https://doi.org/10.17323/jle.2023.16044

A Corpus-Based Investigation of Phrasal Complexity Features and Rhetorical Functions in Data Commentary

Muhammed Parviz [1], **Ge Lan** [2]

[1] Imam Ali University, Tehran, Iran

[2] City University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Introduction: In academic written texts, linguistic and rhetorical features are often interactively used as a vehicle for writers to construct their texts in order to accomplish specific communicative purposes. However, the effective integration of these resources may pose challenges for developing writers.

Purpose: This study employed a corpus-based genre analysis approach to investigate phrasal complexity features and rhetorical functions in data commentaries written by Iranian undergraduate and graduate students. Through this approach, we aimed to examine a relatively unexplored genre of data commentary in terms of its phrasal complexity features, rhetorical functions, and their relationships. By analyzing these relationships, we sought to provide insights into the writing practices of Iranian undergraduate and graduate students in the context of data commentaries.

Method: This study employed a convenient sampling method to select a total of 76 university students, which included 47 undergraduate students and 29 graduate students. The participants were involved in generating a corpus of 380 data commentaries, which were then thoroughly examined and compared. To identify instances of phrasal complexity features, the researchers utilized the AntConc software tool, applying regular expressions (regex) to extract potential occurrences. Additionally, a Python program was developed and implemented to calculate the frequencies of the identified PCFs. The researchers manually annotated the rhetorical function of the data commentaries to determine their specific usage.

Results: Statistical analysis such as Mann Whitney U test and Spearman correlation test, revealed that graduate students significantly utilized more phrasal complexity features including attributive adjectives, nominalizations, and prepositional phrases (of) compared to undergraduate students. However, a qualitative analysis showed that the use of these linguistic features is influenced by the writing topics. Regarding rhetorical functions, graduate students used more moves and/or steps related to presenting and commenting data, while undergraduate students produced more moves or steps concerning personal asides. Moreover, certain phrasal complexity features and the moves and/or steps were found to be correlated, aligning with recent corpus-based studies.

Conclusion: The study concludes with pedagogical implications.

KEYWORDS

phrasal complexity, rhetorical function, academic writing, data commentary, corpus-based approach

Citation: Parviz M., & Lan G. (2023). A Corpus-based Investigation of Phrasal Complexity Features and Rhetorical Functions in Data Commentary. *Journal of Language and Education, 9*(3), 90-109. https://doi.org/10.17323/jle.2023.16044

Correspondence:
Muhammed Parviz,
mohammad.parviz60@gmail.com

Received: September 27, 2022
Accepted: September 15, 2023
Published: September 30, 2023

INTRODUCTION

As an essential medium for developing scientific knowledge, academic writing is seen as an indispensable part of knowledge construction. For this reason, learning academic writing is a demanding job due to the challenges it may pose concerning how to transform knowledge into a disciplinarily appropriate entity for a particular community. Chief among multifarious linguistic resources closely associated with academic writing, phrasal complexity features (PCFs) are considered an important group of linguistic features for academic discourse (Biber & Gray, 2010). Researchers have recently focused on investigating PCFs in various written genres such as argumentative papers (Lan & Sun., 2019), research articles (Parviz et al., 2020), and MA theses (Parkinson & Musgrave, 2014). Closely intertwined with linguistic features are rhetorical functions. In academic written texts, linguistic and rhetorical features are often interactively used as a vehicle for writers to construct their texts in order to accomplish specific communicative purposes. However, the effective integration of the linguistic and rhetorical resources in academic writing may pose challenges to developing writers such as undergraduate students. Among the existing research on grammatical complexity features (including PCFs), a functional-register approach has been frequently used to interpret the rhetorical functions of selected grammatical features in academic written texts (Biber et al., 2022).

In terms of second language (L2) writing research, most existing research focuses on some common genres such as argumentative papers, theses/dissertations, and exam essays. However, as an important written genre, data commentary (DC) has been under-represented in L2 writing research on PCFs and their rhetorical functions. DC is a multimodal way of presenting data and is described as "the verbal comments on visual materials" (Nordrum & Eriksson, 2015, p. 59). DC is important to student writers, especially L2 students, because 1) it can be found in diverse academic written genres (e.g., case studies, research articles, and business proposals), 2) it also includes major writing techniques such as presenting and interpreting experimental data and statistical evidence in visual elements (e.g., images and diagrams), which student writers need to acquire, and 3) recent studies have demonstrated that visual renderings, articulating, and formulating focal points in a multimodal context such as a data commentary can be an arduous undertaking to be experienced and written by numerous student writers and may present a challenge to both proficient writer students and teachers in science fields (Eriksson & Nordrum, 2018; Jalilifar et al., 2019; Nordrum & Eriksson, 2015; Sancho Guinda, 2012; Wharton, 2012, to name but a few). Therefore, the aims of the present study are (a) to compare PCFs and the rhetorical functions produced by Iranian undergraduate and graduate students, and more importantly (b) to analyze the relationship between the PCFs and the rhetorical functions in data commentary.

LITERATURE REVIEW

The Importance of PCFs in Academic Writing

Phrasal complexity features have recently been included as an important subset of grammatical complexity features in many studies in L2 writing and academic writing (e.g., Staples et al., 2016; Parviz et al., 2020). The increasing studies on the PCFs emanate from two important roles they play in writing research. First, PCFs contribute to the expansive representation of grammatical complexity in writing research. In writing studies, scholars argued the integration of PCFs into the analysis of grammatical complexity. Previously, T-units-based measures[1] have been dominated in applied linguistics to assess grammatical complexity of academic writing (Biber et al., 2011). Then, scholars (e.g., Biber & Gray, 2010; Staples et al., 2016) expressed their concerns over analyzing L2 and academic writing with only T-unit-based measures, and this resulted in a call for integrating phrasal complexity[2] into the analysis of grammatical complexity. As the seminal work on phrasal complexity, Biber et al. (2011) proposed a developmental index of writing complexity features, operationalizing phrasal complexity into a set of phrasal features that functions as noun modifiers (e.g., attributive adjectives, pre-modifying nouns, and prepositional phrase). The integration of these PCFs, which represents phrasal complexity, supplements existing measures/features of complexity at the clausal level to make the representation of grammatical complexity more complete in L2 writing research.

Second, the importance of PCFs in academic writing also arises from their close relationship with writing development. Empirical studies have demonstrated that PCFs are often used to create compressed NPs, which is a grammatical characteristic of indicating the advanced stage of writing development (Biber et al., 2011). Scholars conducted large-scale corpus-based studies to show that in academic written registers, there are co-occurring patterns of PCFs to construct compressed NPs (Biber, 1988). Then, researchers added a functional interpretation to the co-occurrence: the PCFs fulfill an important function, packing intensive information in a compact space (Parkinson & Musgrave, 2014). For example, the NP, "*a solid research plan on the topic of grammatical complexity in second language writing*", only includes PCFs, for instance attributive adjectives (e.g., solid) and prepositional phrases as postmodifiers (e.g., of grammatical complexity). The NP carries rich information to expand the meaning of the head noun (i.e., *plan*) in a compressed structure. Researchers claim that when stu-

[1] A T-unit refers to a main clause and all dependent clauses attached to the main clause, which only capture grammatical complexity based on subordinate clauses.

[2] Phrasal complexity primarily refers to compressed noun phrases (NPs), which is the grammatical structure that head nouns are modified by phrasal features, such as premodifying nouns and prepositional phrases (Biber & Gray, 2010).

dents become academically advanced, they tend to use more PCFs, which indicates their writing development (Staples et al., 2016).

Because of the importance of PCFs, there has been a growing research trend of this grammatical structure in L2 and academic writing from 2011 to present. Based on Biber et al. (2011), many relevant studies were conducted in diverse academic written genres, including but not limited to EAP essays (Parkinson & Musgrave, 2014), research articles (Parviz et al., 2020) argumentative essays (Lan & Sun, 2019), PhD dissertations (Ansarifar et al., 2018), first-year composition (Staples & Reppen, 2016), and diverse genres in British Academic Written English Corpus (Staples et al., 2016).

