This Paper is Organized as Follows: Lexical Bundles in Computer Science Academic Texts Produced by Novice and Expert Writers

Wesley Henrique Acorinti & Ana Eliza Pereira Bocorny
Federal University of Rio Grande do Sul

Abstract. Conventional academic writing in English is crucial for scholars seeking publication in peer-reviewed international journals. English for Academic Purposes (EAP) materials often use a "one-size-fits-all" approach that does not cater to disciplinary variation or learner needs (Murray, 2016). Lexical bundles (LBs), defined as recurrent continuous sequences of words, are essential building blocks of academic discourse (Biber et al., 2002). While much research has been done on LBs in academic contexts (Neely & Cortes, 2009; Cortes, 2013; Staples et al., 2013; Gil & Caro, 2019), to the best of our knowledge, few studies have compared the use of LBs in texts produced by expert writers (EWs) and novice writers (NWs) to inform the design of EAP materials. This study explores two corpora: (i) a corpus of 170 undergraduate theses written in English by Brazilian undergraduates in Computer Science and (ii) a corpus of 581 published research articles of the same discipline. The primary focus of the study is the introduction section, wherein the goal is to extract, categorize, and compare the most frequently occurring 4-, 5-, and 6-word transparent bundles based on their rhetorical function. The results reveal four trends: (a) both groups of writers do not use LBs to realize certain steps; (b) both groups use LBs similarly; (c) NWs use LBs that EWs do not, and (d) EWs use LBs that NWs do not. These findings can inform the design of EAP materials that cater to the disciplinary variation and learner needs of different academic contexts.

Plain English Abstract. Academic writing in English is essential for scholars who want to publish their work in respected international journals. Most materials that teach English for Academic Purposes (EAP) do not take into account the specific needs of different academic fields or learners. In this study, we focus on a particular aspect of academic writing, which is the use of lexical bundles (LBs). LBs are common phrases or word sequences that are frequently used in academic writing (Biber et al., 2002). We analysed two sets of texts: undergraduate theses in Computer Science written in English by Brazilian students and published research articles in the same discipline. We extracted, categorized, and compared LBs used in the introduction sections of these texts written by expert writers and novice writers. Our goal was to discover which LBs were used most often and for what purpose. Our findings showed that both expert and novice writers did not use LBs as much as they could have in certain instances. However, both groups used LBs in a similar way for most of their writing. We also found that novice writers used some LBs that expert writers did not, and vice versa. These findings can help create better materials for teaching EAP, which can better serve the needs of learners across different academic disciplines.

Keywords: lexical bundles; English for academic purposes; academic writing; communicative functions.

1 Introduction

Academia widely recognizes English as the predominant language of communication, thereby placing substantial pressure on scholars across diverse fields to publish their research in this language to attain global recognition. According to Hyland and Shaw (2016, p. 5), English is a "near-universal academic lingua franca," and its widespread use cannot be ignored by academics worldwide, who must read, write, and publish in an additional language. Although the reasons behind English's dominance in

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

academia can be debated as "the result of a conspiracy orchestrated by political and economic interests or the legacy of US and British colonialism" (Hyland, & Shaw, 2016, p. 5), the effects are undeniable. This reality requires academics to invest additional time and effort in mastering the language, on top of their research responsibilities.

In an effort to expand their involvement in research communities at an early stage, Brazilian computer science undergraduates are increasingly writing their undergraduate theses in English. This genre, also referred to as "final paper" in English and "*Trabalho de Conclusão de Curso*" or "*TCC*" in Portuguese, comprises the majority of available undergraduate theses written in English on Lume, one of the largest open-access digital repositories in the world, maintained by the Federal University of Rio Grande do Sul.

The majority of universities in Brazil require senior-year undergraduate students to produce a thesis as a prerequisite for earning their bachelor's degree. The purpose of producing this genre is to acquaint students with research methods and writing techniques while also providing them with a comprehensive understanding of a topic they select. The organizational pattern of an undergraduate thesis in Brazil mirrors that of a research article, consisting of an introduction, methodology, results/discussion, and conclusion.

Genres are shaped by the discourse communities they are rooted in, which share a common set of communicative events aimed at achieving specific communicative goals (Swales, 1990). According to Biber and Conrad (2019), genres can be compared based on various situational characteristics such as participants and communicative purposes. It is essential to explore the situational characteristics of the introductions of undergraduate theses to facilitate proper comparisons, especially since the understanding of this genre can vary across cultural contexts.

The objective of our study is to analyse the writing style of computer science undergraduate theses written in English and compare it to that of established researchers who publish in peer-reviewed journals. Our methodology involves a comparison of the introduction sections of Brazilian undergraduate theses and research articles, which both follow the expected conventional rhetorical organization of reviewing previous research, identifying a research gap and stating how the present study addresses that gap (Biber & Conrad, 2019). Furthermore, the introduction section of the two genres conforms to the purpose of providing a rationale for the work and attracting interest to the topic and the reader (Swales & Feak, 2012). Hence, due to those similarities, the comparative approach adopted in this study seems well-suited to our goal of highlighting differences in writing style between newcomers to the scientific community in this field of study and experienced researchers in the same area.

There are, however, notable distinctions in terms of participants and relationships within the two genres being discussed. Senior-year undergraduates, who are novices in the discourse communities of researchers in their field, produce undergraduate theses, whereas more experienced researchers[23] publish their work in peer-reviewed journals. The target audience for these written academic genres also varies slightly. Readers of undergraduate theses generally expect a more extensive and comprehensive discussion of the research topic compared to readers of research articles. Both genres are accessible online, with undergraduate theses being published in Lume as open-access documents. It

[23] In this study, we draw a comparison between the papers produced by *more experienced writers* and those written by *undergraduate students*. The term *more experienced writers* refers to authors who have advanced beyond the undergraduate level, consistent to our survey of the journal from which we source our data. This group comprises, for example, individuals who are either pursuing their Master's or PhD degrees or are already serving as instructors or professors. Given that these authors are subjected to a rigorous peer-review process prior to the publication of their work, a hallmark of a reputable journal, we believe that the research articles examined from this group align with the objectives of our study.

is worth noting that both genres undergo a peer-review process before publication to ensure adherence to conventions of discourse and scientific rigor. In the case of undergraduate theses, they undergo committee review before being included in the open institutional repository.

