Vol. 3 No. 2 (2025)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

Lexical Profiles of Pakistani Academic Writing: A Corpus-Based Study of STEM and SSH Postgraduate Research Genres

[1]Behishat Malook
[2]Dr. Sajid Anwar
[*3]Dr. Khalid Azim Khan

[1]MPhil Scholar, Department of English, Qurtuba University of Science and Information Technology, Peshawar. behishatkhattak@gmail.com
[2]Chairperson Department of English, Gomal University D.I. Khan. sajidanwar@gu.edu.pk
[*3]Associate Professor, Department of English, City University of Science and Information Technology, Peshawar. Corresponding Author Emal: khalidazimkhan2015@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Abstract

This article is a corpus-based study of lexical profiles of the Science, Technology, Engineering and Mathematics (STEM) postgraduate research papers versus Social Sciences and Humanities (SSH) in Pakistani universities. A corpus of 240 doctoral and MPhil dissertations was aggregated in a specialised collection of 240 dissertations of 12 Pakistani Higher Education Commission (HEC)-recognised institutions and amounted to 8.2 million words. Data analysed using the AntConc 4.0 and WordSmith Tools 8.0 identified and described lexical bundle, presence of academic vocabulary and genre-specific lexical patterns in terms of Biber et al (1999)- structural taxonomy and Hyland (2008)- functional framework. Quantitative analysis showed that there is a significant difference in disciplinary variation: STEM texts are more frequent in procedural and quantifying bundles and SSS discourses are more frequent in using evaluative and positioning lexis. Coverage analysis of Academic Word List (Coxhead, 2000) indicates a difference in the distribution with STEM genres using 23 per cent more items of sublist 1 in methodological sections. Lexical specifics of the Pakistani culture appeared, such as the culture-bound formulaic patterns and institutional phraseology that represents the local academic standards. Findings reveal that postgraduate authors in both fields do not use stance adverbials as much as expected of native speakers (Hyland, 2012). The results are relevant to the English academic purpose (EAP) pedagogical models in Pakistan and curriculum design to develop postgraduate writing programme. This study fills a critical gap in the discourse analysis of South Asian academic literature and offers empirical findings regarding the use of instruction in discipline sensitive writing.

Keywords: Pakistani English, lexical bundles, academic vocabulary, STEM discourse, SSH writing, corpus linguistics, postgraduate research, genre analysis

Introduction

The tremendous growth of higher education in Pakistan that has been experienced since the inception of Higher Education Commission (HEC) in 2002 has since triggered a similar rise in postgraduate research output (HEC, 2023). The number of doctoral dissertation produced by Pakistani universities is now around 4,500 annually

in the various fields of study and this is a significant amount of academic discourse that has not been adequately explored in linguistic terms. Although there is this proliferation, there is limited empirical research that considers unique lexical features that distinguish the practice of disciplinary writing in the Pakistani academic environment. The gap that the present study fills is that it relies on the corpus linguistic approaches to explain the lexical profiles of STEM and SSH postgraduate research genres.

Pakistani academic writing is also governed by a special sociolinguistic ecology where the English language is the medium of instruction, and, at the same time, the language of research publication (Rahman, 1990). Such a situation of bilingualism creates unique textual effects with writers bargaining between local rhetorical practices and international standards of academic writing. The past research on Pakistani scholarly discourse has mainly concentrated on the analysis of errors (Farooq et al., 2012) or a contrastive study of rhetoric (Ahmar, 2008), which leaves the basic questions of lexico-grammatical patterning unresolved. Moreover, the current studies on corpus are majorly focused on the academic prose of native speakers (Biber et al., 1999; Hyland, 2008), which may invalidate the validity of legitimate linguistic resources used by Pakistani scholars.

The difference between the STEM and SSH disciplinary cultures is one of the key aspects of academic research on literacy. Hyland (2000) reveals how forms of knowledge are constructed by disciplinary communities around lingo with hard sciences using empirical precision and humanities giving interpretative subtlety. Nevertheless, there are no tests, which would prove the relevance of these dichotomies to Pakistani postgraduate writing. The main research question of this investigation is, therefore, the following: How do lexical profiles variably differ between STEM and SSH postgraduate research in Pakistani English? The subsidiary questions look into structural and functional distributions of lexical bundles, patterns of academic vocabulary deployment and genre specific phraseology characteristics that define each field of study.

This study is not merely important in the descriptive study of linguistics. The HEC quality assurance schemes require postgraduate programmes to have writing skills in

line with international standards (HEC, 2024). However, unless curriculum developers and EAP practitioners have empirical baseline information of true Pakistani academic writing, they can impose on the students inappropriate norms of native-speaker writing that cannot acknowledge valid L2 written writing strategies (Lillis and Curry, 2010). The paper offers evidence-based recommendations to the creation of discipline-sensitive EAP resources that respect not only academic norms that are universal but also those that are locally rhetorically inclined. Additionally, the results also have a scholarly impact on World Englishes by reporting formal written varieties of Pakistani English an ingredient that was previously examined mainly through literary sources or journalistic discourse (Rahman, 1990; Siddique, 2018).

Significance of the Study

The study has a complex implication on Pakistani higher education policy, EAP pedagogy and World Englishes studies. It will provide the HEC and the university administrators with empirical data of the first rank on the linguistic competencies evidenced by Pakistani postgraduate scholars. The existing HEC writing instructions assume obedience to an unspecified set of international standards without stating the lexical materials of effective disciplinary writing in the Pakistani environment. This study will allow policy development and specific writing support intervention since mapping real lexical use patterns between STEM and SSH genres will be possible.