Rhetorical Functions[3] in Academic Writing

In discourse analysis, rhetorical functions have been considered a key component of academic writing. Over the past three decades, rhetorical functions within the domain of genre analysis have captured the attention of genre analysts and rhetoricians, many of which have been based on Swales' seminal work. Swales (1990) provided a widely used analytical framework to analyze discourse structure of a genre, including rhetorical moves and linguistic features. Swales (1990) also discussed communicative moves, which have generated many empirical studies on varied genres, in which academic written genres are a major part. These academic genres include but not limited to textbooks (e.g., Hyland, 2000), theses and dissertations (e.g., Samraj, 2008), research articles (e.g., Yang & Allison, 2003; Parviz, 2023), among others. These studies have shared profound insights into the schematic patterns, discoursal features, and linguistic-textual mechanisms deployed in diverse academic disciplines. The analyses of rhetorical functions, furthermore, provide extensive implications for the teaching and learning of L2 writing. The successful use of rhetorical functions has been regarded as a characteristic of "good-standard" academic and L2 writing.

Furthermore, grammatical features play an important role in the analyses of rhetorical functions of academic writing. They serve as a medium for achieving specific rhetorical functions, making it critical to build a connection between rhetorical functions and grammatical options to enhance L2 students' learning of academic writing (Charles, 2007). While many studies have attempted to build this connection, only a few have examined the connection between rhetorical functions and NPs and PCFs in academic writing. For example, Jiang and Hyland (2015) conducted an analysis on shell nouns to analyze how this structure is used in research articles to express stances across multiple disciplines. More recently, Jiang and Hyland (2021) compared patterns of metadiscursive nouns, which are often used to express viewpoints on research and to interact with intended audiences as members of disciplinary communities.

In register analysis, researchers have applied a register-functional approach to study both grammatical forms and their rhetorical functions based on large-scale corpora (Biber et al., 2022). For instance, Staples and Reppen (2016) compared grammatical complexity features in two academic genres such as argumentative papers and rhetorical analysis. The rhetorical functions of two PCFs (attributive adjectives and premodifying nouns) were interpreted in their study. They found that more premodifying nouns are used in argumentative papers, "reflecting the greater complexity of topics and relationships being expressed, as well as the need to provide more informational support than in the rhetorical analysis" (Staples & Reppen, 2016, p. 28-29). Staples et al. (2016) also explored grammatical complexity (including PCFs) in diverse written genres across four institutional levels in the university setting. They noted that writers at higher institutional levels used more PCFs to meet "increased information packaging demands as they gain disciplinary content knowledge and are expected to convey this knowledge concisely for their audiences" (p. 77).

Research Gaps and Research Questions

This study was conducted in a relatively new genre of data commentary, which has not been explored substantially in terms of phrasal complexity features, rhetorical functions, and their relationships. In this study, we analyzed the use of phrasal complexity features and the rhetorical functions and their relationships in data commentaries written by Iranian undergraduate and graduate students.

Data commentary is selected as the academic written genre due to two specific reasons. First, it has been considered a challenging genre to not only L2 students but also university students in general (Eriksson & Nordrum, 2018; Nordrum & Erikson, 2015). The primary reason for this may be the multimodal nature of data commentary via which writers need to incorporate both textual and visual information in academic and scientific texts (Swales & Feak, 1997). Second, it could be considered a common genre across multiple written genres in university settings. A number of academic written genres across various disciplines can include data commentary, such as case analyses in economics, scientific reports in chemistry, and research articles in applied linguistics. Generally, data commentary is commonly organized in a general-to-specific pattern, with the overall structure classified into location elements and/or summary statements, as well as highlighting statements, discussions of implications,

[3] In this article, we used the terms "rhetorical functions", "rhetorical moves,» «communicative moves,» and «moves and steps» interchangeably to refer to the various strategic techniques that speakers use to achieve their communicative goals. While each term conveys a slightly different nuance, they all refer to similar concepts.

problems, exceptions, recommendations, or other notable aspects of the data (Swales & Feak, 2012).

Locatives, considered an important element, are typically employed as opening statements in data commentary. Their purpose is to indicate the position of the data, providing a succinct overview of findings through visual information. Locatives direct readers' attention to the visual display of information before proceeding to the main text (Swales & Feak, 2012). Following are examples of locatives (*italicized in examples a & b*) which can elucidate how this overarching element is linguistically employed.

> *Table 5 shows the types of internet misbehavior common among university students.*
>
> *Figure 2 shows a honeycomb solid oxide fuel cell (SOFC) unit with air cooling paths.*
>
> *(Adapted from Swales & Feak, 2012, p. 147)*

Location elements, also known as "endophoric markers", play a crucial role in assisting readers in identifying important information presented in visual materials such as tables, figures, charts, and diagrams (Hyland, 2019). As a type of metadiscourse resource, endophorics can also direct the reader's attention to explanatory and interconnected sections of a text. They are often positioned at the end of a sentence using passive voice, as exemplified below (Hyland, 2019; Swales & Feak, 2012).

> *a. The types of internet misbehavior common among university students are shown in Table 4*
>
> *b. A honeycomb solid oxide fuel cell (SOFC) unit with air cooling paths is shown in Figure 2.*
>
> *(Adapted from Swales & Feak, 2012, p. 147)*

Another integral aspect of a data commentary is the use of highlighting statements. These statements are used to emphasize key points that are supported by the data (Swales & Feak, 2012). Highlighting statements also provide opportunities to demonstrate the ability to recognize "trends or regularities in the data, to separate more important findings from less important ones, and to make claims of appropriate strength" (Swales & Feak, 2012, p. 156). In the process of writing highlighting statements, it is crucial to exercise "caution and critical thinking towards the data" while effectively employing appropriate linguistic devices to express this cautious approach (Swales & Feak, 2012, p. 156). However, the concluding section of a data commentary can be challenging, as it requires presenting significant insights derived from the data, distinguishing experienced writers from novices (Swales & Feak, 2012). To tackle this challenge and demonstrate expertise, Swales and Feak (2012, p. 172) proposed incorporating certain elements in the conclusion of a data commentary such as "explanations and/or implications of the data and explanation of the reasoning process that led to the conclusions". Overall, Swales and Feak (2012, pp. 140-141) listed some of the more common purposes of data commentary as shown in Table 1. They highlighted that while many of the aforementioned objectives are typically addressed in a data commentary, it can be challenging to precisely specify the requirements of this genre. However, despite its importance, this genre has receive limited research attention regarding the analysis of PCFs and rhetorical functions.

Given the widespread presence of data commentary across disciplines, there is a need for further research to explore and examine the use of PCFs within this under-represented genre. By investigating the presence and functions of PCFs within data commentary, scholars can uncover valuable insights into how these linguistic features contribute to the overall rhetorical effectiveness of written works in different academic domains. This research would greatly benefit our understanding of language usage and provide guidance for students and academics in mastering the skills required for effective data commentary writing in their respective fields. In addition, it is important to mention that individual PCFs and their corresponding rhetorical functions have been explored in some studies (e.g., Staples & Reppen, 2016). The analyses in such studies are often based on analysis of specific texts to interpret rhetorical functions of individual PCFs. While acknowledging the importance of these analyses, it is also meaningful to detect the connection between PCFs and rhetorical functions from a quantitative perspective to triangulate the existing findings. In this study, we quantified the relationship between the frequencies of PCFs and the

Table 1

Common Purposes of Data Commentary Proposed by Swales and Feak (2012)

Highlight the results of research
Use the data to support a point or make an argument in your paper
Assess theory, common beliefs, or general practice in light of the given data
Compare and evaluate different sets
Assess the reliability of the data in terms of the methodology that produced it
Discuss the implications of the data
Make recommendations

Note. From *"Academic Writing for Graduate Students: Essential Tasks and Skills,"* by J. M. Swales & C. B. Feak, 2012, University of Michigan Press. Copyright 2012 by University of Michigan Press.

frequencies of rhetorical functions (i.e., moves and steps). Recently, this method has been applied in academic writing research. For instance, Lu et al. (2021b) analyzed the relationship between phraseological features and rhetorical functions in research articles. Consequently, our study would supplement the existing analyses from the qualitative analysis on PCFs and their rhetorical functions, presenting the importance of PCFs in the under-explored genre, data commentary.

