



A Corpus-Based Study of Indian English Usage in Academic Writing

Anjali Shah

Ph.D. Scholar (English)

&

Dr. Himadri N. Raval

Assistant Professor (English)

Monark University, Ahmedabad, Gujarat

Abstract

This study investigates distinctive lexico-grammatical and discourse features of Indian English (IndE) in academic writing using a corpus-based approach. The research examines patterns in lexical choice, nominalization, passive use, hedging, reporting verbs and citation practices and compares these with reference international academic corpora. Results indicate that Indian academic writers show higher frequencies of Indian lexical insertions, extended noun phrases and conservative hedging devices, while passive constructions are used strategically in methods and results reporting. Findings also point to discipline-specific variation: humanities texts show more narrative and evaluative lexis, while sciences rely more on dense nominal groups and technical collocations. Pedagogically, the study identifies recurrent learner/writer difficulties collocation mismatch, article and preposition use and register drift which have implications for academic writing instruction, corpus-informed syllabus design and automated feedback tools. Methodologically, the research combines frequency-based quantitative measures (normalized frequencies, keyness, n-grams and collocate statistics) with qualitative discourse analysis of representative concordance lines to interpret pragmatic functions. The study contributes an empirically grounded description of IndE academic register, provides reference datasets for future comparative work and recommends corpus-driven materials for research scholars and EAP instructors in India. The paper argues for a balanced view that recognizes IndE as a legitimate variety while highlighting areas where targeted instruction can improve international intelligibility and publication outcomes.

Keywords: Indian English, Academic Writing, Corpus Linguistics, Lexico-Grammar, Hedging, Nominalization

Introduction

The globalization of higher education and the increasing pressure on Indian scholars to publish in international journals have stimulated research into the characteristics of Indian English (IndE) as it appears in academic writing. Understanding the lexico-grammatical and discursal tendencies of IndE academic register is crucial for both descriptive variety studies and practical interventions in English for Academic Purposes (EAP). Corpus linguistics by enabling large-scale, data-driven description has been widely adopted in studies of academic English internationally, but comparatively fewer studies have focused specifically on Indian academic texts spanning multiple disciplines within the recent decade. (Ackermann et al., 2018; Barua, 2022).



Previous research on IndE shows that it exhibits systematic lexical, morpho-syntactic and discourse patterns that distinguish it from other World English's: localized lexical items, extended noun phrases, differences in article and preposition use and particular pragmatic strategies for stance and evaluation (Kolhapur Corpus studies; Le, 2019). These patterns are consequential in academic contexts where register expectations (conciseness, nominalization, hedging) affect peer review and readership comprehension (Kolhapur Corpus; Le, 2019).

Corpus-based approaches to academic writing have two complementary strengths. Quantitatively, they reveal frequency distributions, collocation patterns and keywords that mark register (Biberian multidimensional analyses and n-gram profiling). Qualitatively, concordance-based reading allows researchers to interpret how forms enact rhetorical functions such as stance, evidentiality and politeness (Ackermann et al., 2018). Applying both lenses to IndE academic writing enables an empirically rich account of how Indian writers realize academic norms and where they systematically diverge.

This paper reports a corpus-based study of Indian academic writing with three interrelated aims:

1. To identify recurrent lexico-grammatical features characteristic of IndE in research-article sections and postgraduate dissertations.
 2. To compare these features with international academic reference corpora to locate divergences and convergences
 3. To discuss pedagogical and editorial implications for EAP instruction and publication support.
- Methodologically, the study constructs a balanced IndE academic corpus, performs frequency and collocation analyses and then interprets salient patterns through discourse-functional analysis.