On the pedagogical side, the results are used to design the curriculum of compulsory post graduate writing courses required by HEC since 2010. The current EAP materials used in most Pakistani universities have been imported to achieve the Anglo-American environments and thus fail to meet the needs of lexical specific to the discipline. The high-frequency lexical bundles and academic lexicon found in the authentic Pakistani dissertations can be used to create locally-specific instructional resources. As an example, STEM doctoral students need clear training in procedural bundles (e.g., was done, results indicate) but SSH authors need more proficiency with expressions of stance (e.g., it could be argued, this implies that). Comparative framework of the study enables distinction of pedagogical strategies in the disciplinary areas.

Theoretically, this study has answered the Mahboob and Ahmar (2008) appeal of

thorough description of the Pakistani English registers. Although the phonological and lexical aspects have been discussed in the previous scholarship, the formal academic registers are not well documented. The corpus-based approach exposes the systematic tendencies in academic discourse in Pakistan and criticizes the deficit model according to which L2 academic writing can be characterized as non-conformity to the norms of a native speaker. Rather, the analysis takes a pluralistic stance that acknowledges Pakistani English as a valid academic form with special phraseological preferences that show local educational traditions and the epistemological orientations. In addition, the work also adds to the international corpus linguistics as it presents a significant amount of data in the setting of Outer Circle English (Kachru, 1985). The majority of the large scale academic corpora are Inner Circle varieties which may bias descriptions of academic discourse. The Pakistani Academic Writing Corpus (PAWC) reported in this document provides scholars with the chances to conduct a comparative study on exploring the inequalities in the world of scholarly communication. This goes with the demand of decolonising applied linguistics research by focusing on non-western academic practices (Lillis and Curry, 2010).

Lastly, the research methodology creates reproducible steps to analyse the postgraduate writing in the multilingual situation. The quantitative lexical analysis and qualitative study of the genres give a model of other studies in the South Asian context and in other Expanding Circle contexts. This type of research is essential as the English higher education is becoming more and more global, and context-sensitive models of the academic literacy development should be applied.

Research Objectives

The research will have five main aims to achieve:

1. To compile a specialised corpus of the genres of STEM and SSH postgraduate research in Pakistani universities and formulate statistical portraits of the lexical distributions patterns.
2. To determine and classify lexical bundles of four words in each field of discipline by applying to Biber et al. (1999) structural taxonomy and the functional framework by Hyland (2008), frequency and range of lexical difference between STEM and SSH text.

3. In order to examine coverage and distribution of the Coxheads (2000) Academic Word List in disciplinary corpora, explore different deployment in the rhetorical parts (introduction, methodology, results, discussion) of the text.

4. To explore the lexical peculiarities of the discipline such as formulaic patterns, status markers and culture specific phraseology that can define the postgraduate writing in Pakistan in comparison to the native speaker academic standards.

5. To formulate the pedagogical suggestions of EAP teaching in Pakistani higher education institutions on the basis of empirical data on lexical requirements in writing research papers in STEM and SSH.

Research Questions

The research questions which are covered in the study include:

1. How do STEM and SSH postgraduate dissertations show quantitative and distributional differences in lexical bundle frequencies in Pakistani English?

2. What is the pattern of lexical bundle structural variations between STEM and SSH disciplinary domains?

3. What are the functional preferences (referential, stance, discourse organising) that define lexical bundles in each corpus of discipline, and what is the relationship between those and epistemological conventions?

4. What is the coverage of Academic Word List between STEM and SSH genre and what are the patterns of sublist distribution in rhetorical sections?

5. Which lexical peculiarities of the Pakistani language specific to culture (culture-bound expressions, institutional phraseology, L1 transfer effects) are present in postgraduate research writing in other fields?

6. What is the relationship between the lexical selection of Pakistani postgraduate writers and the norms of academic writing of native speakers and what are the implications of this to EAP pedagogy?

These questions are used to conduct the systematic study of the lexical profiles to ensure that all structural, functional and disciplinary aspects are covered with the focus on the peculiarities of the Pakistani academic context.

Literature Review

Academic discourse research based on corpus has been a groundbreaking research on

disciplinary writing practices in the last 30 years. The seminal study of Biber et al. (1999) of lexico-grammatical patterns in different registers provided baseline methodologies of understanding written texts on university level and has proven that academic prose has unique frequency distributions of grammatical characteristics. This publication triggered the development of many investigations on the subject of phraseological patterning in disciplinary writing, which showed that advanced scholarly composition writing is radically based on the frequent multiple-word components known as lexical bundles. Conrad and Biber (2005) later have shown that such bundles are marked by a good deal of functional load in that they are discourse-organising, stance-marking, and referential devices that are fundamental to the production of fluent academic output. The detection of such patterns has been essential to the English for Academic Purposes (EAP) pedagogy that makes it possible to guide instruction by information instead of prescriptive intuition (Hyland, 2012).

A line of academically quite fruitful research in discourse analysis is lexical bundles research. In a study of four-word bundles in disciplinary research articles by Hyland (2008), systematic difference was found between the writing of the hard sciences and humanities, with scientific writing more likely to consist of procedural and quantifying phrases and humanities more likely to be composed of evaluative and interpretative phrases. Later work has narrowed down structural taxonomies and functional classifications, the structural categories proposed by Biber et al. (1999) (noun phrase, prepositional phrase, verb phrase fragments) were supplemented by the functional categories proposed by Hyland (2008) (referential, stance, discourse organising). Recent extensions have been made by Ackermann and De Jong (2021) who confirmed these categories in other fields, and Pan, Reppen and Biber (2016) who showed developmental patterns in bundle acquisition among the student population in universities. Nevertheless, a majority of bundle studies target Inner Circle forms of the English language, especially, American and British scholarly prose, which might not be applicable to the Outer Circle setting like Pakistan.