While many recent studies have examined the language used in academic texts written in English, focusing on the significance of lexical bundles (LBs) across various genres such as research article introductions (Cortes, 2013), bachelor dissertations written in English as a second language (L2) (Gil & Caro, 2019), academic lectures (Neely & Cortes, 2009), and proficiency tests (Staples et al., 2013), there remains a scarcity of research comparing the use of LBs in texts produced by expert and novice writers aiming at informing the design of discipline-specific English for Academic Purposes (EAP) materials.

In our study, we build upon the work of Moreno and Swales (2018, p. 77), who identified a "function-form gap" that can be bridged by identifying significant text items or patterns in a specific rhetorical context. LBs have been studied in EAP corpus-based research, but simply relying on frequency data can result in meaningless strings (Swales, 2019). Our research combines Swalesian genre analysis and corpus linguistics in an attempt to bridge the form-function gap while comparing LBs used by EWs and NWs in computer science, revealing differences and identifying language elements that are worth teaching. Therefore, the present study aims to address gaps in current research by extracting, categorizing, and comparing the most frequent 4-, 5-, and 6-word transparent bundles according to their rhetorical function in two corpora, to later propose pedagogical implications.

One corpus investigated contains introduction sections of undergraduate theses written in English by Brazilian computer science senior-year undergraduates, and the other contains introductions of published research articles written by professional academics in the same field. These bundles will then be compared and studied in relation to their rhetorical function proposed by Cortes (2013), which is based on the work of Swales (1990; 2004). Finally, we draw conclusions and concrete pedagogical applications based on our analyses. More specifically, this study seeks to answer three research questions:

(1) To what extent do the patterns of LBs used by undergraduate theses writers differ from those used by experienced academic researchers in published research articles?
(2) How do these differences relate to the merged rhetorical framework proposed by Cortes (2013), and what implications can be drawn for understanding the conventions of academic discourse in Computer Science?
(3) In light of these findings, what pedagogical applications can be recommended for teaching English for Academic Purposes (EAP) to students in computer science?

1.1 Corpus linguistics and EAP

EAP instructors often lack direct exposure to the professional contexts and disciplines of their students, which may hinder their intuitive understanding of language use in specialized domains. As Sinclair (1991) noted, relying solely on human intuition can be a poor guide to understanding how language functions in real-world situations. In fact, Nesi (2013, p. 407) suggested that such instructors "may not have much intuitive understanding of the way language is used in certain specified domains."

Corpus Linguistics is used in English for Specific Purposes (ESP) and English for Academic Purposes (EAP) to examine language features, including register, lexicogrammar, and phraseology

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

within corpora. This approach offers valuable insights into language use in specific contexts and helps identify the linguistic patterns of academic disciplines and professions. By using specialized corpora, targeted language instruction can be developed to better prepare students for their chosen fields of study (Bennet, 2010). As highlighted by Nesi (2013), using corpora to inform the development of pedagogical materials has numerous benefits. Researchers can compare the language found in specialized corpora with that presented in non-corpus-based or non-corpus-informed textbooks, identify discrepancies, and update instructional materials to accurately reflect the language students are likely to encounter in their areas of study.

In EAP, corpus-based discipline-specific materials are increasingly essential (Hamp-Lyons, 2011). To the present, many EAP instructors rely on generic, "one-size-fits-all" materials that are designed for use in a wide range of academic contexts and not tailored to the specific needs of a particular discipline or field (Murray, 2016). As Hyland (2016, p. 20-21) observes, "disciplines are largely created and maintained through the distinctive ways in which members jointly construct a view of the world through their discourse." Therefore, it is crucial to highlight the distinct lexical, grammatical, and rhetorical resources needed to create specialized knowledge within each discipline.

In order to address the gap in existing EAP materials that fail to provide such targeted resources, corpus-based analyses offer valuable insights for developing materials that are closely aligned with the language demands of specific disciplines. By examining language use in specialized corpora, one can identify patterns of vocabulary, grammar, and discourse that are characteristic of particular academic fields. Such insights can inform the development of materials that more accurately reflect the language use in those fields, thereby improving EAP instruction and enhancing learners' ability to succeed in their academic pursuits.

1.2 Lexical Bundles and Swalesian Genre Analysis

Lexical bundles (LBs), as defined by Biber et al. (2002, p. 443), are recurring sequences of words that become "prefabricated chunks," easily retrievable from a writer's or speaker's memory, and used as building blocks for constructing texts. The significance of such units in indicating "competent participation in a given community" (Hyland, 2008, p. 5) is increasingly acknowledged and supported by empirical evidence. Hence, it seems crucial for scholars aiming to publish their work in peer-reviewed international journals to familiarize themselves with LBs and incorporate them in their academic writing. The use of LBs in academic writing not only enhances the coherence and fluency of the text but also reflects the writer's familiarity with the conventions of their field and their ability to communicate their ideas effectively.

Rhetorical moves, as defined by Swales (2004), are discoursal or rhetorical units that perform coherent communicative functions in a given discourse. The concept of moves originates from genre analysis, which was developed by Swales (1981; 1990; 2004) to address the need for teaching ESL students how to enhance their reading and writing skills in academic settings. In his work, Swales (1990) expands upon the concept of discourse community as sociorhetorical networks that emerge to pursue shared objectives and fulfil specific criteria, which encompass common goals, participatory mechanisms, information exchange, community-specific genres, specialized terminology and expertise. Genre, for Swales (1990, p. 58), is conceptualized as "a class of communicative events, the members of which share some set of communicative purposes." By studying the discourse community's genres and their communicative purposes, Swales identifies moves as fundamental units of discourse that serve specific communicative functions in a given genre.

According to Biber, Connor, and Upton (2007, p. 9), the concept of "moves" refers to the fundamental components of texts, acting as building blocks. These moves, in turn, are comprised of "steps," which are described as "multiple text fragments" (Moreno & Swales, 2018, p. 40) and serve the specific purpose of the move they belong to (Biber, Connor & Upton, 2007). A step, as a text fragment, plays a crucial role in introducing new propositional meaning essential for advancing the text and achieving the intended purpose of the genre. Competent readers of a genre recognize the significance of a step, enabling them to infer a specific communicative function without broad generalizations (Moreno & Swales, 2018). Functioning as a smaller functional text within a move, a step acts as an elaborator, operating at a more detailed level than the move itself (Al-Shujairi & Al-Manaseer, 2022).