Finally, we conducted a comparison between the data commentaries produced by undergraduate and graduate students in order to uncover potential linguistic similarities and differences between the two groups. This comparative analysis aimed to shed light on how these groups employ phrasal complexity features and rhetorical functions in their writing. Understanding this comparison is crucial because all student writers, regardless of their academic level, rely on specific linguistic and rhetorical strategies or writing patterns to achieve their writing goals (Lavelle & Bushrow, 2007). Given that undergraduate students are typically in the nascent stages of their academic careers, and they may encounter data commentaries for the first time, it is essential to understand their linguistic and rhetorical approaches in order to provide the necessary guidance and assistance tailored to their needs. On the other hand, graduate students face the complex and often novel task of academic writing at the graduate level (Lavelle & Bushrow, 2007). With their prior experience gained during their undergraduate studies, these students are expected to possess advanced writing skills. By undertaking this comparative analysis, our goal is to shed light on how undergraduate and graduate students utilize phrasal complexity features and rhetorical functions in their data commentaries. This knowledge may empower us to provide targeted support and appropriate resources to both groups, assisting them in improving their writing skills and excelling in their respective academic pursuits. Additionally, this analysis offers valuable insights into academic writing development, allowing for the assessment of proficiency levels, identification of developmental patterns, examination of genre conventions, and contributing to the advancement of writing pedagogy. Accordingly, the following research questions are formulated:

(1) How do undergraduate and graduate students utilize PCFs in data commentaries they produce? Which specific types of PCFs are more prominently used by each group?

(2) How do undergraduate and graduate students employ rhetorical functions in data commentaries they produce? Which specific types of rhetorical functions are more prominently used by each group?

(3) Is there a correlation between the presence of PCFs and the occurrence of rhetorical functions within both the undergraduate and graduate data commentaries?

METHODS

Participants and Context

The present study was conducted at an Iranian state-run university during the winter semester of 2019-2020, which coincided with a severe outbreak of Covid-19 pandemic. Due to the sudden shift from traditional classrooms to on-line education compelled by the pandemic, the authors of this study were required to adjust all their academic activities and face-to-face classes to on-line mode of delivery. Given the challenging circumstances, participant selection for this study was conducted using a convenient sampling method. The researchers selected participants from three existing intact groups, with the first author of the study serving as the instructor for these groups. It is important to emphasize that random assignment or manipulation of group compositions was not employed during the participant selection process.

In this study, a total of 76 students participated, consisting of 47 undergraduate students and 29 graduate students. Among the graduate students, there were 23 MA students and 6 PhD students. The participants included 51 female students and 25 male students, all of whom were native speakers of Persian majoring in English Language Teaching (ELT). The selected participants were enrolled in different writing courses specifically tailored to their academic level. The courses included “Essay Writing” for undergraduate students, “Academic Writing” for MA students, and “Advanced Academic Writing” for PhD students. These courses, which were offered by the English department of the university, were two-credit online writing courses that spanned over the course of the winter semester, with a total of 16 sessions. These courses were conducted online, with each group having one class per week. Each session lasted approximately 90 minutes.

It is also essential to consider the English language proficiency of the participants and the admission process for Iranian students in undergraduate and graduate studies. Iranian students undergo a rigorous and highly competitive university entrance examination (UEE) to gain admission to BA and MA programs. This national examination evaluates their content knowledge as well as their language skills, including reading, grammar, and vocabulary. Admission to doctoral programs also involves meeting strict academic standards, which include performing well in the UEE and evaluating their research achievements during their master’s studies. Therefore, all participants have passed the Iranian nationwide BA, MA, and PhD university entrance examinations, which serve as gatekeeping tests for selection purposes.

Based on their academic background and experiences, it is reasonable to assume that they possess a high command of academic English. Detailed demographic information about the participating students is presented in Table 2.

Furthermore, it should be noted that both undergraduate and graduate students have already completed extensive academic coursework. They have taken various prerequisite writing courses during their undergraduate and graduate studies such as "Introduction to Writing," "Paragraph Development," "Letter Writing," "Essay Writing," and "Advanced Writing." However, the participants' knowledge and experience in writing data commentary may vary, primarily because the genre of data commentary is not commonly taught at the undergraduate level. As a result, undergraduate students (BA) might have prior experience with this type of writing. In contrast, graduate students (MA and PhD), given their advanced academic training, were expected to have a higher level of familiarity and experience in writing data commentary. It is important to acknowledge that the differing levels of experience between undergraduate and graduate students could potentially influence their approach and proficiency in writing data commentary. However, this difference was not considered or analyzed in the current study.

Task

In order to examine phrasal complexity features and rhetorical functions in the data commentaries, a set of 30 suggested topics emphasizing data commentary was initially chosen. These topics were sourced from publicly accessible IELTS materials associated with the IELTS Academic Module writing Task 1 (Cullen et al., 2014), which are seen highly relevant to data commentary writing. The visual nature of this task also aligns with academic writing components, as its discourse modes often mirror those found in authentic university assignments and "reflect some of the features of academic language" (IELTS, 2017)[4]. To ensure impartiality in the writing prompts, we then carefully selected 20 topics that steer clear of sensitive issues related to politics, religion, race, culture, and controversy (samples provided in Appendix A). The visual materials accompanying these prompts were chosen for their familiarity and relevance to daily life, eliminating the need for specialized knowledge.

Procedure

To maintain ethical standards, participants were given a clear explanation of the objectives of the study at the outset. Prior to their involvement, participants were required to provide verbal consent, signifying their willingness to participate and ensuring their rights as subjects were respected. To ensure the research process was feasible and aligned with the participants' busy academic schedules, a pre-selected set of 20 visual materials was utilized. These materials were integrated as in-class activities during eight separate online sessions of the writing course over an eight-week period. This approach aimed to strike a balance between accommodating the participants' demanding academic commitments and maintaining the practicality of the research study. The visual materials included bar/pie charts, diagrams, and line graphs, covering various general topics such as "*adult education, global illiteracy rates, leisure time, food budgets, mobile phones, building construction, cinema attendance, London museums, public transportation, and work performance*". Participants were then required to individually write about their preferred topic in the form of a data commentary, with a minimum length of 150 words and a time limit of 30 minutes. To prevent topic disclosure, participants were not informed beforehand of the specific topic for each session, and it was assumed that the selected visual materials were of general interest and familiar to all participants, aligning closely with their everyday reality (Sancho Guinda, 2012). No participant was allowed to work collectively or in pairs in order to ensure an accurate gauge of how they exploited phrasal complexity features and rhetorical functions in the data commentaries. The commentary writing tasks were presented without any supporting background materials, enabling participants to engage in spontaneous writing without requiring any field-specific knowledge. These data commentary writing tasks were also written without any direct explicit instruction on phrasal complexity features and rhetorical instructions and were gathered as part of the coursework, with no impact on participants' final assess-

Table. 2

Demographics of the Participating Students

Educational Levels	No of the Participating Students	Gender		Total
		Male	Female	
BA	47	12	35	76
Master's	23	10	13	
PhD	6	3	3	

[4] IELTS (2017). IELTS website. Retrieved from: https://www.ielts.org.

ments. After collecting the data commentaries produced by undergraduate and graduate students during the sessions (as shown in Table 3), a corpus was constructed consisting of 380 files. The total number of tokens in the corpus was 72,591 words with an average length of 191.03 tokens. The corpus included two sub-corpora, one for undergraduate students (206 files), and another for graduate students (174 files). It is worth noting that not all graduate students were able to complete all tasks, as some individuals were unable to attend all online sessions of the course.

Linguistic Model of PCFs

The PCFs in the study are derived from the hypothesized index of writing complexity features proposed by Biber et al. (2011). These PCFs are presented in Table 4. The grammatical functions of these PCFs are primarily noun modifiers, except for nominalizations, which can be either head nouns or noun modifiers. The PCFs consist of two pre-noun modifiers (i.e., attributive adjectives and premodifying nouns) and three post-noun modifiers (e.g., prepositional phrases (*of*), prepositional phrases (*others*), and appositive noun phrases). The noun modifiers are at different stages of Biber et al.'s (2011) hypothesized stages of writing development: attributive adjectives and premodifying nouns are at the stages of 2 and 3, respectively, whereas the two types of preposition phrases and appositive noun phrases are at the stages of 4 and 5. Nominalizations are not placed in any stages.

Corpus Processing and PCFs Extraction

The corpus processing involved five steps. First, the original files in the corpus were converted into plain text format using the *AntFileConveter*. Second, the converted files were tagged with part-of-speech (POS) information via the *TagAnt*. The PCFs are based on three core linguistic features: nouns, adjectives, and prepositions. A qualitative check was conducted on precision and recall of the three features to ensure the accuracy of the tags. The rates of precision and recall were all above 90% for the three core linguistic features. Thus, we did not fix the POS tags of the three linguistic features in all files (see Appendix B).