A key motivation is pragmatic: while IndE is increasingly recognized as a legitimate World English variety, Indian authors often face challenges when communicating to global audiences accustomed to particular academic norms (e.g., tight nominalization, compact methodological reporting). By describing corpus-evidenced tendencies and connecting them to communicative effects, this study aims to inform targeted teaching (lexico-grammatical pattern instruction, corpus-based writing materials) and support services (editing, feedback systems) that can enhance clarity and acceptance in international publication venues. The next section reviews relevant research, focusing on corpus constructions, genre-specific features and pedagogical implications for academic writing in Indian contexts.

Review of Literature

Corpus-based studies of academic English have proliferated globally, producing detailed descriptions of register features (nominalization, hedging, passive voice, reporting verbs) and informing pedagogical interventions (Ackermann et al., 2018; Pearson PICAIE project). PICAIE and related international corpora have been instrumental for comparative analyses, offering benchmarks of academic lexis and phraseology across varieties (Ackermann et al., 2018). These projects demonstrate the value of large, genre-balanced corpora to identify normative patterns in academic prose.

Within the Indian context, scholarship over the past decade has combined corpus construction with applied research. Shodhganga theses and institutional studies (e.g., Bhargavi, 2021; Barua, 2022) have used smaller corpora or learner/dissertation collections to analyze grammatical errors, article/preposition use and genre competence among Indian scholars, reporting systematic tendencies such as article omission,



collocation mismatches and over-explicit stance marking (Bhargavi, 2021; Barua, 2022). These findings align with larger IndE descriptions that note morpho-syntactic nativization and lexical assimilation (Kolhapur Corpus work).

Recent corpus-informed studies specifically targeting academic writing have produced insights into lexical density and teaching applications. Mukhopadhyay (2021) used digital lexical tools to estimate lexical profiles of learner academic texts and advocated corpus-driven feedback to scaffold lexical development in adult ESL academic writers. Such studies emphasize practical classroom implications: data-driven materials can make invisible frequency patterns (e.g., favoured reporting verbs, common collocate) explicit for writers and tutors.

Comparative work on academic grammar highlights cross-varietal trajectories: for example, large-scale analyses show evolution in passive voice use, hedging strategies and nominalization across disciplines and over time (Le, 2019). These studies suggest that disciplinary conventions often override national variety differences, but national features remain salient in lexis and certain grammatical choices. In India, discipline-specific variation has been observed humanities and social sciences often display more evaluative language and narrative devices, while STEM texts favour dense nominal groups and formulaic reporting (Le, 2019; PICAIE comparisons).

Methodological reviews of corpus-based academic-writing research identify two recurring needs: more balanced, domain-specific corpora for under-represented varieties and mixed-method analyses that pair quantitative findings with discourse-functional readings. The literature thus motivates the present study's design: constructing a multi-disciplinary IndE academic corpus and combining keyness/collocate metrics with concordance-driven interpretations to account for both form and function.

Pedagogical implications emerge consistently: corpus-informed pedagogy teacher awareness, reference materials and automated feedback can reduce recurrent register problems and improve publication-readiness for non-native authors (Jam al, 2021; Mukhopadhyay, 2021). Indian studies additionally highlight the need for institutional writing support tailored to local research cultures and publication expectations (Barua, 2022). These converging findings set the empirical and applied agenda for the present corpus-based inquiry.

The present study is grounded in an interdisciplinary theoretical framework that draws primarily on Corpus Linguistics, World Englishes theory and English for Academic Purposes (EAP), with supplementary insights from Systemic Functional Linguistics (SFL) and genre theory. Together, these perspectives provide a robust lens for examining the structural, functional and contextual dimensions of Indian English (IndE) as it is realized in academic writing.

Corpus Linguistics as a Theoretical Foundation

Corpus Linguistics (CL) provides the principal methodological and theoretical foundation for this study. At its core, CL views language as a probabilistic system, where patterns of use emerge from frequency, distribution and co-textual association rather than from introspective intuition alone (Biber et al., 2021). In academic writing research, corpus linguistics has enabled scholars to uncover recurrent lexico-grammatical patterns such as nominalization, stance markers and phraseological bundles that characterize disciplinary and register-specific discourse.



