needs. The cost-effective AI of language analysis would aid institutions with low research capacity to incorporate evidence-based practices into language education. The research provided insights on language learning outcome disparities which can assist policymakers in developing standards for the assessment and teaching of language education. Also, this study further developed the global understanding of South Asian contexts of multilingualism and the acquisition of additional languages. The study's methodological framework established a model for further similar investigations, which will aid in the advancement of the educational technology and learner corpus research fields, especially in other developing countries.

LITERATURE REVIEW

Corpus linguistics developed as a new approach and methodological framework in the study of language and altered the manner in which researchers studied language. The corpus-based study of language involved the analysis of a large and systematic collection of texts and provided authentic and empirical evidence for a particular claim about language. In the study of second language acquisition, learner corpora offered a means of investigating interlanguage, errors, and the order of acquisition. The first studies of learner corpora were about the written texts produced by learners of European languages and the framework analyzed in the studies developed several theories of language acquisition. However, the focus of the fields of corpus studies and learner corpus research has moved, trying to account for the varying L1 transfer and cultural contexts of language acquisition of other learners. It has also been of primary interest within the South Asian region as this region has a large population of learners of English (Meyer, 2023).

Rather than the traditional manual coding techniques used in the earlier stages of the technology's development, the applications of artificial intelligence in the analysis of language employed machine learning algorithms on pattern recognition, classification, and prediction. The fields of natural language processing and machine learning worked in concert to annotate morphological, syntactic, and semantic traits of a corpus, and in doing so increased the scale of analysis and corpus construction and the speed. The tools and algorithms of machine learning and natural language processing were used to identify and annotate the instances of error in a sample, and were used to measure complexity, and also to analyze and measure components of discourse (MISNAWATI et al., 2024). The use of artificial intelligence in the processing and analysis of learner language datasets was also a notable success, allowing for the division of the corpus and the analysis of the sub-components in a way that automated tools could measure patterning across multiple variables, freestyle coding of the corpus by a human analyst on the unit of discourse at the micro or even the sub levels. The disparate uses of natural language processing tools in automated analysis of learner language were also the subject of caution by the authors, and drew attention to potential non-encompassing analysis and automation (Martins, 2026).

Research on Pakistani English Learners continues to reveal persistent linguistic issues. Article usage was erroneous and pervasive. Researchers attributed this to the umlauted presence of article systems in the Urdu and other regional languages of the area. Repositional semantics (the area of English's prepositions and their meaning) was also frequently mentioned in relation to L1 interference errors (ones in which English proficiency is absent, and the learner inserts their first language's structures) (Saeed et al., 2023). Errors in subject-verb agreements also occurred in advanced learners (a grammatical structure pertaining to whether the subject and the verb agree in singularity or plurality), which suggests the fossilization of flawed systems of the grammatical pattern). There is an interdependence between the grammatical issues and the limited lexicon on the parts of the learners. There were limited word and collocational usage on the parts of the learners. There was interdependence between productive and receptive disabilities that was repeatedly mentioned. Pakistani learners often had the ability to comprehend complex texts and the simplified language that was by them was produced in their own writing and speaking. How examination orientated instruction was emphasis on rote learning and ineffective communicative competence was critiqued in the Pakistani schooling systems (Abbas et al., 2025).

Thanks to work from Developmental Psychologists conducted longitudinal studies documenting the process of language development from infancy through early childhood, we understand the sequences involved in the acquisition of language and the time periods involved in the different processes in development. It was shown that different parts of language were mastered at different rates, and that different tiers of proficiency could be mastered. Learners also tended to have a significant break in usage to intervene in their development at the lower proficiencies of the language. Although learners intermediately stopped, they tended to go at a more consistent, steady rate when testing their proficiency. More erratic patterns of development were shown for the learners with development through more complex grammatical structures. In working with more advanced grammatical forms of language, the students learned to speak in more simplified formats. It needed to be emphasized that there is a minimum of linear, erratic patterns in development. These patterns and observations also needed to be studied for much longer periods of time (Volling & Cabrera, 2025).

Interest in how learners' language displays a grasp of cohesion, coherence, and overall structure (organizational and topical arrangements) has grown as researchers come to appreciate that a focus only on sentence-level accuracy fails to account for learners' ability to communicate meaningfully and appropriately (Petersen et al., 2024). Documented studies investigated how learners' productions displayed performances around cohesion, coherence, as well as overall structure (topical and organizational arrangements) including academic discourse. Particularly in the case of Pakistani learners, the difficulties espoused in academic writing (e.g., thesis statements, paragraphs, sequencing, and logical flow) demonstrated limited syntactic knowledge and/or a failure to grasp the principles of academic discourse as it is understood in the Western paradigm. Instructed discourse organization, in the views of some, prompted a significant change in the quality of learners' writings because they understood that the development of rhetorical competence was a primary goal that required focused instruction rather than the interventional approach some might consider to be writing as a skill to be developed through exposure (Haidar & Manan, 2021).