Third, *AntConc* was employed to extract the potential instances of the PCFs based on regular expressions (regex). The regex patterns were built based on the following linguistic sequences:

- "adjective + noun" pattern to identify attributive adjectives
- "noun + noun" pattern to identify premodifying nouns
- "noun + preposition" pattern to identify prepositional phrases as postmodifiers (including both *of*-prepositional phrases, and *other* prepositional phrases)
- "a string + nominal suffixes" pattern to identify nominalizations

Table 3

Description of the Corpus

Corpora	File No.	Total Tokens	Mean Length
Undergraduate (**UG**)	206	42,678	207.17
Graduate(**GR**)	174	29,913	171.91
Total	380	72,591	191.02

Table 4

Linguistic Model of PCFs

Stages	PCFs	Position	Examples
1	Attributive adjectives	Pre-noun	[Fundamental] questions
2	Premodifying nouns	Pre-noun	[Research] methods
N/A	Nominalizations	N/A	[Industrialization]
4	Prepositional phrases (*of*) as post-noun modifiers	Post-noun	The analytical approach [of discourse analysis]
4	Prepositional phrases (*others*) as post-noun modifiers	Post-noun	The methodological diversity [in applied linguistics]
5	Appositive noun phrases	Post-noun	A common method, [corpus-based approach], has been ...

– "words in a pair of punctuations" pattern to identify appositive NPs (i.e., parentheses, square bracket, dashes, and commas).

AntConc utilized these regex patterns to extract relevant cases of the PCFs from the corpus. There are two noteworthy points: (1) the list of suffixes to extract nominalizations was based on five high-frequency nominal suffixes in academic registers in Biber et al. (1999)[5], including "-ment", "-ity", "-tion", "-ness", and "-ance", "-ship"; (2) the patterns to extract appositive noun phrases were based on how this phrasal feature was extracted in related previous studies (Lan & Sun, 2019). Then, the extracted potential cases of the PCFs were saved in an Excel file for the manual adjustment.

Fourth, a manual adjustment was conducted. The extracted potential cases in the Excel file were not all accurate cases of the PCFs. In Example1, the second case (e.g., to have their **meal with** their friends) is a "noun-preposition" sequence but the prepositional phrases (i.e., with their friends) are not to modify the head noun (i.e., meal).

To ensure the accuracy of the data, two researchers conducted manual checks on all instances in the Excel file and removed any inaccurate cases of PCFs. During the pilot coding phase, the researchers collaboratively examined 100 instances of each PCF to establish inter-coder reliability, which yielded a high agreement rate of 93.7%. Following this, the two researchers independently examined the remaining instances of PCFs. Given that a larger number of inconsistent instances were found in the prepositional phrases (other) category during the pilot phase, extra attention was given to ensure the accuracy of this specific PCF. In addition, two PCFs (i.e., attributive adjectives and premodifying nouns) might occur more than one time in a case, such as coordinate attributive adjectives (e.g., a valid and reliable method) or multiple noun sequences (e.g., a corpus research method). These cases were marked in the Excel file and were used to adjust frequencies of the PCFs in the next step.

Fifth, a Python program was applied to count the frequencies of the PCFs based on the Excel file with all the adjusted cases of PCFs. The program output is a dataset with filenames and all the frequencies of the PCFs associated with specific files. We then adjusted the frequencies based on the marked cases of coordinate attributive adjectives and multiple noun sequences in the Excel file. By the end of this step, we considered that the frequencies of the PCFs were accurate for statistical analysis.

Coding of Rhetorical Functions

The schematic structures were analyzed via a multi-step process. First, the researchers examined relevant studies in which textual move organization of data commentary were presented (e.g., Eriksson & Nordrum, 2018; Jalilifar et al., 2019; Nordrum & Eriksson, 2015; Sancho Guinda, 2012; Swales & Feak, 2012). Move-step schemas, along with illustrative examples and functional language representative of rhetorical structures of previous research, were then selected, extracted, and saved in a separate single file. This facilitated adoption and adaption of possible functional labels/ names for the moves and steps in our data analysis. Finally, following previous recommendations from the leading researchers (e.g., Biber et al., 2007), the researchers rigorously scrutinized the entire datasets to achieve a clear understanding of the texts' structure, patterns, communication functions, and linguistic signals.

Drawing on seminal text analytical methods (Biber et al., 2007; Monero & Swales, 2018), we assigned discourse roles to various (sub) move types based on function-form relations. While *function* is realized by clause or sentence, *form* is realized by lexico-grammatical constituents in text segments (Monero & Swales, 2018). By utilizing this approach, we aimed to gain a comprehensive understanding of how different linguistic elements contribute to the overall structure and organization of the discourse.

The viewpoint of communicative purposes of move as well as linguistic clues were also the central notion for our analysis. By definition, move is "a discoursal or rhetorical unit in a text that performs a coherent and distinctive communication function in written or spoken discourse" (Swales, 2004, pp. 228-229). On the other hand, moves can be the outcome of the hybrid of multiple elements or sub-moves (steps) which are recognized both rhetorically and linguistically. Steps are, therefore, the many text fragments that "together, or in some combination, realize the move" in such a way that "the steps of a move primarily function to achieve

Example 1

Case	Pre-context	Keyword in context	File
1	...decade_NN (_(1980_CD)_) only_RB five_CD	**percent_NN of_IN** people_NNS spent_VVD...	UG1
2	...more_RBR willing_JJ to_TO have_VH their_PP$	**meal_NN with_IN** their_PP$ friends_NNS...	UG1

[5] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. New York: Longman.

the purpose of the move to which it belongs" (Biber et al., 2007, p. 24).

The researchers followed the general steps outlined in Biber et al. (2007) to conduct a pilot analysis in order to refine the coding schema and minimize the potential risk of human errors. To achieve this, two rounds of pilot coding were performed on a subset comprising 25% of the corpus (*n*=95). The first round included 5% of the files (*n*=19), while the second round encompassed 20% of the files (*n*=76). Each of the researchers independently coded the texts, annotating them in the margins. Following each round, the researchers engaged in discussions to address discrepancies and ensure agreement on the rhetorical interpretations of the coded text segments. This process continued until both researchers reached a consensus. To assess inter-coder reliability, Cohen's Kappa coefficient was calculated, resulting in a Kappa coefficient of 0.865, indicating almost perfect agreement between the coders.

Considering our pilot coding, we developed a coding scheme as the benchmark for the coding of moves and steps for the remaining 75% of the corpus (see Table 5). The coding scheme for the data commentary comprised four rhetorical moves and certain rhetorical steps associated with the moves, as illustrated in Table 5. The only exception was that Move 4 (i.e., concluding visual information) has no associated steps identified. The detailed description of the coding scheme of rhetorical functions and examples taken from the datasets are presented in Appendix C.

Statistical Analysis of PCFs and Rhetorical Functions

Statistical analyses were conducted using SPSS. The frequencies of the PCFs and the counts of the rhetorical functions (moves and/or steps) were firstly normalized to 100 words. Then, the normality of the PCFs and rhetorical moves/steps were examined based on the Shapiro-Wilk test (see Appendix D). The test results revealed that the PCFs and moves/steps did not follow a normal distribution, except for attributive adjectives. To answer the first and second research questions, Mann Whitney U tests were employed to explore the differences in phrasal complexity features and rhetorical functions between the undergraduate and graduate data commentaries, respectively. After that, a Spearman correlation test was utilized to explore the relationship between phrasal complexity features and rhetorical moves/steps, generating a correlation matrix to answer the third research question.

RESULTS AND DISCUSSION

RQ1: How do undergraduate and graduate students utilize PCFs in data commentaries they produce? Which specific types of PCFs are more prominently used by each group?

To answer the first research question, the results of the Mann-Whitney U Test showed that four out of six PCFs had significant differences on the normalized frequencies between the undergraduate corpus and the graduate corpus. These four phrasal complexity features were attributive adjectives, nominalizations, noun-noun sequences, and prepositional phrases (of). Table 6 also presents the p values for these four PCFs (e.g., 0.008 for nominalizations, 0.000 for noun-noun sequences). Based on the comparison of the mean ranks of these four PCFs, we categorized the four PCFs into two groups. Group1 consisted of three PCFs (i.e., attributive adjectives, nominalizations, and prepositional phrases (of)) that were found to be significantly more prevalent in the corpus of graduate data commentaries. On the other hand, Group 2 only included one PCF, noun-noun sequences, which were used significantly more in the corpus of undergraduate data commentaries.