LBs, rhetorical moves, and steps play crucial roles in academic writing. LBs contribute to the fluency and coherence of the text, while a framework of rhetorical moves and their corresponding steps help structure and effectively present information within specialized discourse. These elements indicate the writer's level of engagement within the discourse community relevant to the text. Therefore, it is relevant to understand and employ conventional LBs, moves, and steps to produce academic writing pieces that showcase the writer's competence and expertise in their field of study.

Swales and Feak's (2012, p. 331) influential textbook, *Academic Writing for Graduate Students*, highlights how research paper introductions serve as a response to two types of competition: competition for readers and competition for research space. They introduce the create-a-research-space (CARS) model framework from Swales (1990) to help learners understand this pattern. However, in 2004, Swales himself suggested an update to this widely circulated model. He cautioned against its improper rigidity and mechanistic application, emphasizing the importance of contextualizing its use within specific disciplinary and cultural contexts.

Cortes (2013) presents a novel approach that relates LBs to move analysis by integrating two frameworks developed by Swales (1990; 2004). Specifically, for Move 1, "establishing a territory", Cortes (2013) utilizes the steps introduced by Swales in 1990, while for Move 2 and 3, she selects the steps introduced in Swales' 2004 work. A summary of the moves and steps used by Cortes (2013) can be found in Table 1. The merging of the two frameworks allows for a more nuanced and comprehensive analysis of the connection between moves and steps with LBs in academic writing, therefore, contributing to a better understanding of the communication patterns within specific discourse communities.

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

Table 1: *Moves and steps in research article introductions (Cortes, 2013, p. 37, adapted from Swales, 1990, 2004)*

Move 1 Establishing a territory
Step 1 Claiming centrality
Step 2 Making topic generalizations
Step 3 Reviewing items of previous literature
Move 2 Establishing a niche
Step 1A Indicating a gap or
Step 1B Adding to what is known
Step 2 Presenting positive justification
Move 3 Presenting the present work
Step 1 Announcing present research descriptively and/or purposively
Step 2 Presenting research questions or hypothesis
Step 3 Definitional clarifications
Step 4 Summarizing methods
Step 5 Announcing principal outcomes
Step 6 Stating the value of the present research
Step 7 Outlining the structure of the paper

Overall, Swales (1981; 1990; 2004) developed a text analysis method that helps English language learners improve their research article reading and, most importantly, writing skills. This approach has been widely used by researchers to uncover the generic rhetorical structure of different genres. However, there is a research gap observed by Moreno and Swales (2018, p. 41) referred to as the "function-form gap." This gap can be bridged by identifying the most significant text items or patterns (form), such as lexical bundles, in a specific rhetorical context (function), i.e., a move and step, that aid readers in interpreting a particular communicative function with accuracy. As Moreno and Swales (2018) argue, this objective is relevant to all text fragments that realize a significant communicative function, except for cases where the function is not linguistically indicated, such as implicit causal logical functions.

Swales (2019) notes that LBs have become a prevalent topic in EAP corpus-based research, particularly for larger corpora where there is a preference for four-word bundles. However, simply relying on frequency data to generate LBs lists may result in meaningless strings that are of "little meaning or use to English language teachers and learners" (p. 77). In contrast, Simpson-Vlach and Ellis's (2010) study on academic formulas list went beyond to employ a mixed methodology that included frequency data, statistical measures, and polling of EAP teachers and testers, showing potential pedagogical value, according to Swales (2019). Unlike studies that "provide frequency data without any consideration of possible or potential pedagogical uptake" (Swales, 2019, p. 77), our research aims to combine Swalesian genre analysis and corpus linguistics to compare LBs used by EWs and NWs in computer science, revealing differences between the writing of the two groups, identifying what is worth teaching, and proposing concrete pedagogical applications to our findings.

All in all, the introduction section of Brazilian undergraduate theses and research articles adheres to the established rhetorical organization. Rhetorical moves, as defined by Swales (2004), are

coherent communicative units that contribute to the overall structure of the discourse. Within these moves, steps act as essential building blocks, consisting of multiple text fragments that serve specific purposes within the move (Biber, Connor & Upton, 2007). Steps play a crucial role in introducing new propositional meaning and establishing textual progression. Thus, recognizing the significance of steps allows competent readers to infer the specific communicative function they serve (Moreno & Swales, 2018). To accurately interpret communicative functions, it is essential to bridge the gap between form and function by identifying significant text items or patterns within a specific rhetorical context (Moreno & Swales, 2018). Readers can rely on elements such as LBs, which are recurring word sequences serving as prefabricated chunks. LBs enhance the fluency and cohesion of the discourse, acting as fundamental building blocks for constructing texts (Biber et al., 2002). By recognizing and utilizing these linguistic patterns, readers can better understand and engage with the intended meaning and purpose of the text, while writers can demonstrate competent participation within a given community (Hyland, 2008).

2 Methodology

The primary objective of this study was to enhance the development of discipline-specific EAP materials. To achieve this, two corpora were compiled to extract LBs and categorize them according to the rhetorical functions they convey. The quantitative analysis was conducted using SketchEngine (Kilgarriff et al., 2014) and the qualitative analysis involved manual categorization of transparent LBs, which will be explained in this section.

LBs are textual segments identified using software based on their frequency and dispersion, irrespective of their idiomaticity and structural status (Biber et al., 2021). LBs play a significant role in language instruction, not only in constructing grammatically correct sentences but also in employing "well-established lexical expressions in appropriate contexts" (Biber et al., 2021, p. 982). Furthermore, when considering terminology studies (Krieger & Finatto, 2004), LBs relate to the phraseology present in diverse genres within specialized discourse. It can be contended that LBs function as phraseological components and consequently, contribute to the formation of a stereotyped linguistic structure, leading to an autonomous semantic interpretation of the meanings derived from the structure's constituents.

This study draws on two corpora of introduction sections, one comprising undergraduate theses by novice writers (NWs) and the other consisting of published research articles authored by more experienced researchers and hence expert writers (EWs), both from the field of computer sciences. The data were collected with the aim of contrasting the results and identifying the differences in the use of LBs across the two groups.