RESEARCH METHODOLOGY

This research utilized corpus linguistics methods to examine language development of English language learners in Pakistani universities. The researchers gathered written and spoken discourses of 150 undergraduate students at public and private universities in the Punjab and Sindh provinces. Over a six-month period, subjects created academic essays, gave oral presentations which were audio recorded, and completed standardized assessments of their language proficiency. The researchers assembled these academic and oral language samples (from the presentations) into a specialized learner corpus of 500,000 words. They, then, employed natural language processing (NLP) tools and AntConc, a concordance program, and machine learning algorithms in Python, to detect and analyze language development patterns, language processing errors, and other features of language development. They intentionally selected and analyzed features of operational language development in the learners' corpora: vocabulary, accuracy, and other grammatical features, language complexity and syntactic structures, and the overall coherence and fluency of the discourses. Following the selection of European standards, the researchers annotated the corpora with accuracy, error type, and complexity measures to described the words and language structures of various parts of speech (POS). The researchers used quantitative and qualitative methods to analyze the data to identify patterns from the annotated corpora. The AI-based NLP tools provided researchers with tools to identify systematic features of second language learners' language acquisition processes (interlanguage) and measures of the rate of language development during the data collection period. The researchers applied interrater reliability measures to confirm the accuracy of their annotated data in the results when 20% of the annotated data were independently reviewed by two other researchers. The mixed methods design granted obtained observations and understanding about the university students in Pakistan and their experience of learning the language in the two provinces.

RESULTS AND DATA ANALYSIS

Quantitative Analysis

Table 1: Lexical Diversity Measures Across Proficiency Levels

Proficiency Level	Type-Token Ratio	Lexical Sophistication Index	Academic Vocabulary %
Low (n=45)	0.42	2.8	12.4%
Intermediate (n=68)	0.56	3.9	18.7%
Advanced (n=37)	0.71	5.2	26.3%

The lexical diversity analysis revealed substantial differences across proficiency levels, with advanced learners demonstrating significantly higher type-token ratios compared to their low-proficiency counterparts. The lexical sophistication index, measuring the use of low-frequency vocabulary items, showed progressive increases from low to advanced groups, indicating that proficient learners accessed broader vocabulary ranges. Academic vocabulary usage also demonstrated clear stratification, with advanced learners incorporating more than double the proportion of academic terms compared to low-proficiency students. These findings confirmed that lexical development represented a critical dimension of overall language proficiency, with vocabulary knowledge distinguishing learners across competency levels.

Table 2: Grammatical Accuracy by Error Categories

Error Category	Low Level (%)	Intermediate (%)	Advanced (%)	Overall Frequency
Article Usage	28.4	16.2	8.1	3,847 errors
Preposition	24.7	18.9	11.3	3,521 errors
Subject-Verb Agreement	19.3	12.4	6.7	2,198 errors
Verb Tense	15.8	10.7	5.9	1,876 errors
Word Order	11.8	7.8	3.2	1,354 errors

Grammatical error analysis identified five predominant error categories across the corpus, with article usage emerging as the most frequent problem across all proficiency levels. The data demonstrated clear improvement patterns as learners advanced through proficiency stages, though even advanced students exhibited persistent article and preposition errors. Subject-verb agreement errors showed the most dramatic reduction from low to advanced levels, suggesting this feature responded well to instruction and practice. Verb tense errors decreased steadily across groups but remained present even among advanced learners, particularly in complex narrative contexts. Word order violations proved least common overall and showed the sharpest decline with increased proficiency, indicating that syntactic sequencing rules were acquired relatively early in the developmental process.