In terms of the effect size of the six PCFs, the noun-noun sequences had the largest eta^2 as 0.172. However, these

Table 5

Rhetorical Functions Found in Data Commentary

Move 1: Presenting Visual Information

Step 1: Providing an explanatory note to set the scene

Step 2: Indicating the location of the data

Move 2: Highlighting Visual Information; Comparing and Contrasting Key Points

Step 1: Describing the facts (with/without providing statistical evidence)

Move 3: Commenting on Visual Information

Step 1: Personal asides

Move 4: Concluding Visual Information

Table 6

Mann-Whitney U Test on the PCFs

Linguistic Features	P Values	Effective Size (eta2)	Mean Rank (UG)	Mean Rank (GR)	Description
Attributive adjectives	0.010*	0.017	176.351	205.252	GR > UG
Nominalizations	0.008*	0.007	175.920	205.773	GR > UG
Noun-noun sequences	0.000*	0.172	230.893	139.924	GR < UG
Prepositional phrases (*of*)	0.000*	0.019	162.308	222.066	GR > UG
Prepositional phrases (*other*)	0.869	0.074	188.652	190.514	NA
Appositive NPs	0.094	0.000	184.572	195.401	NA

Note. The test is based on a two-tailed assumption with 0.05 as the alpha level. "*" marks the PCFs with significant difference between the two groups. The sum of the effect size for the PCFs is small (eta2 =0.290).

suggested a small effect size[6]. The remaining PCFs also had a small effect size such as attributive adjectives (eta^2 = 0.017) and nominalizations (eta^2 = 0.007). The total cumulative effect size of the PCFs was also small with the eta^2 = 0.290. This suggested that the six PCFs can only explain 29% of the variance of the data commentary difference between undergraduate and graduate students. Given that writing differences can arise from a number of factors (e.g., linguistic features beyond the PCFs), we concluded that it is not unexpected to get small effects for the PCFs, because only PCFs are not supposed to explain a large portion of the variance of the writing difference in our research context.

The results further suggested that Iranian graduate students produced three types of PCFs (i.e., attributive adjectives, nominalizations, and prepositional phrases (*of*) as post-noun modifiers) much more but noun-noun sequences much less than their undergraduate counterparts. In terms of nominalizations, few studies have included such phrasal complexity features in recent studies on grammatical complexity. The only study is Staples et al. (2016), who presented a similar finding that L1 English students used nominalizations in their writing when they become academically advanced. For attributive adjectives, Staples et al. (2016) found that L1 graduate students used these linguistic features more than their undergraduate counterparts in academic writing, whereas Parkinson and Musgrave (2014) found that EAP students heavily relied on attributive adjectives in their writing. Our finding is similar to Staples et al. (2016).

In the context of data commentary, it was found that Iranian graduate students used more attributive adjectives than undergraduate students. Qualitative analysis indicated that undergraduate students employed a limited range of attributive adjectives in their data commentary prompts (e.g., public transports, selected countries); However, graduate students used a broader usage of attributive adjectives beyond those provided in the prompts and related to the description of data (e.g., "*the highest illiteracy percentage*", "*the different degrees of the effects*"). Thus, it can be concluded that graduate students exhibited a more extensive and proficient use of attributive adjectives in their data commentary.

Concerning noun-noun sequences, Ansarifar et al. (2018) reported that Iranian PhD students used these linguistic features more frequently in their dissertations compared to MA students in their theses. However, our findings reveal a contrasting pattern, indicating that Iranian graduate students employed noun-noun sequences less frequently than their undergraduate counterparts. The higher frequency of noun-nouns sequences in undergraduate data commentaries can be primarily attributed to topic influence, too. This is especially evident in the case for proper nouns such as "*London Museum*" and certain formulaic sequences such as "*mobile phones*" and "*food budget*", because undergraduate students relied on using them in their data commentary. In contrast, graduate students tend to paraphrase these formulaic sequences, aiming to avoid repetitive use of the same lexical chunks. For instance, they rephrased "*food budget*" as "*the budget on their food* " or "*the budget allocated for food* ".

As for prepositional phrases (*of*), Parkinson and Musgrave (2014) found no significant difference between L2 undergraduate students and MA TESOL students. Our finding showed that Iranian undergraduate students used such linguistic features more frequently than Iranian graduate students. This can be explained by the prevalence basic formulaic sequences in undergraduate data commentaries. When a noun is followed by a prepositional phrase (*of*), this construction primarily functions as a post-noun modifier as in "*the result of a survey*" and "*the plan of our research project*" (Lan & Sun, 2019). However, there are situations where L2 students may not be fully aware of their use of this noun modifier when writing. For example, two common and basic formulaic sequences associated with describing quantity,

[6] The interpretation of the effect size is based on Cohen (1988), with eta2 less than 0.300 suggesting a small effect, between 0.300 and 0.500 suggesting a medium effect, and greater than 0.500 suggesting a large effect.

namely, "*a lot of* " and "*the number of* " can be found in undergraduate data commentaries. In contrast, this is not the case in graduate data commentaries. In fact, Iranian graduate students show a tendency to employ a variety of prepositional phrases (*of*), such as "*the cost of education*" and "*the survey of adult education*".

Last but not least, it is important to emphasize that a greater use of a grammatical feature does not always suggest a greater writing development (Staples & Reppen, 2016). Several relevant studies have demonstrated that the repetition of formulaic sequences in writing does not necessarily indicate a greater writing development (Lan & Sun, 2019; Staples & Reppen, 2016). This is particularly the case in our study. Although undergraduate students used more noun-noun sequences and prepositional phrases (*of*), we cannot concluded that their writing development is greater than that of the graduate students. In contrast, graduate students produced a greater diversity of noun-noun sequences and prepositional phrases (*of*), and had less reliance on formulaic sequences when writing about various topics. This suggested a stronger proficiency in utilizing noun modifiers.

RQ 2: How do undergraduate and graduate students employ rhetorical functions in data commentaries they produce? Which specific types of rhetorical functions are more prominently used by each group?

In order to answer the second research question, we conducted a rank-based non-parametric test (Mann-Whitney U) to measure the significance of differences between the variables. The results indicated that graduate students used significantly more M1 (i.e., presenting visual information), M1S2 (i.e., indicating the location of data), and M2 (i.e., highlighting visual information; comparing and contrasting key points) than undergraduate students. Conversely, undergraduate students used significantly more M3 (i.e., commenting on visual information) and M3S1(i.e., personal asides) than graduate students.

Table 7 also presents the *p* values for the five rhetorical moves and steps, for example 0.000 for M1 (i.e., presenting visual information), 0.000 for M1S2 (i.e., indicating the location of data), and 0.000 for M2 (i.e., highlighting visual information). Based on the comparison of the mean ranks of these rhetorical functions, we categorized them into two groups: (a) Iranian graduate students significantly used more M1, M1S2 and M2; (b) undergraduate students significantly used more M3 and M3S. The two groups suggested a preference of rhetorical functions: graduate students included more moves and/or steps related to presenting and interpreting data while undergraduate students included more moves and/or steps related to expressing their personal opinions. However, based on the statistical interpretation of "M4", there was no significant difference observed in the use of M4 between undergraduate and graduate data commentaries. This suggested that both groups tended to employ M4 in a similar manner.

Regarding effect size, we wish to discuss the rhetorical functions included in the analysis, M1, M2, M3, and M4. Among them, M1 had the largest eta^2 as 0.156, followed by M2 eta^2 as 0.089, indicating small effect size for both rhetorical functions. M3 and M4 also had small effect size in our analysis (eta^2 = 0.002 and eta^2 = 0.000), respectively. Taken together, the cumulative effect size of the four rhetorical functions can be considered medium, with the eta^2 = 0.466. This suggested that the four functions can explain around 46% of the variance of the data commentary differences between undergraduate and graduate students. While other factors may also contribute to differences in data commentaries, we concluded that the four rhetorical functions can explain fairly large portion of the variance in our research context.