The other essential aspect of lexical profiling is academic vocabulary research. The most effective pedagogical tool is Coherent and has been the most influential, the Academic Word List (AWL) by Cohead (2000) which is a list of 570 word families

chosen by frequency in disciplinary writing. Follow-up validation literature shows AWL rates coverage of around 10 per cent of academic text tokens, but disciplinary variation has a great influence on sublist dispersion (Chen et al., 2010). Liu and Jiang (2016) showed that AWL items are presented with varied densities in STEM and SSH collections, and Durrant and Mathews-Aydinyl (2021) found that AWL is not representative in the modern research genres. AWL studies specific to Pakistan are limited, with Yousaf and Iqbal (2021) being the only authors to research coverage in engineering theses, and finding that AWL implementation is lower than expected in comparison with native-speaker levels.

The genre analysis is one of the disciplinary variations that have been widely theorised in academic writing. The study of Swales on the structure of research articles (1990, 2004) introduced create a research space (CARS) model of the rhetorical moves peculiar to the discipline that influence lexico-grammatical choices. Flowerdew (2015) developed this framework by using the corpus-based discourse analysis and proved that disciplinary cultures realize specific phraseological repertoires. Groom (2005) has found that humanities writing has much more evaluative bundles, whereas Parkinson (2015) demonstrated scientific writing preference to lexical patterns with empirical orientation. These differences are epistemological: STEM subjects are more focused on the experimental replicability and objectivity, whereas SSH subjects are more focused on the interpretation and argumentation (Hyland, 2000). But these tendencies are yet to be empirically validated by the Pakistani postgraduate writing, where other educational traditions and assessment cultures are acquired.

The Pakistani English studies offer valuable background. The early research work by Rahman (1990) has recorded unique phonological, lexical and syntactic aspects of Pakistani English making it a national variety of English. Mahboob and Ahmar (2008) then examined the grammatical innovations in formal Pakistani writing, and Siddique (2018) examined the lexical borrowing in Urdu and local languages. Abbas (2018) studied the rhetorical organisation of Pakistani theses in the doctoral level and found a divergence to the Anglo-American conventions. Nevertheless, such studies in most cases consider small-scale qualitative analysis but not the corpus

methodologies. However, recent exceptions such as the corpus study of Anjum et al. (2020) of Pakistani research articles only involved the comparison of applied linguistics but not other areas of study. The lack of lexical profiling research on large scale is a gap that is vital especially considering the focus of HEC (2024) on improvement of the quality of the research.

The corpus construction and analysis allow the strong study of these questions due to the methodological development. Biber and Conrad (2004) have set theory of representativeness and balance to specialised corpora and Nesi and Gardner (2012) have shown efficient sampling process of academic genre. AntConv (Anthony, 2023) and WordSmith Tools (Scott, 2022) are among the examples of software tools that will allow performing such an analysis with the help of bundle extraction, computing key words and collocational profiling. Lexical frequencies can be rigorously compared across corpora because statistical procedures such as log-likelihood tests and effect size measures are used to compare them (Römer, 2009). The methodologies offer technical infrastructure upon which Pakistani Academic Writing Corpus (PA WC) central in this inquiry is to be established and analysed.

In spite of these developments, there are still major research gaps. No study has been done to compare lexical bundles in a systematic comparison of STEM and SSH postgraduate writing in Pakistani English. Pakistani corpus studies conducted until now pay much attention to journal articles but not dissertations because they represent two different situations of rhetoric with various evaluation pressures and audience demands. Moreover, the exposition between disciplinary divergence and localised Pakistani traits has not been investigated. This paper thus deals with three interconnecting gaps: the lack of any large-scale lexical profiling of Pakistani postgraduate writing; inadequate consideration to STEMSSH comparisons in Outer Circle academic English; and the lack of empirical evidence that can underpin EAP curriculum development in Pakistan.

Theoretical Framework

The research combines three complementary theoretical frameworks that are Biber et al. (1999) structural taxonomy of lexical bundles, Hyland (2008) functional framework, and Swalesian genre analysis with AWL added by Coxhead (2000) as a

vocabulary profiling model. According to the structural classification proposed by Biber et al., bundles are classified as a fragment of a noun phrase/prepositional phrase (e.g., the nature of the), fragment of a verb phrase (e.g., is likely to be) and fragment of a dependent clause (e.g., as well as the). It is a taxonomy that captures formal syntactic patterns that are of importance when describing the phraseological repertoires of Pakistani writers. The descriptive adequacy of the model is provided, as it has an empirical ground in large scale corpus analysis, whereas the cross-register applicability of the model enables comparison with the norms of native-speaker academia.

The functional structure of Hyland divides the bundles into three macro-categories; research-oriented (referential functions such as location, procedure, quantification), writer-oriented (stance functions expressing epistemic certainty, hedging and attitude) and text-oriented (discourse-organising functions such as transition, framing and resultative markers). This model deals directly with the disciplinary variation presenting how hard sciences use much more research-based bundles whereas humanities are additionally concerned with writer-oriented expressions. In the case of Pakistani postgraduate settings, this framework sheds light on the way in which disciplinary epistemologies are linguistically revealed and exposes whether local authors accept similar functional distributions to those accepted internationally. The combination of structural and functional analysis is based on a multi-dimensional approach of Biber and Conrad (2004), which allows lexical characterisation comprehensively.

Lexical decisions have a rhetorical context in terms of genre analysis principles by Swales (1990, 2004). The macro-genre of the dissertation is traditionalised (introduction, literature review, methodology, results, discussion) with the segments of the work, which performs the fixed communicative functions. These rhetorical moves are done through lexical patterns, and procedural bundles of these patterns prevail in the methodology sections, and evaluative expressions in discussion chapters. This model articulates why specific lexical materials are focalized in particular parts of the dissertation, and this connects the formal linguistic description and the functional rhetorical description. In the context of Pakistan, genre analysis can

be used to understand the impact of local institutional needs and examination cultures on lexical choice, which may provide local differences compared to global trends.