Table 3: Syntactic Complexity Measures

Measure	Low Level	Intermediate	Advanced	Native Baseline
Mean Length of T-unit	8.3	11.7	15.2	17.8
Clauses per T-unit	1.2	1.6	2.1	2.4
Dependent Clauses %	18.4	31.2	47.6	54.3
Complex Nominals per Clause	0.7	1.1	1.8	2.2

Syntactic complexity analysis demonstrated progressive sophistication across proficiency levels, approaching but not reaching native speaker baselines even among advanced learners. The mean length of T-units increased substantially from low to advanced groups, reflecting the ability to produce longer, more elaborate grammatical structures. Clause complexity showed similar developmental patterns, with advanced learners employing significantly more clauses per T-unit than lower-proficiency students. The proportion of dependent clauses revealed particularly striking differences, as advanced learners used dependent clauses more than twice as frequently as low-proficiency learners. Complex nominal structures also increased with proficiency, though the gap between advanced learners and native speakers remained considerable, suggesting this feature developed later in the acquisitional sequence.

Table 4: Discourse Coherence Indicators

Indicator	Low Level	Intermediate	Advanced

Cohesive Devices per 100 words	3.2	5.8	8.4
Topic Consistency Score	2.1/5	3.4/5	4.2/5
Logical Connectors Usage	1.8	3.6	5.9
Paragraph Organization Score	2.3/5	3.6/5	4.4/5

Discourse-level analysis revealed systematic differences in organizational and coherence features across proficiency groups. Advanced learners employed nearly three times as many cohesive devices per hundred words compared to low-proficiency students, demonstrating greater attention to textual connectivity. Topic consistency scores, measuring the maintenance of central themes throughout texts, showed steady improvement with proficiency advancement, though even advanced learners scored below optimal levels. Logical connector usage displayed similar patterns, with advanced students utilizing substantially more transitional expressions to signal relationships between ideas. Paragraph organization scores indicated that structural competence developed alongside linguistic proficiency, though this remained an area where even advanced learners showed room for improvement.

Table 5: Provincial Comparison of Language Features

Feature	**Punjab (n=87)**	**Sindh (n=63)**	**Statistical Significance**
Overall Accuracy Rate	76.4%	73.8%	Not significant
Lexical Diversity (TTR)	0.58	0.56	Not significant
Syntactic Complexity	12.3	11.8	Not significant
L1 Transfer Errors %	31.2%	34.7%	Marginally significant

Cross-provincial analysis revealed minimal differences in most language measures between Punjab and Sindh students, suggesting that regional variations had limited impact on overall proficiency development. Accuracy rates, lexical diversity, and syntactic complexity showed statistically insignificant differences between the two provinces, indicating comparable learning outcomes despite potential variations in educational contexts. However, L1 transfer errors demonstrated marginally significant differences, with Sindh students showing slightly higher rates of mother tongue interference. This finding suggested that L1 backgrounds, which varied more within Sindh due to greater linguistic diversity, influenced specific error patterns even when overall proficiency levels remained similar across provinces.

QUALITATIVE ANALYSIS

Theme 1: L1 Transfer Interference

The influence of the first languages of learners, especially Urdu, Sindhi, and Punjabi, were perceptible, involving a semantic, lexical, and grammatical analysis of the English language. Obvious translation strategies, contrary to the accepted English semantic and syntactic structure, resulted in a phrase or sentence that was not idiomatic. Language transfer, in negative phonological discourse, where the first language's (L1) patterns were inappropriately patterned to English and the learners spoke in English were phonologically were particularly evident. Grammatical transfer was evident in the word order of English questions, the placement of adjectives and clauses, and the omissions of adjuncts. While these transfer effects were consistent, the more advanced learners were more self-aware, self-correcting and monitored the transfer syntactic structures to English more than lower-level learners.

Theme 2: Overgeneralization of Grammatical Rules

In a number of situations, learners did not observe the appropriate context of grammatical rules, which resulted in a number of systematic errors that were reflective of the incomplete mastery or the incomplete disengagement of a grammatical rule. In English, there were regular past tense markers inappropriately used on verbs and plurals that were uncountable. Overgeneralization of article usage by using the definite article for proper nouns was a norm. A number of learners displayed a lack of preposition usage by using single prepositional patterns that did not align in context with other required structures in a sentence. These errors, while novice, were typical of learners even at advanced levels. Particularly, this was evident when cumulative cognitive pressure tended to create situations where spontaneous production of English language was required.

Theme 3: Lexical Fossilization Patterns

Certain types of fossilization were indicated through the corpus prolific occurrence of the same type of lexical error wherein the learners misused words, and repeatedly did so, regardless of the exposure of the learners to the correct usage, and the learners' misuse of words was actually a class of error for a set of words that are commonly paired, e.g., affect/effect, accept/except, and principle/principal. The influence of false cognates from Urdu resulted in the systematic misuse of several words that are similar in their phonological composition that were also of diverse meanings. Collocation errors were also clearly the result of an L1 pattern of collocation in learners' native Urdu, and were clearly the result of the absence of a conjunction. The learners' lexical repertoire was also evidently very limited, which resulted in an overuse of the same words in their writing, and through the high frequency of words they employed, their writing was actually imprecise and their errors were also lexical gaps that educational interventions were not able to alleviate.