Table 7

Mann-Whitney U Test on the Rhetorical Functions

Rhetorical Functions	*P* Values	Effective Size (eta^2)	Mean Rank (UG)	Mean Rank (GR)	Description
M1	0.000*	0.156	150.330	236.420	GR > UG
M1S1	0.245	0.004	194.709	183.261	NA
M1S2	0.000*	0.100	158.788	226.282	GR > UG
M2	0.000*	0.089	159.699	225.192	GR > UG
M2S1	0.875	0.000	188.577	190.604	NA
M2S2	0.000*	0.092	176.000	205.670	GR > UG
M3	0.006*	0.020	200.266	176.604	GR < UG
M3S1	0.000*	0.129	216.524	157.133	GR < UG
M4	0.987	0.000	189.431	189.580	NA

Note. *A Bonferroni adjustment was applied to the alpha value (i.e., 0.05), and the adjusted alpha value is 0.006. "*******" marks the moves/steps with significant difference between the two groups. The sum of the effect size for the four rhetorical functions (i.e., M1, M2, M3, M4) is medium (eta^2 = 0.466).*

These findings suggested the difference of using the four rhetorical functions had a medium practical influence, which deserve teaching attention in real classroom instruction.

Moreover, the results demonstrated that both undergraduate students and graduate students incorporated four moves and four steps to construct their data commentary. The occurrences of some of these moves and steps in both datasets may not only corroborate previous studies on data commentaries (e.g., Eriksson & Nordrum, 2018; Nordrum & Erikson, 2015; Swales & Feak, 2012) but it may also show that both groups conform to recognizable patterns of discourse organization of data commentaries.

Regarding the rhetorical functions, M1, M1S2, and M2 are highly favored by graduate students. M1 specifically serves the purpose of explicitly stating the objectives of visual content and is achieved through two distinct steps. The first step (i.e., M1S1, providing an explanatory note to set the scene) makes a brief note to begin reporting visual information that follows and apprises the readers of data that is going to be discussed, while the second step (M1S2) helps readers notice further salient information presented in non-text materials. As an opening tactic, M1 is commonly employed to commence the data commentary prior to discussing methods of highlighting visual information. This finding may indicate that data commentary typically begins with a general overview to enhance readers' understanding of the text before highlighting key points visually.

Known as an endophoric marker and a text organizer, M1S2 is utilized to demonstrate the position of the non-verbal information and draw readers' attention to important information presented in visual modes (Hyland, 2019; Swales & Feak, 2012). The greater propensity for employing M1S2 may suggest that the references to charts, tables, diagrams, and figures can form a preferred vehicle for drawing attention to a very brief summary of significant information that non-material texts show. The higher incidence of this rhetorical device could be, moreover, attributed to the directive force of endophoric markers in multimodal contexts. This finding echoes Sancho Guinda's (2012) conclusion that multimodal environments can influence endophorics counts, compared to simply verbal texts. On the whole, the high rate of occurrences of M1S2 may show how student writers establish metadiscoursal links between text constituents, argument, and target reader (see Hyland, 2019). Nevertheless, these metadiscursive items (i.e., M1 and M1S2) can function as prefabricate expressions or ready-made language chunks (e.g., *the chart shows that*... and *according to the table...*) that help writers start their writing. This could explain why these devices outnumber other rhetorical choices.

Recognized as the backbone of data commentary writing, M2 is a core rhetorical device via which key points are compared, contrasted, and accentuated normally by relevant evidence, statistics, and examples. The greater propensity for this rhetorical choice among graduate students may reflect their heightened awareness of the nature, characteristics, and rhetorical conventions of the text type, enabling them to highlight and focus on important aspects in the data commentaries they produced. In sum, the increased preference for M1, M1S2, and M2 may not only indicate graduate students' appreciation of the importance of these cardinal rhetorical moves in data commentary but it also demonstrates the extent to which graduate writers are preoccupied with aiding their readers in directing, processing, and unfolding non-text elements. Thus, these findings can support earlier studies that similarly identified the rhetorical functions as the most common and essential moves in data commentaries (e.g., Eriksson & Nordrum, 2018; Nordrum & Erikson, 2015; Swales & Feak, 2012).

M3 (i.e., commenting on visual information) and M3S1 (i.e., personal asides) were highly preferred by undergraduate students, on the other hand. The greater exploitation of M3 may suggest they are far more likely to include subjective explanations, judgments, and comments on their data description. Another reason why this rhetorical option is preferred is that undergraduate students strove to establish their identities as knowledgeable individuals (Sancho Guinda, 2012). However, the presence of M3 is not in agreement with the common purposes and structure of data commentary writing stated by Swales and Feak (2012). Additionally, M3S1 was solely based on undergraduate students' personal judgments or explanations of the data visually presented and was not directly related to the main topic being discussed. M3S1 can also present an opportunity to the writers to address the target readers subjectively and straightforwardly by asides and interrupting the ongoing discussion to make a statement on what has been mentioned (Hyland, 2019). The presence of M3S1 might give rise to some cases of pragmatic infelicity or deviance from academic writing (Sancho Guinda, 2012); yet, it can function as "an important reader-oriented strategy" (Hyland, 2019) that helps writers/speakers accentuate personal viewpoints, determine the nature social relationships, and exhibit relatively innovative linguistic performance. However, not being a move common and cardinal to data commentary, M3S1 can reflect undergraduate students' overt stance-taking through which they express their stance openly via a description and an evaluation of data. Swales and Feak (2012) regard stance as an important element in academic writing since it enables writers to reveal not only what they think but also what they know. In all, this finding can present evidence for previous research in which M3S1 was found in data commentary (Sancho Guinda, 2012).

Finally, the low incidence of M4 by both groups may suggest that students struggle to provide a brief synopsis of the key information visually presented at the conclusion of their data commentaries. One possible explanation for this could be in line with Swales and Feak (2012, p. 172), who consider M4 to be a challenging segment that distinguishes proficient writers from less proficient ones, as it requires the ability to engage in "some original thinking". This further implies that the skill of producing a clear and concise visual data summary demands not only critical thinking skills but also the capacity to identify and prioritize relevant information effectively. Therefore, students who can effectively exploit M4 and present the data coherently and meaningfully are more likely to excel in writing data commentaries compared to those who struggle with this move.

RQ 3: Is there a correlation between the presence of PCFs and the occurrence of rhetorical functions within both the undergraduate and graduate data commentaries?

In order to answer the third research question, we applied a Spearman correlation matrix among the PCFs and the rhetorical moves/steps. This matrix helped us determine the potential relationship between the intended linguistic features and rhetorical functions. As shown in Table 8, the results indicated that there was a correlation between M1 and attributive adjectives (ρ= 0.223) as well as prepositional phrases (*of*) (ρ= 0.212). M1S2 was also correlated with attributive adjectives (ρ= 0.210). M2 was another rhetorical function correlated with attributive adjectives (ρ= 0.216) and prepositional phrases (*of*) (ρ= 0.214). Finally, there was a correlation between M3S1 with prepositional phrases (*of*) (ρ= - 0.285).

The results demonstrated a pattern of *head nouns + prepositional phrases (of)* for M1 (i.e., presenting visual information). Few studies have correlated the PCFs and rhetorical moves or steps, but two recent studies have revealed the relationship between rhetorical functions and phraseological features (i.e., p-frames), which is related to our study. For instance, Lu et al. (2021a) found that the frequent p-frames (*the NOUN of*) largely contributed to the presentation of research information in the introductions of research articles in social sciences. Lu et al. (2021b) suggested that this was particularly the case for research articles in applied linguistics. For example, some frequent p-frames associated with presenting research purposes were *the purpose of* and *the basis of*. In our study, *head nouns + prepositional phrases (of)* fulfilled the function of presenting visual information. This PCF pattern is frequently used to present visual information in data commentary, as in the examples: *the number/amount/quantity/percent/rate(s) of.* Despite the genre differences, our findings can be consistent with Lu et al. (2021a) and Lu et al. (2021b).

Moreover, there was a connection between M1S2 (i.e., indicating the location of the data) and attributive adjectives. Two possible interpretations can be given: First, although locatives generally present a set of data or figures without concentrating solely on any particular item, they can be embedded in a presentational move or step to focus on a comparison of categories, groups, and data or report a relationship between variables (Lim, 2011). This finding may echo Lim's (2011) interpretation that indicating the location of the data is characterized by certain linguistic features such as (comparative) adjectives even though our genre of study (data commentary writing) is different from that of Lim (2011). Our second interpretation is the student writers' data commentary might be influenced by the presence of attributive adjectives in the writing prompts/topics, as discussed already. However, Lim (2011, p. 738) considered the locatives as a *space-saving strategy* to present a relatively abridged report.