To create the Corpus of Computer Sciences Final Paper Introductions from Lume (CorCompInt-Lume), all undergraduate theses written in English under the "Computer Science - Undergraduate degree" collection on Lume were located through website filters for subject (Computer Science), language (English) and genre (undergraduate thesis) and, subsequently, downloaded[24]. The introduction sections were then manually identified, copied, and pasted into a Word document (Microsoft, 2022). After that, the data was cleaned to remove any elements that did not pertain to the section's prose or reflect the students' writing skills, such as page numbers, titles, subtitles, graphs, tables, images, captions, formulas, footnotes, and block quotes.

[24] Obtaining consent from the students was not necessary for the process, as all undergraduate theses made available on Lume are open-access and covered under the Creative Commons BY-NC-SA 2.5 Deed.

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

The Corpus of Computer Sciences Introductions from PLOS One (CorCompInt-PLOS), consisting of text produced by expert writers (EWs), was generated using AntCorGen (Anthony, 2022), a freeware program that facilitates the creation of section- and discipline-specific corpora from peer-reviewed, open-access articles published in international peer-reviewed journals available on the PLOS One platform. Subsequently, both corpora were uploaded to and compiled on Sketch Engine (Kilgarriff et al., 2014). Table 2 displays the number of tokens, types, and texts in each corpus used in this study.

Table 2: *Corpora metadata*

Corpus	Number of tokens	Number of types	Number of texts
CorCompInt-Lume	173,371	148,310	170
CorCompInt-PLOS	646,236	532,235	581

Three criteria were considered for the extraction of n-grams using Sketch Engine (Kilgarriff et al., 2014): (i) the extension of the word sequences (n-gram length), (ii) their raw frequency in the corpus (minimum frequency), and (iii) the number of texts in which the sequences occurred in the corpora (dispersion). For CorCompInt-Lume, the extraction criteria were as follows n-gram length of 5, 6, and 7; minimum frequency of 6; and dispersion of 3. Given the larger size of CorCompInt-PLOS, we employed a more conservative approach to extract a similar number of LBs as we did from CorCompInt-LUME. This allowed us to compare only the most frequent LBs. Thus, for CorCompInt-PLOS, the extraction criteria were as follows: n-gram length of 5, 6, and 7; minimum frequency of 15; and dispersion of 8.

The total number of LBs meeting the described criteria in CorCompInt-Lume was 125 and in CorCompInt-PLOS, 120. However, due to the objectives of this study, we worked with a subset of LBs that met the established criteria and were easily classified as transparent lexical bundles. These so-called transparent lexical bundles contained a semantically significant collocation node (Frankenberg-Garcia et al., 2019).

The two authors collaborated to manually categorize the LBs into the merged Swalesian framework for moves and steps proposed by Cortes (2013; see Table 1). The most significant functions conveyed by the chosen bundles were prioritized, and the categorization was performed by associating LBs with specific steps based on their collocation nodes. LBs that could not be clearly linked to certain steps were disregarded. This methodological decision was made because we believed it would be more useful and pedagogically beneficial to concentrate on bundles that were transparent in conveying rhetorical functions. While acknowledging that less transparent bundles, such as discourse organizers that occur across sections and moves, and stance bundles that express the authors' opinions are also important (Hyland, 2008), we delimited our scope and left them for further research.

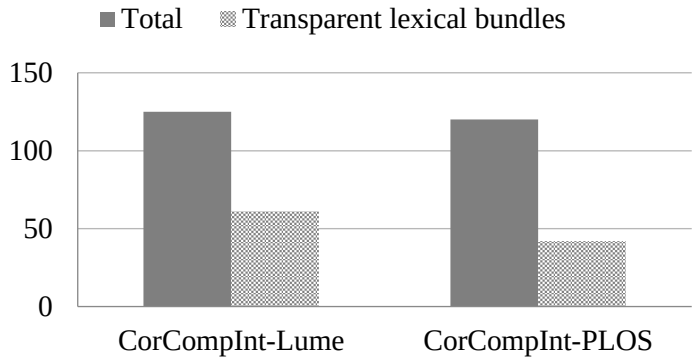
Overall, this study followed a corpus-driven approach to the data in which we combined a bottom-up corpus linguistics methodology with a top-down move-analysis approach. First, two corpora were designed and compiled, then 4-, 5- and 6-word bundles were extracted, analysed and sorted. Then, the transparent LBs were categorized into each step of an introduction. Finally, we observed four trends in our data, which are explored in the section below, to later assess pedagogical implications.

3 Results and discussion

This study aims to compare the use of LBs in Computer Science by NWs and EWs to gain insights into the pedagogical applications of language data, which can then inform the development of discipline-specific EAP materials. In this context, NW refers to writers of undergraduate theses who are Brazilian undergraduate students of computer science, while EW refers to other researchers in the same field who have published papers in journals from the PLOS One platform.

Figure 1 shows the total number of LBs extracted from each corpus, along with the number of LBs selected for this study from each corpus. A total of 245 LBs were extracted, with 125 from CorCompInt-Lume and 120 from CorCompInt-PLOS. In selecting bundles for the study, 61 LBs (48.8%) were chosen from CorCompInt-Lume, and 42 LBs (33.6%) were chosen from CorCompInt-PLOS[25]. These selected bundles are referred to as transparent lexical bundles because they allow for more accurate identification of the rhetorical function they realize. We classified the transparent lexical bundles based on the primary moves and steps conveyed by them, as outlined in the merged framework proposed by Cortes (2013). This classification was made by observing the collocation nodes and their immediate context.

Figure 1: *Total and transparent numbers of LBs.*



Four trends were identified when comparing the data extracted from both corpora. First, both groups of writers showed a lack of use of LBs to realize certain steps. Second, there were instances of similar and/or identical use of LBs by both groups of writers. Third, NWs used some LBs that EWs did not use. Finally, there were some LBs used by EWs that were not used by NWs.

The first trend observed in the data appears to reveal that none of the transparent LBs expressed six steps from the merged Swalesian framework by Cortes (2013). These six steps include "making topic generalizations" (M1 - S2), "reviewing previous literature" (M1 - S3), "adding to existing knowledge" (M2 - S1B), "presenting research questions or hypotheses" (M3 - S2), "providing definition clarifications" (M3 - S3) and "summarizing research methods" (M3 - S4). These findings raise the question of whether these steps are not typically expressed using a formulaic structure or whether they are not present in written works within this discipline. While it is reasonable to assume that certain steps

[25] See Appendices for the raw frequency, normalized frequency and document frequency of transparent lexical bundles extracted from CorCompInt-PLOS and CorCompInt-Lume.