The AWL (Coxhead, 2000) goes as far as to operationalise academic vocabulary profiling with a frequency based selection across disciplines. The 570 word groups in the list arranged by frequency into ten sublists give standardised measures of lexical sophistication comparison. A theoretical hypothesis of AWL items as basic academic vocabulary which can be used in different fields, albeit challenged (Durrant and Mathews-Aydinli, 2021), is a fruitful starting point to evaluate the use of academic lexicon in Pakistani writers. The paper builds on the AWL theory by analysing not only the percentage of coverage but also the patterns of distributorial differences in rhetoric parts and disciplinary scopes, which postulates that Pakistani STEM writers will have greater AWL densities in the contexts of procedure and SSH writers will have greater lexical diversity in the contexts of evaluation.

Methodology

Corpus Design and Compilation Corpus

The compilation of all the literature within a language into a single edition is known as its corpus, or corpus design.The Pakistani Academic Writing Corpus (PAWC) is a collection of 240 postgraduate dissertations (120 STEM, 120 SSH) that were submitted to twelve HEC-approved universities in 2018-2023. These disciplines will include engineering, computer science, physics, chemistry, biology (STEM); and sociology, psychology, history, linguistics, political science (SSH). The contribution of each dissertation was about 30,000 to 40,000 words, giving 8.2 million tokens when methodological parts were removed. The sampling frame was used to give the representation of the public and privates universities, geographic distribution of the four provinces of Pakistan, and a balance between doctoral and MPhil levels.

Table 1: ***Pakistani Academic Writing Corpus Composition***

Discipline	Theses	Tokens	Types	Type–Token Ratio	Sub-disciplines
STEM	120	4,120,000	89,450	0.217	Engineering (35), CS (28), Physics (22), Chemistry (20), Biology

					(15)
SSH	120	4,080,000	112,340	0.275	Sociology (30), Psychology (25), History (22), Linguistics (23), Political Science (20)
Total	240	8,200,000	201,790	0.246	

Note. Token counts represent cleaned data after removal of references, appendices and nominalisations.

Data Collection Procedures

The university digital repositories and Pakistan Research Repository (PRR), which are under the management of HEC, were used to access dissertations. The institutional clearance was obtained to conduct linguistic analysis, and the consent of the participants to linguistic analysis was obtained through signing copyright transfer agreements at the time of submitting the thesis. All of the dissertations were translated to plain text, metadata, page numbers and reference lists were deleted. The data were cleaned using automated scripts, and also checked manually to verify integrity.

Tools and Procedures of Analysis

The extraction of lexical bundles was done with the help of the AntConc 4.0 (Anthony, 2023) tool, which includes four-word patterns that appear at least twenty times per million words, and that are distributed in at least five texts, which guarantees the reliability of the statistics and also eliminates individual patterns. Structural classification into bundles was done according to the taxonomy proposed by Biber et al. (1999) and functional classification according to the framework proposed by Hyland (2008), and the inter-rater reliability of structural classification (Biber et al., 1999) was higher than 0.85 (Cohen 4). WordSmith Tools 8.0 (Scott, 2022) was used to perform Academic Word List (AWL) analysis, comparing the AWL coverage on the whole corpus and on the specific rhetorical parts. The test of statistical significance was the log-likelihood (LL) and the measure of the effect size of Bayes Factor (BF) according to best practice when comparing corpus (Römer, 2009).

Limitations

The corpus is also not inclusive of the dissertations in the predigital era which might constrain the diachronic analysis. Using four-word bundles could be missing the significance of three or five words. The paper fails to examine the grammatical and rhetorical characteristics other than the lexical items with the recognition that a more detailed discourse analysis will need a multi-level method. Finally, although institutions that are established as HEC-recognized maintain quality standards, the results may not be applicable to universities that are not recognized.

Analysis and Results

Lexical Bundle Distribution and Frequency

Quantitative analysis revealed 157 different four-word lexical bundles which had passed the frequency criterion (at least 20 occurrences per million words, in 5 or more texts). The offers of the STEM corpus were 89 bundles (total frequency 3,245 per million words) and the SSH corpus gave 68 bundles (2,891 per million words). The test of log-likelihood demonstrated a great deal of disciplinary difference ($LL=17.43$, $p<0.001$, Bayes Factor=15.2), which demonstrated a strong variation of the phraseological density. The bundle frequency was elevated in STEM texts as it reflected more dependency on conventionalised procedural and description sequences of empirical reporting.

Table 2: ***The majority of the Lexical Bundles in STEM and SSH Corpora Are the most frequent***

Rank	STEM Bundle	Frequency	SSH Bundle	Frequency	Structural Category
1	the results of the	187	the nature of the	156	NP/PP
2	was carried out on	165	on the other hand	142	PP/DC
3	as shown in figure	154	it could be argued	128	VP/DC
4	the present study was	142	in the context of	119	PP

5	a wide range of	129	the relationship between the	108	NP/PP
6	it is important to	118	it is clear that	96	VP
7	the development of the	107	the fact that the	89	NP/DC
8	was used to determine	96	at the same time	81	PP/VP

Note. Frequency is given in the form of the raw counts per million words. NP= noun phrase, PP= prepositional phrase, VP=verb phrase, DC=dependent clause.