Theme 4: Discourse Organization Challenges

At the macro level, the learners' organization of their text clearly indicated the overarching problem. The learners were doing essays, and the learners' organization of their essays was very problematic along the lines of there being no clear thesis statements, nor did the learners' organization exhibit a logical flow of their ideas or a clear conclusion or effective conclusion was there. For particular paragraphs, there was a weak development along the lines of the grouping of ideas, along with the learners' weak and vague presentation of ideas which was in the form of presentation of a single sentence that was paragraph-sized, or they did not adequately develop the topic sentence. The organization of the essays was along a pattern that came out as being in a sequence that lacked coherence, which made it difficult for the readers to follow along as the readers were lost along the sequence of ideas that were not interrelated or connected. There also seemed to be an influence of culture as reflected in the patterns of argumentation that seemed to be to not explicitly take a stance in the argument, and there was an overuse of hedging that was excessive, along the statements that were made. This absence of organization of discourse was also evident at a higher level than the sentence level, which indicates that the development of rhetorical skills which were at a very basic level needed more than the mere surface level of linguistic proficiency.

Theme 5: Code-Switching Phenomena

The practice of code-switching, that is, the ability to use multiple languages in a communicative act with ease, was noted in all of the use of speech during the sampling, between English and the mother tongue. Some students brought into their speech token and expressions in Urdu and/or other regional expressions of the Pakistani students, and/or other expressions, especially, in very culturally unique and/or deeply emotional terms when in English and/or other expressions that were equivalent. Intra-sentential switching occurred at syntactic boundaries, following consistent grammatical constraints and not at random. Code-switching frequency correlated inversely with proficiency, as advanced learners relied less on L1 support during English production. Of Code-switching persisted even informal and of other contexts among even less proficient speakers of English, which suggests a bilingual communication strategy rather than merely a deficiency marker, code-switching of other languages was even evidenced.

Theme 6: Strategic Competence Development

The use of other compensatory strategy suggested the presence of a developing type of strategic competence in the students in the presence of other types of communicative or other speech or even linguistic constraints. The use of circumlocution to enable other students to communicate or even express in other terms of meaning without essential vocabulary expressing, however, even at times this led to the overall being imprecise and/or vague expressions of communication, meaning in a communication. When even particular words of vocabulary use was not even able to communicate. Especially the self-monitoring and the other observable repair of other sequences in advanced learners of production were noted and this even as of the overall language awareness, metalinguistic awareness of the learners of the other advanced level. The request or the appeal for interruption of a communication of other speech in the use of speech even in the mere form of question and of other expressions in request for uttering and additional terms clarification of other speech.

DISCUSSION

The findings confirmed highly multifaceted developmental trajectories with respect to language learning among Pakistani university students due to the intricacy and involvement of cross-cutting dimensions as opposed to simple and linear growth. The quantitative analyses confirmed and justified the effectiveness of the measures as developmental indicators due to the clear stratification of the various levels of one’s proficiency with respect to the various dimensions of lexical diversity, grammatical accuracy, syntactic complexity, and discourse coherence. The presence of certain error types even among the advanced levels of proficiency suggested that those linguistic features might have resisted acquisition despite the years of instruction and exposure that those learners had due to the risk of fossilization. The presence of limited variation among the provinces deviated from the initial expectation of the impact of the region on learning and achievement among students due to the suggestion that the specific geography was of secondary importance and that the greater contribution was from institutional variables and individual differences. The themes, appropriated from qualitative research, provided explanation for the patterns and distribution of the results from the quantitative methods and confirmed the presence of transfer effects, overgeneralization, as well as strategic compensations that together acted in the formation of the linguistic outcomes. The incorporation of AI-powered techniques to identify the emergent patterns validated the use of corpus-based research, but also demonstrated the significance of human interpretative skills for contextual understanding.