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

may not be realized through formulaic language, caution should be exercised when interpreting this result due to the limitations of our samples.

Table 3 shows that NWs and EWs use similar or identical LBs for the steps of "indicating a gap" (M2 - S1A) and "outlining the structure of the paper" (M3 - S7). When it comes to expressing the function of M2 - S1A, both groups of writers tend to use the 4-word bundle *best of our knowledge*, as illustrated in examples 1 and 2. The group of transparent LBs that express the function of "outlining the structure of the paper" (M3 - S7) is the most frequently used in both corpora. This indicates that M3 - S7 is not only reliant on formulaic language but is also an obligatory rhetorical step of papers in computer science.

Ex. 1. 'To the best of our knowledge, no full multi-objective linear model for UTRP has been published before.' (CorCompInt-Lume)

Ex. 2. 'To the best of our knowledge, it is the first attempt to combine these two methods to improve the performance of NER on Chinese EMR.' (CorCompInt-PLOS)

Table 3: *Similar and/or identical use of LBs by NWs and EWs*

Rhetorical structure		CorCompInt-Lume	CorCompInt-PLOS
		Lexical bundles	Lexical bundles
M2 - S1A	Indicating a gap	best of our knowledge the best of our knowledge the best of our	the best of our knowledge the best of our best of our knowledge To the best of To the best of our knowledge To the best of our
M3 - S7	Outlining the structure of the paper	is organized as follows work is organized as work is organized as follows This work is organized This work is organized as this work is organized of this work is organized as this work is organized as follows this work is organized as of this work is organized This work is organized as follows The remainder of this The remaining of this is structured as follows document is organized as rest of this work remainder of this work is remainder of this work	is organized as follows paper is organized as paper is organized as follows this paper is organized of this paper is organized this paper is organized as follows this paper is organized as of this paper is organized as The remainder of this The remainder of this paper remainder of this paper The remainder of this paper is remainder of this paper is the paper is organized of the paper is organized the paper is organized as follows the paper is organized as of the paper is organized as

remainder of this work is organized	the structure of the
document is organized as follows	The rest of the paper is
This work is divided	The rest of the paper
the organization of the	rest of the paper is
The remainder of this work is	rest of the paper
The remainder of this work	rest of this paper
The rest of this work is	The rest of this
The rest of this work	The rest of this paper
rest of this work is organized	remainder of this paper is organized
rest of this work is	rest of this paper is
this document is organized as	rest of the paper is organized
this document is organized	The rest of this paper is
are presented in Chapter	rest of this paper is organized
presents an overview of	paper is structured as follows
of this document is	is structured as follows
	paper is structured as
	The remainder of the

While the term "paper" was exclusively employed to describe texts in CorCompInt-PLOS (as in example 3), NWs in CorCompInt-Lume referred to their written pieces as "work" (as in example 4) or "document" (as in Example 5). This divergence in terminology could potentially be attributed to the distinct genres produced by each group. Despite this difference, in both corpora, writers employ the same pattern of construction, which might characterize it as a lexical frame, i.e., “discontinuous sequences in which words form a ‘frame’ surrounding a variable slot” (Gray & Biber, 2013, p. 109), as suggested in Table 4.

> Ex. 3. ‘This paper is organized as follows: section State of the art introduces the state of the art about characterization methodology and selection of simulation intervals.’ (CorCompInt-PLOS)

> Ex. 4. ‘This work is organized as follows: in chapter 2 the MOSFET technology is reviewed, and the FinFET architecture detailed, along with comparisons.’ (CorCompInt-Lume)

> Ex. 5. ‘The remainder of the document is organized as follows: In Section 1.1 a formal definition of the problem is given.’ (CorCompInt-Lume)

Table 4: *Lexical frame "The * of this * is * as follows"*

The	*	of	this	*	is	*	as	follows
	remainder			**paper**		organized		
	rest			**work**		structured		
				document				

(ISSN 2754-0820)

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

The third trend in comparing LBs reveals a difference in their use by NWs to realize the steps of "presenting positive justification" (M2 - S2) and "announcing present research descriptively and/or purposively" (M3 - S1). In contrast, EWs do not use any transparent LBs to express these two functions. This finding is displayed in Table 5 and suggests that NWs may be oversimplifying a construction that is more complex and less formulaic than what is employed by EWs.

NWs tend to use certain bundles, such as *The goal of this work is* (example 6) and *objective of this work is*, to perform the obligatory step of announcing the current research. In contrast, looking at the noun word list of CorCompInt-PLOS in Sketch Engine (Kilgarriff et al., 2014), the most frequent words used by EWs that relate to this step are "purpose" (202.71 occurrences PMW) and "goal" (184.14 occurrences PMW), neither of which appear in the n-grams list that meet our criteria for LBs in this corpus.

Table 5: *LBs used by NWs, but not used by EWs*

		CorCompInt-Lume
	Rhetorical structure	Lexical bundles
M2 - S2	Presenting positive justification	a better understanding of the
		a better understanding of
		better understanding of the
M3 - S1	Announcing present research descriptively and/or purposively	goal of this work
		goal of this work is
		goal of this work is to
		The goal of this
		objective of this work is
		objective of this work is to
		objective of this work
		The main objective of
		The main goal of
		The goal of this work is
		The goal of this work
		main objective of this work
		main objective of this work is
		main objective of this
		The main objective of this
		The main objective of this work
		This work focuses on
		the focus of this

Ex. 6. '<u>The goal of this work is</u> to present concrete benchmark results of different libraries, APIs, platforms, and implementation techniques for the matrix decomposition problem, and further analysis into the ways that such tools improve (or not) the efficiency of the implementation for similar problems.' (CorCompInt-Lume)

Table 6 highlights LBs used by EWs but absent in the use of NWs when performing the steps of "claiming centrality" (M1 - S1), "announcing principal outcomes" (M3 - S5), and "stating the value of the present research" (M3 - S6), according to the established cut off points. Surprisingly, only four LBs were found in this category: *plays an important role, results show that the, the results of the* and *contributions of this paper*, all of which are exemplified below.