A central theoretical assumption of CL is that repeated linguistic behaviour reflects underlying norms. From this perspective, the frequent occurrence of particular forms in Indian academic texts (e.g., extended noun phrases or specific reporting verbs) is not incidental but indicative of stabilized conventions within the IndE academic register. This aligns with usage-based theories of language, which argue that linguistic competence emerges from exposure to patterned input over time (Ellis, 2016).

Corpus linguistics distinguishes between corpus-based and corpus-driven approaches. While corpus-based studies test predefined hypotheses, corpus-driven research allows patterns to emerge inductively from the data. The present study adopts a hybrid position, using established categories from prior research (e.g., hedging, passivation) while remaining open to emergent features specific to Indian academic writing. This approach is consistent with recent academic writing research that advocates combining quantitative evidence with qualitative interpretation (Ackermann et al., 2018).

World English's and Indian English

The study is also theoretically informed by World Englishes (WE), particularly Kachru's (1992) model of Inner, Outer and Expanding Circles. Indian English occupies a central position in the Outer Circle, where English has institutionalized functions in education, administration and academic knowledge production. From a WE perspective, IndE is not a deficient approximation of British or American English but a legitimate, rule-governed variety shaped by local sociocultural and educational contexts. This theoretical stance is crucial for interpreting corpus findings. Features such as article variation, prepositional preferences or localized collocations should not be automatically labeled as "errors." Instead, they can be understood as nativized norms, especially when they appear consistently across large datasets. As Schneider (2014) argues, mature postcolonial varieties develop their own endonormative standards, including in formal written registers.

World English's theory also recognizes the tension between local legitimacy and global intelligibility, particularly in academic publishing. While IndE academic writing reflects localized conventions, international journals often implicitly privilege Inner Circle norms. The present study situates itself within this tension, seeking not to delegitimize IndE features but to identify those patterns that may affect cross-cultural academic communication.

English for Academic Purposes (EAP)

The framework of English for Academic Purposes provides the pedagogical and applied dimension of this study. EAP views academic writing as a socially situated practice, where language choices are shaped by disciplinary expectations, audience awareness and institutional power structures (Hyland, 2018). From this perspective, academic writing is not merely grammatical correctness but the ability to deploy appropriate rhetorical and linguistic resources to construct knowledge and claim authority. Corpus-based EAP research has demonstrated that successful academic writing relies heavily on formulaic language, including lexical bundles, reporting structures and stance expressions. For Indian academic writers many of whom operate in an ESL context limited exposure to such phraseological patterns can result in overgeneralization or restricted lexical range. The theoretical contribution of EAP lies in linking corpus



findings to instructional relevance, emphasizing how empirical descriptions of language use can inform teaching materials and feedback mechanisms.

This study adopts an EAP-informed view by interpreting corpus patterns not only descriptively but also diagnostically, identifying areas where instruction can support greater rhetorical effectiveness without undermining varietal identity.

Systemic Functional Linguistics (SFL)

To interpret the functional significance of corpus findings, the study draws selectively on Systemic Functional Linguistics, particularly Halliday's (1994) conception of language as a resource for making meaning. SFL's three metafunction, interpersonal and textual are especially relevant to academic discourse.

1. Ideational metafunction explains the prevalence of nominalization and dense noun phrases in academic writing, which allow complex processes to be packaged as abstract entities.
2. Interpersonal metafunction underpins the analysis of hedging and stance, revealing how writers negotiate certainty, authority and politeness.
3. Textual metafunction informs the examination of cohesion and information flow, especially in relation to passivization and thematic progression.

By integrating SFL insights, the study moves beyond surface frequency counts to explain why certain linguistic choices are favoured in Indian academic writing and how they function rhetorically.