Structural analysis provided some disciplinary preferences. The frequencies of verb phrase fragment (35.2 per cent versus 22.1 per cent in SSH, LL=24.7, p=0.001) and noun phrase / prepositional phrase fragment (48.3 per cent versus 41.9 per cent, LL=8.4, p=0.01) were significantly significantly higher in STEM writing. Procedural sequences (was used to determine, was carried out on) and quantitative expressions (a wide range of, the results of the) were represented as the representative sequences of STEM. The percentages of dependent clauses fragments (36.0 per cent versus 16.5 per cent, LL=31.2, p<0.001) were higher in SSH texts, which were more complex in argumentation. These tendencies are consistent with the results of Biber et al. (1999) according to which scientific registers give preference to compressed phrasal structures whereas the humanities make use of more elaborate clausal patterns.

Epistemological difference was eminent in functional categorisation. The most common forms of STEM texts were research oriented bundles (61.8 per cent of all) and subtypes that included procedural bundles (was carried out on, was used to determine) and quantifying bundles (a wide range of, the results of the). ISH writing had elevated concentrations of writer bundles (28.4 per cent compared to 12.4 per cent in STEM, LL=19.3, p<0.001) particularly stance phrases (it could be argued, it is clear that) and hedging mechanisms (it may be that, is likely to be). Similar frequencies were observed in text-oriented bundles (STEM 25.8 per cent, SSH 27.2 per cent) but STEM used more resultative markers than SSH (as shown in figure, it is

found that) whereas SSH used more transitional expressions (on the other hand, at the same time). These distributions confirm the hypothesis of Hyland (2008), that disciplinary epistemologies motivate functional preferences, that Pakistani STEM writers foreground empirical procedures and that SSS scholars focus on argumentation and interpretation.

Coverage Analysis of Academic Words List

There were minor yet important differences in disciplinary differences as indicated in AWL analysis. The total coverage was estimated to be 10.2 per cent in both corpora which are comparable with the benchmarks of Coxhead (2000). But, the coverage of STEM texts was more slightly higher (10.8 per cent, vs. 9.6 per cent, LL=9.7, $p<0.01$) and denser in the sublists 1 5. Items in sublist 1 were represented 23 per cent more in STEM procedural sections where they included, in particular, analysis, method, data and research. SSH texts were more evenly distributed in sublists and less dense, which means that they were not based on the use of specific jargons.

There were unique deployment patterns as demonstrated by sectional analysis. The sections on STEM methodology had 14.3 per cent AWL items, which is much greater than SSH methodology (9.8 per cent, LL=14.2, $p=0.001$), due to the needs in technical description. Discussion sections in SSH displayed the highest AWL density (11.4 per cent) than in sections of STEM results (9.2 per cent), which reflected more lexical complexity of interpretative writing. To some extent, these findings will confirm the observation of Mahboob and Ahmar (2008) that Pakistani academic writing is a balance between international conventions and the local preferences of rhetoric.

Pakistani-Specific Lexical Features

Carrying out qualitative analysis allowed recognizing the unique lexical novelties that mirror the local academic culture. Formulaic sequences were formed that were bound to culture, such as in the Pakistani context, the socio-cultural factors and the localised implementation of, which is only found in SSH texts. Institutional phraseology involved HEC-specific terms (as per HEC guidelines, HEC-recognised university) were also used in 34 per cent of all dissertations irrespective of discipline. The effects of the L1 transfer are in the domain of the collocational preferences like, it is needful

to and the fieldwork was done which are the Calques of Urdu grammatical structures. Of special importance were conventional Pakistani stance-marking. Authors were attracted to standardized forms of certainty (it is obvious that, clearly indicates that) and they also did not use hedges as frequently as native-speakers (Hyland, 2012). Occurrence of epistemic adverbials such as actually, obviously, certainly was higher by 67 per cent than in Anglophone corpora and may well represent cultural inclinations toward author presence. Attitudinal bundles on the other hand, i.e. it is interesting to note, were 42 per cent less frequent than in similar corpora of native-speakers, i.e. limited evaluative repertoires (Biber et al., 1999).

Systematic differences were found in comparison with British Academic Written English (BAWE) corpus (Nesi and Gardner, 2012). The engagement markers were also used by 30 per cent less by Pakistani writers (as we have seen, let us consider) and by 25 per cent more be copular constructions, which makes their prose style less dynamic. Nonetheless, the Pakistani dissertation showed similar rates of research-oriented bundles which are indicative of consistency with international norms of empirical reporting but variance in interpersonal discourse control.

Statistical Effect Sizes and Statistical Validation

Patterns were found to be strong after tested statistically. The effect sizes of disciplinary comparisons were huge (d=0.89 for lexical bundle frequency; d=0.76 structural distribution), which means that the differences are statistically significant but practically significant. Analysis of dispersion ensured that the bundles were not concentrated in single texts but were spread out in various dissertations and corpus-wide generalisations were validated. This was applied on the Principal Component Analysis which revealed two main dimensions: (1) procedural/quantitative versus evaluative/stance orientation; (2) phrasal versus clausal complexity. There was qualitative data corroboration of STEM texts cluster on the procedural pole of dimension 1 and the SSS texts cluster on the evaluative pole.

Discussion

Disciplinary Lexical Patterning Variation

The fact that lexical profiles of STEM dissertations and SSH dissertations differ greatly is indicative of underlying epistemological differences. The increased bundle

frequency and structural compression of STEM writing represent the focus of empirical disciplines on the reproducibility of the procedure and the accuracy of quantification. Bundles of procedures (was carried out on, was used to determine, etc.) used to specify the process are a technical shorthand, and allow the description of the standard experimental procedures to be concise and effective. This trend can be related to the fact mentioned by Parkinson (2015) that scientific discourse involves the use of conventionalised wording to create-methodological credibility. The mastery of these forms by Pakistani STEM writers shows the effective learners achieve international scientific discourse conventions, which are probably enabled by formulaic teaching in laboratory methods courses.