This fourth trend suggests that NWs may not be familiar with the formulaic nature of the phraseology used to fulfil the rhetorical functions of these three steps. This finding underscores the necessity of explicit instruction on the use of LBs in academic writing and highlights the potential benefits of providing NWs with access to formulaic language resources to enhance their proficiency in constructing texts that fulfil the rhetorical steps in a formulaic manner within the computer science field.

Table 6: *LBs used by EWs, but not used by NWs*

		CorCompInt-PLOS
	Rhetorical structure	Lexical bundles
M1 - S1	Claiming centrality	plays an important role
M3 - S5	Announcing principal outcomes	results show that the
		the results of the
M3 - S6	Stating the value of the present research	contributions of this paper

Ex. 7. ‘Software development effort estimation plays an important role in the software engineering field.’ (CorCompInt-PLOS)

Ex. 8. ‘The results show that the hybrid model proposed in this paper has higher prediction accuracy than other models.’ (CorCompInt-PLOS)

Ex. 9. ‘Specifically, we use reverse annealing to explore local minima near an initial state defined by the results of the previous iteration of the algorithm.’ (CorCompInt-PLOS)

Ex. 10. ‘The contributions of this paper are listed as follows.’ (CorCompInt-PLOS)

In conclusion, the analysis of the data extracted from both corpora revealed four distinct trends. Firstly, both groups of writers exhibited a lack of using specific lexical bundles (LBs) to realize certain steps. Secondly, there were instances of similar or identical use of LBs by both groups. Thirdly, NWs used LBs that were not utilized by EWs. Lastly, there were LBs used by EWs that were not employed by NWs. These findings indicate potential differences in the usage and understanding of LBs between the two groups. Additionally, the study highlights the importance of explicit instruction and access to formulaic language resources to enhance academic writing proficiency, particularly within the computer science discipline.

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

4 Pedagogical implications

The main contribution of this paper is to inform EAP material designers of relevant linguistic features in a specialized genre. Incorporating the findings discussed in the section above into the design of EAP materials is essential for enabling NWs to produce more formulaic and conventional texts. This, in turn, increases the likelihood of their research being recognized within the specific disciplinary community of computer scientists.

In Brazil, there is a rising demand for the creation of resources that cater to the needs of the expanding community of emerging researchers in computer science. This is evident due to the notable inclination among undergraduate students in this discipline to actively engage in research communities through the English language from the early stages of their academic journey. A significant indicator of this trend is the growing number of students choosing to compose their undergraduate theses in English. As previously noted, the most extensive collection of English-language undergraduate theses on Lume is centred around this particular field.

Despite the significance of corpus linguistics in identifying patterns in authentic language use and advocating for the integration of corpus data into EAP classes, its actual implementation in classrooms globally is still in the early stages. A host of challenges, including time constraints, large numbers of students in class, and technological barriers have been recognized as impediments to its widespread adoption (Kavanagh, 2021). Moreover, many educators lack a basic understanding of the underlying principles of corpus linguistics. Additionally, a significant number of papers in this field often leave readers in the dark about potential classroom applications of the extracted data, as they do not propose any tasks or activities that incorporate the findings.

Previous studies that investigated the pedagogical work with LBs, found that these units are not easily acquired in the short term (Cortes, 2007). However, once students become proficient in using them, there is a positive impact on their writing grades (Kazemi et al., 2014). Therefore, we now move on to suggest a task based on our results that can enlighten EAP material designers and instructors to work with data extracted from our corpora.

Figure 2: *Tasks.*

The framework below presents moves and steps that can occur in the introductions of research articles.

Table 1: *Moves and steps in research article introductions (Cortes, 2013, p. 37, adapted from Swales, 1990, 2004)*

Move 1 Establishing a territory
Step 1 Claiming centrality
Step 2 Making topic generalizations
Step 3 Reviewing items of previous literature
Move 2 Establishing a niche
Step 1A Indicating a gap or
Step 1B Adding to what is known
Step 2 Presenting positive justification
Move 3 Presenting the present work
Step 1 Announcing present research descriptively and/or purposively
Step 2 Presenting research questions or hypothesis
Step 3 Definitional clarifications
Step 4 Summarizing methods
Step 5 Announcing principal outcomes
Step 6 Stating the value of the present research
Step 7 Outlining the structure of the paper

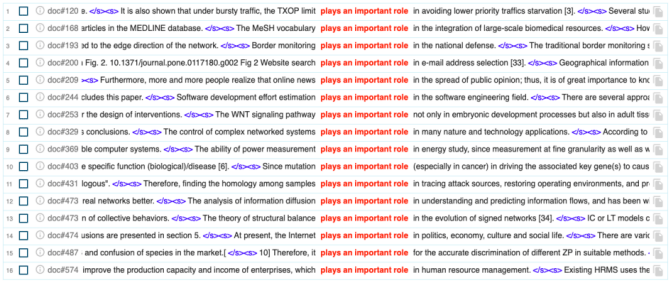
(a) Based on your prior experience with texts in your field, mark with a check (✓) which moves and steps you believe are obligatory in computer science research.

(b) Match the concordance lines (Screenshots 1, 2 and 3) with the steps from the framework you think the authors are conveying with the formulaic language in red.

Screenshot 1:
KWIC "plays an important role"

Move ( )

Step ( )



Screenshot 2:
KWIC "results SHOW that the"

Move ( )

Step ( )

Screenshot 3:
KWIC "contributions of this paper"

Move ( )

Step ( )

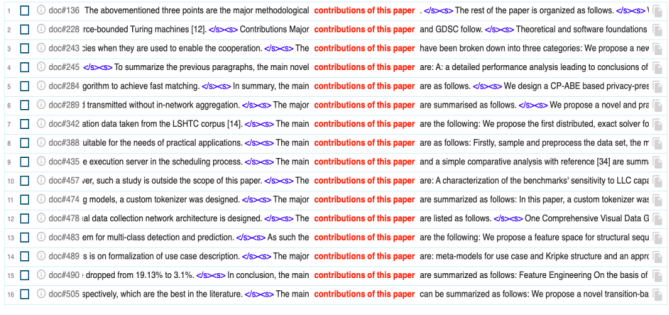


Figure 2 presents tasks that exemplify indirect data-driven learning (Johns, 1990), where students are presented with concordance lines extracted beforehand by the teacher, allowing them to analyse and

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

draw conclusions independently. The primary objective of the tasks is to enhance learners' rhetorical awareness, perception of disciplinary variation, and recognition of formulaic language. This is achieved specifically by introducing and highlighting LBs used by EWs that are absent in NWs writing so that these linguistic features can become a part of the repertoire of the latter group.