Genre Theory and Academic Discourse

The study is informed by genre theory, particularly Swales' (1990) view of genres as communicative events with shared purposes and conventions. Academic writing is not monolithic. Genre theory helps explain why certain features such as passives in Methods sections or evaluative reporting verbs in Literature Reviews are more prominent in specific contexts.

Applying genre theory allows the study to interpret variation within the corpus as contextually motivated rather than inconsistent. It also supports comparisons with international corpora by providing a principled basis for section-wise and discipline-wise analysis.

Integrative Perspective

Taken together, these theoretical perspectives position the present study at the intersection of description, interpretation and application. Corpus linguistics provides empirical rigor; World English's ensures ideological balance; EAP supplies pedagogical relevance; and SFL and genre theory enable functional interpretation. This integrated framework allows for a nuanced understanding of Indian English academic writing as both a localized academic practice and a participant in global knowledge production.

Findings and Discussion

The findings of this corpus-based study provide further empirical support for the growing body of research that positions Indian English (IndE) academic writing as a systematic, norm-governed register, rather than a deficient approximation of Inner Circle academic English. Consistent with earlier studies



(Barua, 2022; Mukhopadhyay, 2021), Indian academic texts in the present corpus demonstrate a strong preference for lexical density, nominalization and formal syntactic constructions, particularly within STEM disciplines. This tendency reflects both global academic conventions and localized pedagogical traditions that equate formality with scholarly credibility.

One significant insight emerging from this study is the overgeneralization of rhetorical strategies, especially in the use of passive voice and hedging devices. While passivation aligns with international norms in Methods and Results sections, its uniform application across all sections particularly in humanities writing suggests limited genre sensitivity. Similar observations have been reported in comparative corpus studies, which note that non-native academic writers often transfer stylistic conventions across genres and sections without adequate differentiation (Hyland, 2018; Le, 2019).

The restricted range of reporting verbs observed in the corpus further supports findings from previous research indicating limited exposure to evaluative and discipline-specific rhetorical resources among ESL academic writers (Ackermann et al., 2018). This has implications for authorial stance and critical engagement, particularly in literature review sections where evaluative positioning is essential.

Importantly, several features traditionally labelled as “errors” such as article variation or prepositional preferences appear consistently across texts and disciplines, reinforcing arguments from World English’s scholarship that such patterns may constitute endonormative features of Indian English rather than performance deficiencies. However, the study also highlights the pragmatic challenge faced by Indian scholars when communicating within international publication contexts that implicitly privilege Inner Circle norms.

The discussion underscores the need to reconceptualise Indian academic writing through a pluralistic lens, balancing recognition of varietal legitimacy with targeted pedagogical intervention to enhance global intelligibility and rhetorical effectiveness.

Conclusion

This study set out to examine the linguistic and rhetorical characteristics of Indian English usage in academic writing through a corpus-based analysis of texts produced. By combining quantitative corpus techniques with qualitative discourse interpretation, the research provides a comprehensive and empirically grounded account of how Indian academic writers deploy lexico-grammatical resources across disciplines and genres.

The findings demonstrate that Indian academic writing is characterized by high lexical density, extensive nominalization, frequent passivation and cautious stance marking. These features largely align with global academic conventions, while also reflecting localized educational practices and sociolinguistic influences. The study confirms earlier research suggesting that Indian English academic discourse has developed relatively stable norms, reinforcing its status as a legitimate and functional academic variety (Barua, 2022; Mukhopadhyay, 2021).

The study identifies areas where Indian academic writing diverges from international reference corpora in ways that may affect clarity, persuasiveness and publication success. These include overuse of certain grammatical structures, limited reporting verb repertoires and insufficient genre sensitivity. Rather



than framing these tendencies as deficits, the study argues for corpus-informed pedagogical strategies that make disciplinary and rhetorical conventions explicit to emerging scholars.

The implications of this research extend to English for Academic Purposes pedagogy, academic writing support services and policy-making in higher education. By integrating corpus evidence into teaching and mentoring practices, institutions can better support Indian scholars in navigating the demands of global academic communication without undermining their linguistic identity.