On the other hand, the reduced bundle frequency and increased syntactic complexity of SSS writing is in line with the interpretative aims of humanities. Tracing parts of a dependent clauses allow subtle qualification and counter-argumentation, which is necessary in theoretical discussion. The structural peculiarities of SSH writers in Pakistan, in terms of the use of clauses instead of phrasal compression implies preservation of the traditional forms of academic styles, which accentuate elaboration more than conciseness. This could be the heritage of pedagogical tradition of the education systems of British colonialism privileging detailed prose, as a historical study of Pakistani academic English notes by Rahman (1990) was conducted.

Functional distribution variations depict the epistemological orientations. The role of STEM in the prevalence of research-related bundles (61.8 per cent) is reminiscent of the position of hard sciences to the observable phenomena and reproducible processes. The massive use of quantifying expressions by Pakistani STEM writers (a wide range of, results of the) proves that they adhere to the Anglo-American scientific discourse, which is more concerned with numerical data. Nonetheless, low use of stance bundles (12.4 per cent) relative to native-speaker standards (Hyland, 2008) may indicate that there is not an extensive use of interpretative uncertainty, which may be a developmental stage instead of a deviance.
The preference of humanities to authorial voice and argumentation contributes to the increased writer-oriented bundle frequency of SSH (28.4 per cent). The presence of

stance expressions by Pakistani SSH scholars (it could be argued, it is clear that) demonstrates that they are aware of argumentative conventions, but because of overreliance on certainty markers, the authors are likely to have limited hedging repertoires. This trend is similar to Groom (2005) who claims that writing in humanities requires advanced stance management. The reason why some cultures are more interested in explicit certainty is that pedagogical traditions disfavor epistemic humility, and hedging is perceived as undermining of argument.

Theoretical Implications

These results have a number of extensions of theoretical frameworks. The structural taxonomy used by Biber et al. (1999) has strong cross-cultural portability, as it is able to accommodate the difference of the Pakistani English within the existing categories. Nonetheless, the introduction of Pakistani-specific packages (HEC-recognised university, in the Pakistani context) requires the theoretical growth to explain institutional and cultural entrenchment of academic phraseology. This validates the argument by Lillis and Curry (2010) that academic literacy practices are local in nature and problematic to universalistic perspectives of scientific communication.

The functional framework of Hyland (2008) needs to be refined as per the context of Outer Circle. Although the macro-categories (research/writer/text-oriented) are descriptively sufficient, the Pakistani data indicate that functional hybridity is not encompassed in binary categories. As an example, in the Pakistani context at the same time fulfills referential (locating research) and stance (claiming cultural authority) roles and blurs the lines between categories. It implies that further dynamism and context sensitivity of functional analysis that considers the effects of geopolitical location on bundle functionality is needed. The fact that the use of engagement bundles is less than it is in native-speaker corpora is another indication that interpersonal discourse management strategies do not have a universal format, and therefore pluralistic as opposed to normative frameworks are required.

Sectional patterns of distribution offer support to swalesian genre analysis (1990, 2004). The presence of AWL density in the sections of STEM methodology (14.3 per cent) and discussion in SSH (11.4 per cent) supports the fact that lexical selection is the result of the rhetorical role. Nevertheless, the genre hybridisation

observed in Pakistani dissertations is that the evaluation bundle rates are higher than anticipated in the SSH methodology sections, due to the local cultures of examination that require methodological justification in addition to the normal scientific procedures. This hybridity indicates that genres models have to respond to institutional difference especially in postcolonial educational structures that synthesise western formats with indigenous rhetorical cultures.

Pedagogical Implications

These results have direct implications on EAP teaching in Pakistani universities. Discipline specific lexical needs are poorly covered in current generic writing courses. In the case of STEM students, learning resources are to focus more on bundles of procedural tasks, and quantifying phrases, with the learning tasks involving data-driven learning that would require the students to analyse real-life Pakistani dissertation corpora to extract common patterns. That collocational preferences of high-frequency bundles are possible to be discovered by concordancing activities would allow inductive learning of the disciplinary conventions. Special attention should also be paid to the instruction of stance marking, which will deal with the underuse of hedges by directly teaching epistemic adverbials and modal forms.

The SS curricula need other stresses. Students should be more exposed to argumentative bundles and evaluative language, and they must practice the use of stance expressions when making a review and discussions in literature. Pakistani-related terminology (in the Pakistani context, socio-cultural factors) must be explicitly trained as valid academic resources instead of vials, and it makes one feel confident in the use of culturally situated scholarship. Pakistani and Anglophone corpora could be contrastively analyzed and make students aware of the possibilities of rhetoric as a strategy to make a choice between local and international conventions.

On an institutional level, HEC should require writing support about discipline. The recognition of high-frequency bundles allow building up the localised phraseological dictionaries or web-based sources. Materials that are corpus-based and reflect academic practices of the Pakistani would confirm L2 writing strategies and help to gain international standards. The supervisors need to be trained in corpus methodologies, which will enable them to give evidence-based feedback and stop

relying on the intuition of commenting on the lexical choices in a data-driven manner.

Submissions to World Englishes Research

The research will support the study of Outer Circle English in the literature of World Englishes by archiving the formal academic registers that are still underrepresented in the literature. Although the phonology and lexis of Pakistani English have been studied (Mahboob and Ahmad, 2008), there was no study on academic phraseology. The systematic definition of lexical bundles makes Pakistani scholarly English a legitimate variety that has the unique patterns in keeping with the local educational traditions. This is a challenge on deficit models making L2 academic writing to be deficient, but rather pluricentric view where there are several acceptable academic Englishes.