5 Conclusion

This study aimed to compare the usage of 4-, 5-, and 6-word bundles between two distinct corpora: one composed of texts written by NWs (CorCompInt-Lume) and another containing texts written by EWs (CorCompInt-PLOS). The former comprises introduction sections of undergraduate theses produced by novice writers in the field of computer science from Brazilian universities. The latter consists of introduction sections from research articles published in PLOS One, written by experienced writers in the same field. In total, 245 LBs were extracted, and 42% of them were classified based on the rhetorical functions they linguistically realize. The goal was to use language data extracted from a corpus to inform the design of discipline-specific EAP materials, aiming to address the specific needs of students rather than relying on generic "one-size-fits-all" EAP resources widely available.

Upon contrasting the two corpora, four distinct trends emerged in the use of LBs. These trends are as follows: (i) both groups of writers showed a lack of LBs usage to realize certain rhetorical functions expressed; (ii) both groups of writers demonstrated similar and/or identical usage of LBs; (iii) NWs utilized LBs that EWs did not use; and (iv) EWs used LBs that were not used by NWs.

The first trend in the data may not be the primary concern for EAP material designers who are aiming to teach conventional language chunks. This is because our results yielded no LBs that clearly realize some steps. In fact, our results suggest a lack of formulaic structures in realizing the following steps: "making topic generalizations" (M1 - S2), "reviewing previous literature" (M1 - S3), "adding to existing knowledge" (M2 - S1B), "presenting research questions or hypotheses" (M3 - S2), "providing definition clarifications" (M3 - S3), and "summarizing research methods" (M3 - S4). The non-identification of LBs in these steps within the scope of our study does not necessarily imply their absence in the texts. Despite their unconventional realization through LBs in our corpora, the emphasis on these steps in teaching the rhetorical structure of research articles should persist, remarking that some steps do not seem to follow a formulaic pattern in the field of computer science.

Still regarding the first trend, it is worth mentioning that our results do not merge with those of Cortes (2013), who investigated LBs in a corpus of research article introductions across thirteen disciplines, being computer science one of them. The author did find LBs that express the six steps previously mentioned. This discrepancy might indicate that our results corroborate the fact that disciplinary variation exists and that a closer look at a discipline-specific sample can reveal different patterns.

As per the second trend, both groups of writers showed similar and/or identical usage of LBs in realizing the rhetorical functions related to "indicating a gap" (M2 - S1A) and "outlining the structure of the paper" (M3 - S7).

The third and fourth trends in the data highlight the teaching gap - the areas in which EAP professionals should pay close attention while making decisions about what to bring into their classrooms. The third trend indicates that NWs may oversimplify a construction that experienced writers build in a more complex and less formulaic manner. These constructions are responsible for "presenting positive justification" (M2 - S2) and performing the rhetorical function of "announcing present research descriptively and/or purposively" (M3 - S1). In contrast, the fourth trend suggests that

NWs may not be aware of the formulaic nature of the phraseology used to perform the rhetorical functions of "claiming centrality" (M1 - S1), "announcing the principal outcomes" (M3 - S5) and "stating the value of the research" (M3 - S6).

To bridge the function-form gap and support the development of effective academic writing skills, EAP material designers should focus on addressing the third and fourth trends identified in the data. The third trend highlights the oversimplification of certain steps by NWs in "presenting positive justification" (M2 - S2) and "announcing present research descriptively and/or purposively" (M3 - S1). The fourth trend reveals NWs' lack of awareness of formulaic phraseology that could improve the construction of their research text by conventionally expressing the steps of "claiming centrality" (M1 - S1), "announcing principal outcomes" (M3 - S5), and "stating the value of the present research" (M3 - S6).

In contrast, the second trend, which describes the similar and/or identical use of LBs by both groups of writers, should not be the main concern of EAP instruction. While it is important for students to acquire these LBs through exposure to high-standard works, it may be worth considering lexical and frequency discrepancies based on the specific goals of the lesson.

When designing materials, it is important to consider the language elements where NWs differ from EWs. By analysing the written production of both groups, for example, if NWs already demonstrate proficiency in employing certain linguistic constructions, it may be unnecessary to design instructional materials that solely target those aspects. Instead, instructional efforts should be directed toward identifying and addressing the distinctive language patterns exhibited by NWs compared to their expert counterparts. This approach allows for more targeted and effective instruction that addresses the specific linguistic needs and challenges faced by NWs, facilitating their development toward higher levels of writing proficiency.

It is also important to note that the lexical differences observed in the study may be related to the fact that we compared the introduction sections of two different genres: undergraduate thesis and research articles. For example, LBs containing the words “work” and “document” were exclusively found in CorCompInt-Lume, while LBs with the word “paper” were exclusively found in CorCompInt-PLOS. This highlights the importance of considering genre-specific language use when designing EAP materials.

This research is not without its limitations. The main one is that our samples, which are rooted in different genres and sourced from only two databases: PLOS One and Lume. Further research is needed to explore the relationship between LBs and moves and steps from other sources, genres and sections of computer science academic texts, beyond the scope of this study.

The results presented in this paper could be valuable for future research aimed at developing an automated method for classifying LBs in different sections, building upon the foundation established by the present study. By incorporating machine learning techniques, an automatic classification system for LBs could be developed, facilitating more efficient analysis and interpretation of these linguistic units.

Furthermore, it is crucial to consider alternative approaches for studying LBs in order to gain a more comprehensive understanding of these units. It is also important not to overstate the significance of conducting discipline and section-specific studies. Investigating LBs that are more recurrent in certain disciplines and specific sections of research articles holds immense value. This is particularly true when it comes to informing the design of EAP writing materials. These discipline and section-specific studies will contribute significantly to enhancing our understanding of the utilization of

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

formulaic language in academic writing and its role in disciplinary variation, especially in shaping the development of targeted EAP writing resources.