Future research may expand the corpus diachronically, incorporate reviewer and editorial feedback or explore multimodal academic genres. Such work would further enhance understanding of Indian English as a dynamic participant in international knowledge production.

References

- Halliday, M. A. K. *An Introduction to Functional Grammar*. 2nd ed., Edward Arnold, 1994.
- Swales, John M. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.
- Kachru, Braj B. *The Other Tongue: English across Cultures*. 2nd ed., University of Illinois Press, 1992.
- Biber, Douglas, Susan Conrad and Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998.
- Hyland, Ken. *Disciplinary Discourses: Social Interactions in Academic Writing*. University of Michigan Press, 2004.
- Schneider, Edgar W. *Postcolonial English: Varieties around the World*. Cambridge University Press, 2007.
- Hyland, Ken. *Academic Discourse: English in a Global Context*. Continuum, 2009.
- Biber, Douglas, et al. *Longman Grammar of Spoken and Written English*. Pearson Education, 2011.
- Swales, John M. and Christine B. Feak. *Academic Writing for Graduate Students*. 3rd ed., University of Michigan Press, 2012.
- Schneider, Edgar W. *English around the World: An Introduction*. Cambridge University Press, 2014.
- Bernaisch, Tobias and Christopher Koch. "Attitudes towards Englishes in India." *World Englishes*, vol. 35, no. 1, 2016, pp. 1–18.
- Ellis, Nick C. "Usage-Based Approaches to Language Acquisition." *The Handbook of Applied Linguistics*, edited by Robert B. Kaplan, 2nd ed., Wiley-Blackwell, 2016, pp. 1–20.



Publishing URL: <https://www.researchreviewonline.com/upload/articles/paper/RRJ807870.pdf>

- Ackermann, Kirsten, Yu-Hua Chen and Michael Myer. "Developing the Pearson International Corpus of Academic English (PICAЕ)." *International Journal of Corpus Linguistics*, vol. 23, no. 3, 2018, pp. 1–27.
- Kunchukuttan, Anoop, et al. "The IIT Bombay English–Hindi Parallel Corpus." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association, 2018, pp. 439–446.
- Hyland, Ken. "Metadiscourse and Stance in Academic Writing." *Applied Linguistics*, vol. 39, no. 1, 2018, pp. 1–23.
- Biber, Douglas, Susan Conrad and Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. 2nd ed., Cambridge University Press, 2021.
- Barua, Nivedita Malini. "Academic Writing Needs of Research Scholars in an Indian University." *Journal of English Language Teaching*, vol. 64, no. 5, 2022, pp. 33–44.
- Hyland, Ken. *Second Language Writing*. 2nd ed., Cambridge University Press, 2019.
- Mukhopadhyay, Lopamudra. "Generating Lexical Estimates in Academic Writing in English Using Technology: An Exploratory Study of Adult Indian Learners." *Language Learning & Technology*, vol. 28, no. 2, 2021, pp. 85–104.
- Le, Thi Ngoc. "Passive Voice and Grammatical Evolution in Global Academic English: A Corpus-Based Study." *English for Specific Purposes*, vol. 77, 2019, pp. 45–61.
- Ackermann, Kirsten. "Phraseology in Academic English: Corpus-Based Insights." *Journal of English for Academic Purposes*, vol. 41, 2019, pp. 1–12.
- Hyland, Ken and Feng Jiang. "Academic Discourse and Globalization." *Journal of English for Academic Purposes*, vol. 34, 2018, pp. 1–12.
- Biber, Douglas. "Complexity in Academic Writing." *Written Communication*, vol. 36, no. 2, 2019, pp. 131–159.
- Mukherjee, Joybrato. "World Englishes and Corpus Linguistics." *The Cambridge Handbook of World Englishes*, edited by Daniel Schreier et al., Cambridge University Press, 2020, pp. 308–329.