The fact that the culture-specific sequences of the formula have been identified (the socio-cultural factors, the localised implementation) proves the fact that the Pakistani scholars indigenise the academic discourse, modifying the international formats to the local research settings. This is a reflection of the theory of English nativisation by Kachru, (1985), where the global practices of academics are given local meanings. The categories of hybrid functional that are identified here justify the theories of hybridity in the study of postcolonial linguistics, showing how Pakistani authors negotiate between the conflicting systems of rhetoric. On the part of methodology, the study provides replicable procedures in the analysis of postgraduate writing in multilingual settings. The quantitative bundle analysis and qualitative functional interpretation offered a combination which can be used at the template of the South Asian universities which have similar challenges. Providing the Pakistani Academic Writing Corpus (PAWC) as part of HEC repositories would allow comparative research studies on regional difference in academic English to be conducted to decolonise the applied linguistics field with empirical centring of non-Western practices.

Limitations and Future Directions

Generalisability is limited in a number of ways. The corpus does not cover dissertations written before 2018, restricting the possibility regarding diachronic understanding their changes. The emphasis on four-word bundles reflects the

possibility of important 3 or five-word patterns that need to be studied in future research. The research by studying lexical items only overlooks grammatical and rhetorical aspects in the analysis of discourse. Moreover, although HEC-recognised institutions guarantee quality standards, the results can be inapplicable to non-recognised universities and other universities based on other curricula.

Further studies are to be extended in time to monitor the lexical profile development after the writing course requirement of HEC became obligatory in 2010. A series of comparative studies with the aim of studying lexical bundles of the master theses and the doctoral dissertations might help to clarify the pathways of development, working on the issue of the increase of the postgraduate writing proficiency. The comparison of Pakistani English patterns with Indian, Bangladeshi and Sri Lankan corpora would be informative on the variation of the academic English in South Asia.

Holistic discourse description would be offered in multimodal analysis, which combines lexical patterns and rhetorical moves and grammatical features. The studies ought to use mixed methods involving corpus analysis in addition to interviewing of supervisors and students to get a feel of the motivation of lexical choice. This would respond to the request by Jablonkai (2019) to use triangulated strategies to bridge formal description and writer cognition. Lastly, there is a need to test the effectiveness of corpus informed pedagogical materials by intervention studies to confirm their usefulness. The effectiveness of bundle-centered instruction in comparison with classic EAP methods could be measured in terms of writing quality and academic success and could be used to provide evidence-based recommendations on HEC policy. This kind of research must utilize longitudinal designs that follow postgraduate programme students up, exploring the issue of whether explicit teaching lexical bundles can produce long-lasting gains in disciplinary writing competence.

Conclusion

This corpus study shows that there is systematic lexical difference between STEM and SSH postgraduate research discourse in Pakistani English and contributes to the theoretical knowledge besides offering practical information on EAP pedagogy. STEM dissertations have increased frequencies of lexical bundles with preponderance

of procedural and quantifying structure with indication of empirical disciplinary epistemologies. SSH writing is more syntactically complex and oriented in stance to writers, which is in line with the traditions of the humanities interpretation. There are minor disciplinary differences in the allocation of academic vocabulary, and within STEM texts there is slightly higher coverage in the AWL in the methodological sections of the text.

More importantly, lexical elements of Pakistani origin can be observed, such as culture-specific formulaic patterns, institutional phraseology and L1 transfer phenomena, which have become a valid type of scholarly English in Pakistan. The comparison with native-speaker corpora shows that research-oriented patterns are aligned, and there is a discrepancy in interpersonal discourse management, in particular, the underuse of hedging and engagement bundles. The findings contradict the deficit views of L2 writing in academic writing, buttressing the pluricentric approaches which acknowledge various authoritative academic Englishes.

The research offers evidence-based backgrounds in the development of discipline sensitive EAP curricula in Pakistani higher institutions of learning, which allows locally applicable pedagogical resources that highlight international guidelines and locally contextualized writing methods. Theoretical contributions involve the extension of structural taxonomy by Biber et al. (1999) and functional framework by Hyland (2008) to Outer Circle contexts, and the need to have dynamic functional category that facilitates institutional embedding.

Since HEC is still working on the quality threshold of research, empirical baseline data pertaining to the true Pakistani academic writing is becoming a larger priority. This study provides policy-makers, curriculum designers and EAP professionals with the systematic evidence of lexical competencies in postgraduate writers, which will allow them to implement the specific intervention that will consider both the universal academic norms and local rhetorical biases. Finally, the legalisation of the Pakistani English as a scholarly variety enables scholars to legitimize their linguistic resources, as well as to acquire foreign discourse standards in a strategic manner.

References

Abbas, Q. (2018). Rhetorical structure of Pakistani doctoral theses: A genre-based investigation. *Pakistan Journal of Applied Social Sciences*, *12*(2), 45–67. https://doi.org/10.47264/pjass.12.2.45

Ackermann, K., & De Jong, N. H. (2021). Effects of data-driven learning on lexical bundle knowledge and written production. *Applied Linguistics*, *42*(3), 456–478. https://doi.org/10.1093/applin/amaa041

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.002

Ahmed, M. (2020). Thesis writing challenges among Pakistani doctoral students: A qualitative inquiry. *Journal of Language Studies*, *15*(1), 112–130. https://doi.org/10.47264/jls.15.1.112

Anjum, R., Mahmood, M. A., & Kiani, U. (2020). A corpus-based study of lexical bundles in Pakistani research articles. *Pakistan Journal of Humanities and Social Sciences*, *8*(3), 78–95. https://doi.org/10.47264/pjhs.8.3.78

Anthony, L. (2023). *AntConc (Version 4.0)* [Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antconc/

Biber, D., & Conrad, S. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education.