CONCLUSION

This study demonstrated the effectiveness of integrating AI technologies and corpus-based methodology in exploring learner language development in the context of higher education in Pakistan. Using IL cross-measure corpus research methodology and multiple dimensions of IL proficiency, the study painted a complete and more intricate picture of the English language IL of the Pakistani learners than any other study on the same subject, thereby confirming the study's objectives. Using a metaphor, Pakistani learners seemed to complete the English language proficiency developmental cycle as learners of other countries, while reflecting some distinct differences. The differences stemmed primarily and predominantly from the learners' multilingual education environment and educational experience. The gaps in proficiency levels illustrated the differences in the target population and the necessity for tailored frameworks to meet the varying levels of proficiency. The research study demonstrated the power and usefulness of automated tool technologies to analyze and interpret research data in large volumes; and at the same time illustrated the unchangeable importance of human expertise in data analysis to the qualitative level. The study has provided educational policy makers in higher education in Pakistan with the much-needed empirical data, and at the same time advanced the educational research methodology in IL corpus to be applied in other similar contexts.

RECOMMENDATIONS

Educational institutions ought to adopt the incorporation of corpus-informed curricula, especially with respect to error patterns and developmental issues that this study addressed, and especially concerning the recurring issues with articles, prepositions, and discourse organization. Teacher education needs to include corpus-informed teaching and data-driven methods of teaching so that teachers are able to make evidence-informed decisions for instruction, rather than relying on intuition. Universities ought to maintain learner corpus databases that track the language of students over time, thereby creating resources for the institution for curriculum and pedagogical improvement, as well as for the institution to support research. Policymakers ought to integrate competences in discourse as being equal to grammatical competences in teaching and assessment policies, and to review policies relevant to assessment to provide more holistic views of communicative competence. Further studies need to apply this same methodology to tracking individual students over time to determine the effectiveness of corpus-informed interventions and determine the role of explicit instruction in removing features of language that have simply fossilized over time.

REFERENCES

Ahmad, S., Akram, M., & Ali, M. (2021). The Impact OfEsl Teachers' Use of Motivational Language On Students Learning: A Study Of Esl Learners At Elementary Level In Pakistan. Palarch's Journal of Archaeology of Egypt/Egyptology, 18(10), 1187-1196.

Abbas, S. G., Mahjabeen, A., Akbar, G., & Lodhi, K. (2025). English Education and Social Development In Pakistan: Investigating Motivation, Challenges, And Future Prospects Among Students. Annual Methodological Archive Research Review, 3(3), 22-33.

Andleeb, D. N., Fatima, D. N., Ali, A., KHAN, D. T., Tanvir, M. M., Iqbal, D. S., Mahmood, W., Ashfaq, D. M. S., & Ahmad, S. (2025). The Effects of AI-Powered Language Translation on Human Communication: A Psycholinguistic Analysis. TPM–Testing, Psychometrics, Methodology in Applied Psychology, 32(S4 (2025): Posted 17 July), 1671-1682.

Haidar, S., & Manan, S. A. (2021). English in Pakistan: Language policy, features and present-day use. In English in East and South Asia (pp. 242-255). Routledge.

Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. Educational Technology & Society, 26(1), 112-131.

Kumar, D., Ahmad, S., & Lodhi, K. (2024). Exploring the Role of Digital Technology in Enhancing Learning Experiences in Pakistani Classrooms. International Journal of Language and Literary Studies, 8(2), 380-393.

Mahlobogoane, M. G. (2024). The representation and distribution of conjunctions in selected Sepedi home language textbooks: a corpus-based Investigation University of Pretoria].

Martins, H. F. (2026). The Role of Artificial Intelligence in Linguistic Corpus Analysis. In Harnessing AI for Multigenerational English Language Learning (pp. 193-232). IGI Global Scientific Publishing.

Meyer, C. F. (2023). English corpus linguistics: An introduction. Cambridge University Press.

MISNAWATI, M., Sahril, N., & TAHIR, S. Z. B. (2024). Corpus linguistics today: A qualitative approach. Research and Innovation in Applied Linguistics, 2(1), 45-62.

Petersen, I. T., Apfelbaum, K. S., & McMurray, B. (2024). Adapting open science and pre-registration to longitudinal research. Infant and child development, 33(1), e2315.

Saeed, F., Rashid, A., & Rasheed, A. (2023). Navigating the Linguistic Divide: Challenges Faced by Pakistani Students in American and British English. Inception-Journal of Languages & Literature, 3(2), 116-128.

Volling, B. L., & Cabrera, N. J. (2025). Loving, Laughing, and Learning: How Father–Child Relationships Contribute to Children's Development in Early Childhood. Annual Review of Developmental Psychology, 7.

Wang, S., Zhang, H., & Sardar, N. (2023). English Teaching Methods in Chinese and Pakistani Educational Institutes: A Comparative Review. Qlantic Journal of Social Sciences, 4(4), 332-345.