The results further revealed the pattern of *attributive adjectives + head nouns*. Few p-frames based on this pattern can be observed in Lu et al. (2021a) and Lu et al. (2021b) for the rhetorical function of presenting research information. This is contradictory to our finding. However, taking a close look at our corpus, we found that when presenting the visual information, it is unavoidable to use some keywords associated with the topics such as *secondary school* and *mobile phone*. As topic would have information on the pattern of

Table 8

Correlation Matrix between the PCFs and the Rhetorical Functions

Rhetorical functions	Attributive adjectives	Nominalization Forms	Noun-noun sequences	Prepositional phrases (*of*)	Prepositional phrases (*other*)	Appositive NPs
M1	**0.223**	0.081	-0.099	**0.212**	0.023	0.017
M1S1	-0.090	0.058	-0.008	-0.107	0.073	0.107
M1S2	**0.210**	0.039	-0.101	0.187	-0.009	-0.056
M2	**0.216**	-0.125	-0.060	**0.214**	-0.022	0.068
M2S1	0.004	-0.045	-0.030	0.021	-0.039	-0.130
M3	-0.092	0.051	0.039	-0.141	-0.052	-0.035
M3S1	-0.138	-0.019	0.003	**-0.285**	-0.039	-0.058

Note. The number in bold only are the correlation coefficients (ρ) with noticeable absolute values.

attributive adjectives + *head nouns* (e.g., Staples & Reppen, 2016), we found that this pattern has been used to present visual information mostly associated with the keywords of the topics. Nevertheless, the pattern of *attributive adjectives* + *head nouns* + *prepositional phrases (of)* can be excluded from the analysis because few cases are found to present visual information in the corpus.

Another connection between rhetorical functions and the intended linguistic resources has to do with M2 (i.e., highlighting visual Information; comparing and contrasting key points) and *attributive adjectives* + *head nouns* pattern. We have two possible interpretations for this pattern and rhetorical connection: a) this pattern suggests that attributive adjectives are used as a communicative vehicle to describe head nouns to meet the rhetorical functions of highlighting and comparing information as in the examples of *the highest percentage/rate/number/rank* vs. *the lowest percentage/percent/rank/bar*; b) as mentioned earlier, the use of attributive adjectives could be associated with the topics, for example *mobile phone* and *developed countries*. It is not easy to avoid using these attributive adjectives when highlight and/or compare the key visual information as in *the percentages for mobile phone users are close to each other*.

Moreover, the results revealed a connection between M2 and prepositional phrases (*of).* This pattern can also align with the p-frame (*the NOUN of*) in Lu et al. (2021a) and Lu et al. (2021b). As Lu et al. (2021a) concluded, the p-frame (*the NOUN of*) contributed to a wide range of rhetorical functions, not only the presentation of research information as discussed above, but also discussing counter claims and/or indicating research gaps in research articles. These functions involved the comparison and contrast between previous literature and recent studies. In our study, the genre is data commentary, however. A reasonable inference is that the pattern – *head nouns* + *prepositional phrases (of)*- can be further used for comparing visual information in the data/table/charts. In contrast to M1, M2 has many cases of *attributive adjectives* + *head nouns* + *prepositional phrases (of)* pattern, for instance *the highest amount of leisure time*.

Finally, the results indicated a negative correlation between M3S1 (e.g., personal asides) and prepositional phrases (*of*). Lu et al (2021a) suggested that some rhetorical functions that are related to *personal asides*, discussed values and limitation of research. This function involved personal comments/ideas on the research. Lu et al. (2021b) mentioned that writers in social sciences tended to include p-frames that express their stance for participating in knowledge construction in particular disciplines. The frequent p-frames with relevant rhetorical functions are *it is important to VERB*, *it has been VERB in*, and *this is the first NOUN*. Based on the examples, we can see that this rhetorical function related to personal stance is expressed mostly by clausal structures instead of nominal structures based on PCFs. Although our genre is different from Lu et al. (2021a), the expression of personal asides can possibly be universal across different written genres. We also found personal asides have rarely been based on nominal constructions with PCFs, but mostly with verbal structures, for example: *in my opinion*, *it is because* ... and *I believe we cannot have a conclusion that* .... This could explain why there was a negative correlation between prepositional phrases (*of*) and M3S1.

CONCLUSION

This study uncovered the use of PCFs and rhetorical functions in data commentaries written by L2 student writers in the Iranian university context. As the L2 student writers become more academically advanced, they tend to use more diverse and abundant PCFs (e.g., nominalizations, prepositional phrases as post-noun modifiers) to construct their data commentary description. This suggests that higher academic levels can be associated with advanced linguistic competence in data commentary writing. Moreover, writing proficiency is reflected not only in linguistic abilities but also in discourse dimensions, where advanced L2 student writers demonstrate a better ability to handle different rhetorical functions (e.g., presenting, highlighting, and/or comparing visual information). Additionally, there is a connection between linguistic features and rhetorical functions, indicating the importance of both dimensions in academic writing development. These insights have practical implications for guiding the instruction and assessment of data commentary writing in educational settings. By understanding students' linguistic and rhetorical patterns, educators can target their instruction to specific areas that need improvement. Finally, this study lays the groundwork for future research; further exploration of additional factors influencing the use of PCFs and rhetorical functions, along with investigations into the impact of instructional interventions on students' writing skills, could broaden our understanding of this field.

This study has some limitations. We need to acknowledge that the discussion of the relationship between the PCFs and rhetorical moves/steps is primarily based on recent studies, because a few prior studies have explored the connection between linguistic features and rhetorical functions in relation to phrasal complexity features in academic writing. Therefore, this study can be regarded as an initial and exploratory investigation. To enhance our understanding of this relationship, future studies should consider incorporating additional research to provide further insights. It is also important to acknowledge that the varying levels of experience between undergraduate and graduate students could potentially impact their approach and proficiency in writing data commentary. This difference was not considered and analyzed in the study. Additionally, although we attempted to minimize topic influence by selecting 20 topics, it is essential to recognize that topic influence cannot be completely eliminated. However, using only one topic for research design can help reduce the influence of topic.

Previous studies have suggested various methods for teaching PCFs, including awareness-raising instruction and data-driven learning. In line with these findings, our study proposes an integrated approach that combines grammatical instruction on PCFs with relevant rhetorical moves and steps. This holistic approach allows students to not only understand the grammatical aspects of PCFs but also comprehend how to use them strategically to achieve specific rhetorical purposes. Utilizing corpora is an effective way to facilitate this integration. Corpora can provide valuable resources of authentic writing in specific genres, such as data commentary and research article. By incorporating corpora into the teaching process, students gain access to real-world examples and can analyze how PCFs are appropriately used in their target genres. This enhances their understanding and application of PCFs within the context of academic writing. Moving from the linguistic instruction to the rhetorical instruction can effectively help students use PCFs appropriately in the written genres that they need to work on in their specific disciplines.

The authors would like to thank the students who actively and voluntarily participated in this research project. They also thank the anonymous reviewers.

DECLARATION OF COMPETITING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Muhammed Parviz: Corpus building/data collection, discourse analysis of rhetorical moves/steps, manuscript drafting.

Ge Lan: Corpus tagging, extraction and analysis of linguistic features, statistical analysis, manuscript drafting.

ACKNOWLEDGMENTS

REFERENCES

Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes, 31*, 58-71. https://doi.org/10.1016/j.jeap.2017.12.008

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D., Connor, U., & Upton, T. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins Publishing.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2-20. https://doi.org/10.1016/j.jeap.2010.01.001

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5-35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., Staples, S., & Egbert, J. (2022). *The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application*. Routledge.

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes, 6*, 289-302. https://doi.org/10.1016/j.jeap.2007.09.009

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Publishers.

Cullen, P., French, A., & Jakeman, V. (2014). *The official cambridge guide to IELTS for academic and general training*. Cambridge University Press.

Eriksson, A., & Nordrum, L. (2018). Unpacking challenges of data commentary writing in master's thesis projects: an insider perspective from chemical engineering. *Research in Science & Technological Education, 36* (4), 499-520. https://doi.org/10.1080/02635143.2018.1460339

Hyland, K. (2000). *Disciplinary discourses: Social interaction in academic writing.* Pearson.

Hyland, K. (2019). *Metadiscourse: Exploring interaction in writing*. Continuum.

Jalilifar, A., Parviz, M., & Don, A. (2019). Exploring phrasal complexity features in graduate students' data commentaries and research articles. *Journal of English Language Teaching and Learning*, *24*(11), 118-152. https://doi.org/10.22055/.2019.17801

Jiang, F. K., & Hyland, K. (2015). 'The fact that': Stance nouns in disciplinary writing. *Discourse Studies*, *17*(5), 529–550. https://doi.org/10.1177/1461445615590719

Jiang, F. K., & Hyland, K. (2021). 'The goal of this analysis ...': Changing patterns of metadiscursive nouns in disciplinary writing. *Lingua*, *252*, 103017. https://doi.org/10.1016/j.lingua.2020.103017

Lan, G., & Sun, Y. (2019). A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course. *Journal of English for Academic Purposes*, *38*, 14-24. https://doi.org/10.1016/j.jeap.2018.12.001.