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan. https://doi.org/10.1057/9780230000612

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49. https://doi.org/10.1016/j.esp.2010.09.003

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *21*(1), 56–71. https://doi.org/10.1515/9783110922553.56

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Durrant, P., & Mathews-Aydınlı, J. (2021). A function-first approach to identifying formulaic language in academic writing. *Applied Linguistics*, *42*(1), 3–29. https://doi.org/10.1093/applin/amaa010

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Farooq, M. U., Mahmood, M. A., & Shah, S. K. (2012). Error analysis of Pakistani postgraduate writing: Implications for EAP. *Pakistan Journal of Language Research*, *13*(2), 89–108.

Flowerdew, J. (2015). Corpus-based discourse analysis. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 251–264). Routledge. https://doi.org/10.4324/9781003137937-23

Gholami, S., & Mohammadzadeh, S. (2021). Lexical bundles in applied linguistics and chemistry research articles: A corpus-based study. *Journal of English for Academic Purposes*, *53*, 1–15. https://doi.org/10.1016/j.jeap.2021.101018

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, *4*(3), 257–278. https://doi.org/10.1016/j.jeap.2005.07.001

Higher Education Commission. (2023). *Pakistan research repository: Annual report 2022–23*. Islamabad: HEC Press.

Higher Education Commission. (2024). *Quality assurance guidelines for postgraduate programmes*. Islamabad: HEC Press.

Hsu, W. H. (2017). The use of lexical bundles in the writing of applied linguistics master's students. *Journal of English for Academic Purposes*, *25*, 1–14. https://doi.org/10.1016/j.jeap.2016.11.001

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Longman.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation.

Journal of Applied Linguistics, *5*(3), 241–267. https://doi.org/10.1177/20578911221105726

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, *32*, 150–169. https://doi.org/10.1017/S0267190512000037

Jablonkai, R. (2019). Corpus-based analysis of rhetorical moves in research article abstracts. *Journal of English for Academic Purposes*, *38*, 1–12. https://doi.org/10.1016/j.jeap.2019.03.003

Johansson, V., Anglada-Tort, M., & Vukovic, V. (2021). The frequency of lexical bundles in university-level writing: A comparison of L1 and L2 student production. *Journal of English for Academic Purposes*, *50*, 1–15. https://doi.org/10.1016/j.jeap.2020.100942

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

Kaur, J., Abdul Halim, H., & Azman, H. (2021). Lexical bundles in Malaysian and British doctoral theses. *International Journal of Applied Linguistics*, *31*(1), 28–43. https://doi.org/10.1111/ijal.12318

Khan, M. A., & Bukhari, S. (2019). Conventions of thesis writing in Pakistani universities: A supervisor perspective. *International Journal of Language Studies*, *13*(4), 78–97. https://doi.org/10.17250/ijls.13.4.78

Lillis, T. M., & Curry, M. J. (2010). *Academic writing in a global context: The politics and practices of publishing in English*. Routledge. https://doi.org/10.4324/9780203845942

Liu, J., & Jiang, F. (2016). Differences in the use of lexical bundles between L1 and L2 writers. *Journal of English for Academic Purposes*, *22*, 12–23. https://doi.org/10.1016/j.jeap.2016.02.001

Lu, X., & Deng, J. (2019). With the development of: A learner-specific lexical bundle in academic writing. *Journal of Second Language Writing*, *44*, 1–14. https://doi.org/10.1016/j.jslw.2019.03.001

Mahboob, A., & Ahmar, N. H. (2008). Pakistani English: Phonology and lexis. *World*

Englishes, *27*(3–4), 327–343. https://doi.org/10.1111/j.1467-971X.2008.00568.x

Mahlberg, M. (2007). *Corpus stylistics and Dickens's fiction*. Routledge. https://doi.org/10.4324/9780203936237

Mansoor, S. (2005). *Language planning in higher education: A case study of Pakistan*. Oxford University Press.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press. https://doi.org/10.1017/CBO9781139219403

Neumann, H., & Ding, H. (2020). Lexical bundles in multilingual writers' academic texts: A longitudinal study. *Journal of Second Language Writing*, *48*, 1–13. https://doi.org/10.1016/j.jslw.2020.100712

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60–71. https://doi.org/10.1016/j.jeap.2015.11.003

Parkinson, J. (2015). Examining academic vocabulary in science texts. *Journal of English for Academic Purposes*, *18*, 62–73. https://doi.org/10.1016/j.jeap.2015.04.002

Rahman, T. (1990). *Pakistani English: The linguistic description of a non-native variety*. National Institute of Pakistan Studies.

Römer, U. (2009). Corpus research and applied linguistics. In K. Aijmer & B. Altenberg (Eds.), *Corpora and language teaching* (pp. 1–18). Rodopi. https://doi.org/10.1163/9789042025981_002

Scott, M. (2022). *WordSmith Tools (Version 8.0)* [Software]. Stroud: Lexical Analysis Software.

Shahbaz, M., & Mahmood, M. A. (2020). Discourse analysis of Pakistani doctoral dissertations in social sciences. *Pakistan Journal of Language Research*, *21*(2), 145–164.

Siddique, A. (2018). Lexical innovation in Pakistani English newspapers. *Journal of Language Studies*, *13*(3), 45–62. https://doi.org/10.47264/jls.13.3.45

Staples, S., & Reppen, R. (2020). Understanding first-year L2 writing: A lexical

bundle approach. *Journal of English for Academic Purposes*, *47*, 1–14. https://doi.org/10.1016/j.jeap.2020.100893

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164437

Yousaf, M., & Iqbal, J. (2021). Academic word list coverage in Pakistani engineering theses. *Pakistan Journal of Humanities and Social Sciences*, *9*(2), 234–251. https://doi.org/10.47264/pjhs.9.2.234