Lavelle, E., & Bushrow, K. (2007). Writing approaches of graduate students. *Educational Psychology*, *27*(6), 807-822. https://doi.org/10.1080/01443410701366001.

Lim, J. M. H. (2011). Paving the way for research findings: Writers' rhetorical choices in education and applied linguistics. *Discourse Studies, 13*(6), 725-749. https://doi.org/10.1177/1461445611421364.

Lu, X., Yoon, J., & Kisselev, O. (2021). Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes, 61,* 63-83. https://doi.org/10.1016/j.esp.2020.10.001.

Lu, X., Yoon, J., Kisselev, O., Casal J., Deng, J., & Nie, R. (2021). Rhetorical and phraseological features of research article introductions: Variation among five social science disciplines. *System*, *100*, 102543. https://doi.org/10.1016/j.system.2021.102543.

Moreno, A. & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes, 50*, 40-63. https://doi.org/10.1016/j.esp.2017.11.006.

Nordrum, L., & Eriksson., A. (2015). Using a small, specialized corpus for formative self-assessment practices. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment. Studies in corpus linguistics* (vol. 70, pp. 59-84). John Benjamins Publishing Company.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes, 14*, 48-59. https://doi.org/10.1016/j.jeap.2013.12.001.

Parviz, M. (2023). A comparative corpus-based investigation of results sections of research articles in Applied Linguistics and Physics. *ICAME Journal*, *47*(1). https://doi.org/10.2478/icame-2023-0005.

Parviz, M., Jalilifar, A., & Don, A. (2020). Phrasal discourse style in cross-disciplinary writing: A comparison of phrasal complexity features in the results sections of research articles. *Círculo de Lingüística Aplicada a la Comunicación*, *83*, 191-204. https://doi.org/10.5209/clac.70573.

Samraj, B. (2008). A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes, 7*, 55-67. https://doi.org/10.1016/j.jeap.2008.02.005.

Sancho Guinda, C. (2012). Proximal positioning in students' graph commentaries. In K. Hyland & C. Sancho Guinda (Eds.), Stance and Voice in Written Academic Genres (pp.166-183). Palgrave Macmillan.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183. https://doi.org/10.1177/074108831663152.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexicogrammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17-35. https://doi.org/10.1016/j.jslw.2016.02.002.

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.

Swales, J. M. & Feak, C.B. (1997). From information transfer to data commentary. In Miller, T. (Ed.), *Functional approaches to written text: Classroom applications*. United States Information Agency.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills.* University of Michigan Press.

Wharton, S. (2012). Epistemological and interpersonal stance in a data description task: Findings from a discipline-specific learner corpus. *English for Specific Purposes, 31*(4), 261-270. https://doi.org/10.1016/j.esp.2012.05.005.

Yang, R., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. English for Speci*fi*c Purposes, 22(4), 365-385. https://doi.org/10.1016/S0889-4906(02)00026-1.

APPENDIX A

A list of 20 writing topics used in the study.

(1) The charts below show the results of a survey of adult education. The first chart shows the reasons why adults decide to study. The pie chart shows how people think the costs of adult education should be shared.

(2) The chart shows the number of mobile phones and landlines per 100 people in selected countries.

(3) The chart below shows estimated world illiteracy rates by region and by gender for the year 2000.

(4) The chart shows the percentage of their food budget the average family spent on restaurant meals in different years. The graph shows the number of meals eaten in fast food restaurants and sit-down restaurants.

(5) The bar chart shows the number of visitors to three London Museums between 2007 and 2012.

(6) The bar graph shows the annual number of rides taken by two forms of public transports in the city of Williamsville.

(7) The bar chart below shows the results of a survey conducted by a personnel department at a major company. The survey was carried out on two groups of workers: those aged from 18-30 and those aged 45-60, and shows factors affecting their work performance.

(8) The line graph below gives information on cinema attendance in the UK.

(9) The chart below shows the Higher Colleges of Technology graduates in the UAE.

(10) The chart below shows the amount of leisure time enjoyed by men and women of different employment status.

(11) The diagrams below show the changes that have taken place at West park Secondary School since its construction in 1950.

(12) The pie charts and the table show the types of living accommodation occupied by 25-year-olds in London during the 1990s and 2010s and the availability of different types of accommodation in London during the same two periods.

(13) The pie charts below show the share of Oscar winners by film genre for 2003 and 2008.

(14) The line graph below shows the percentage of tourists to Scotland who visited four different attractions in Edinburgh.

(15) The chart below gives information on the percentage of British people giving money to charity by age range for the years 1990 and 2010.

(16) The pie charts below show the online sales for retail sectors in New Zealand in 2003 and 2013.

(17) The chart below shows the changes that took place in three different areas of crime in New Port city center from 2003-2012.

(18) The chart below shows the annual number of rentals and sales (in various formats) of films from a particular store between 2002 to 2011.

(19) The chat below gives information about Southland’s main exports in 2000 and future projections for 2025.

(20) The graph and table below show average monthly temperatures and average number of hours of sunshine per year in three major cities.

APPENDIX B

The precision and recall rates of the three target features

Target features	Precision	Recall
Adjectives	95.3%	95.9%
Nouns	96.2%	97.8%
Prepositions	98.9%	99.9%

Note: The precision and recall are generated based on 5% files of the corpus.

APPENDIX C

The detailed description of the coding scheme of rhetorical functions as well as examples taken from the datasets are presented.

Moves	Definitions	Examples from the Datasets
M1: Presenting visual information	Announcing the purpose(s) of visual prompts/forms	The bar graph representation provides data description on a survey of adult education with reference to the reasons why they have made their minds up to continue their studies. (GR)
M1S1: Providing an explanatory note to set the scene	Making a brief note to begin reporting visual information that follows	In today's report I am going to compare people's food budget in restaurant meal during 1970 until 2000...(UG)
M1S2: Indicating the location of the data	Showing the position of the non-verbal information	As bar chart shows, these factors have different effects on each range of age. (GR)
M2: Highlighting visual information; Comparing and contrasting key points	Revealing data at several levels of details including trends and regularities	From 2009 onwards, it has experienced an increase and in 2012 it records 14 million visitors in a year. Victoria and Albert Museum, on the other hand, shows a constant annual record from 2007 until 2008. The record is noticeable and is about 13 million visitors each year. However, this annual record undergoes a gradual decrease in the following two years....(GR)
M2S1: Describing the facts (with/without providing statistical evidence)	Stating information (with/without providing statistical evidence)	The least important fact among those mentioned, however, is meeting people. With respect to sharing the cost of the courses most probably is that it should be shared on the basis that tax payers pay 25%, employers 35% and individuals 40%. (GR)
M3: Commenting on visual information	Making comments on the non-verbal information	I believe landlines are not as important as mobile phones, we can carry them everywhere but landlines don't have this property... (UG).
		It is strange that the numerical distance between sexes in Latin America is very low (less than two percent)...(GR).
M3S1: Personal asides	Expressing personal opinions	I'm not trying to say that fast foods are disaster, but it is very harmful if it becomes your everyday meal. (UG)
M4: Concluding visual information	Affording a short account or summary of the key points	To conclude, we can say that Dubai and Abu Dhabi are the two most popular colleges in UAE and the women occupied a much greater role in UAE... (GR)

APPENDIX D

Shapiro-Wilk Test on Normality

	Statistics	*P*-value	Description
Attributive adjectives	0.993	0.081	normal
Appositive NPs	0.313	0.000	not normal
Noun-noun sequences	0.925	0.000	not normal
Nominalizations	0.912	0.000	not normal
Prepositional phrases (*of*)	0.958	0.000	not normal
Prepositional phrases (*others*)	0.969	0.000	not normal
M1	0.929	0.000	not normal
M1S1	0.684	0.000	not normal
M1S2	0.882	0.000	not normal
M2	0.992	0.052	normal
M2S1	0.904	0.000	not normal
M2S2	0.278	0.000	not normal
M3	0.526	0.000	not normal
M3S1	0.496	0.000	not normal
M4	0.605	0.000	not normal