6 References

Al-Shujairi, Y., & Al-Manaseer, F. (2022). Backgrounding the discussion section of medical research articles. *Open Journal of Modern Linguistics*, 12, 71-88. 10.4236/ojml.2022.121008

Anthony, L. (2022). *AntCorGen* (Version 1.2.0) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software/antcorgen/

Bennet, G. R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Ann Arbor: University of Michigan Press.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd Ed.). Cambridge: Cambridge University Press.

Biber, D., Connor, U. & Upton, T. A. (2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins Publishing.

Biber, D., Conrad, S, & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of Spoken and Written English*. Amsterdam: John Benjamins Publishing.

Cortes, V. (2007). Teaching lexical bundles in the disciplines: an example from a writing intensive history class. *Linguistics and Education*, 17(1), 391-406.

Cortes, V. (2013). The purpose of this study is to: connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12(1), 33–43.

Frankenber-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23–39.

Gil, N. N., & Caro, E. M. (2019). Lexical bundles in learner and expert academic writing. *Bellaterra Journal of Teaching & Learning Language & Literature*, 12(1), 65-90.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109-135.

Hamp-Lyons, L. (2011). English for Academic Purposes. In E. Hinkel (Ed), *Handbook of Research in Second Language Teaching and Learning*, Vol. II. (pp. 89-105). Routledge.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.

Hyland, K. (2016). General and specific EAP. In K. Hyland, & P. Shaw (Eds), *The Routledge Handbook of English for Academic Purposes*. (pp. 17-29). Oxon: Routledge.

Hyland, K., & Shaw, P. (2016). Introduction. In K. Hyland, & P. Shaw (Eds), *The Routledge Handbook of English for Academic Purposes*. (pp. 1-13). Oxon: Routledge.

Johns, T. (1990). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10(1), 14-34.

Kavanagh, B. (2021). Bridging the gap from the other side: how corpora are used by English teachers in Norwegian schools. *Nordic Journal of English Studies*, 20(1), 1–35.

Kazemi, M., Katiraei, S., Rasekh, A. E. (2014). The impact of teaching lexical bundles on improving Iranian EFL students' writing skill. *Procedia - Social and Behavioral Sciences*, 98(1), 864-869.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7-36.

Krieger, M. G., & Finatto, M. J. B. (2004). *Introdução à Terminologia: Teoria e Prática*. Contexto.
Microsoft. (2022). *Word* (Version 16.60) [Computer Software]. Redmond: Microsoft. Available from https://www.microsoft.com/en-us/microsoft-365/microsoft-office
Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *Journal of English for Specific Purposes*, 50(1), 40–63.
Murray, N. (2016). An academic literacies argument for decentralizing EAP provision. *ELT Journal*, 70(4), 435-443.
Neely, E., & Cortes, V. (2009). A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value*, 1(1), 17-38. https://www.e-revistes.uji.es/index.php/languagevalue/article/download/4731/4783/
Nesi, H. (2013). ESP and Corpus Studies. In B. Paltridge & S. Starfield (Eds) *The Handbook of English for Specific Purposes* (pp. 406-426). Chichester: John Wiley & Sons, Inc.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics*, 31(4), 487-512.
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, 12(1), 214–225.
Swales, J. (1981). *Aspects of Article Introductions*. Birmingham: The University of Aston.
Swales, J. M, & Feak, B. C. (2012). *Academic Writing for Graduate Students* (3rd Ed.). Ann Arbor: University of Michigan Press.
Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
Swales, J. M. (2004). *Research Genres: Exploration and Applications*. Cambridge: Cambridge University Press.
Swales, J. M. (2019). The futures of EAP genre studies: a personal viewpoint. *Journal of English for Academic Purposes*, 38(1), 75-82.

6 Appendices

6.1. Appendix One: Raw frequency, relative frequency, and document frequency of transparent lexical bundles extracted from CorCompInt-PLOS

Lexical bundle	Raw frequency	Relative frequency	Document frequency
is organized as follows	110	170.21645	110
paper is organized as	98	151.64739	98
paper is organized as follows	97	150.09996	97
this paper is organized	42	64.99174	42
the best of our knowledge	42	64.99174	40
the best of our	42	64.99174	40
best of our knowledge	42	64.99174	40

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

of this paper is organized	41	63.44431	41
this paper is organized as follows	40	61.89689	40
this paper is organized as	40	61.89689	40
of this paper is organized as	39	60.34947	39
The remainder of this	38	58.80205	38
The remainder of this paper	35	54.15978	35
remainder of this paper	35	54.15978	35
The remainder of this paper is	33	51.06494	33
remainder of this paper is	33	51.06494	33
To the best of	31	47.97009	30
the paper is organized	30	46.42267	30
of the paper is organized	30	46.42267	30
To the best of our knowledge	29	44.87525	28
To the best of our	29	44.87525	28
the paper is organized as follows	29	44.87525	29
the paper is organized as	29	44.87525	29
of the paper is organized as	29	44.87525	29
the structure of the	27	41.7804	23
The rest of the paper is	26	40.23298	26
The rest of the paper	26	40.23298	26
rest of the paper is	26	40.23298	26
rest of the paper	26	40.23298	26
rest of this paper	25	38.68556	25
The rest of this	23	35.59071	23
The rest of this paper	22	34.04329	22
results show that the	22	34.04329	13
remainder of this paper is organized	22	34.04329	22
rest of this paper is	21	32.49587	21
rest of the paper is organized	21	32.49587	21
The rest of this paper is	20	30.94845	20
rest of this paper is organized	18	27.8536	18
is structured as follows	18	27.8536	18
contributions of this paper	18	27.8536	18
paper is structured as follows	17	26.30618	17
paper is structured as	17	26.30618	17
the results of the	16	24.75876	15
plays an important role	16	24.75876	15
The remainder of the	15	23.21133	15

2.2 Appendix Two: Raw frequency, relative frequency, and document frequency of transparent lexical bundles extracted from CorCompInt-Lume

Lexical bundle	Raw frequency	Relative frequency	Document frequency
is organized as follows	58	334.54269	58
work is organized as	39	224.95112	39
work is organized as follows	36	207.64718	36
This work is organized	20	115.35955	20
This work is organized as	18	103.82359	18
this work is organized	17	98.05561	17
of this work is organized as	17	98.05561	17
this work is organized as follows	17	98.05561	17
this work is organized as	17	98.05561	17
of this work is organized	17	98.05561	17
goal of this work	15	86.51966	14
goal of this work is	15	86.51966	14
This work is organized as follows	15	86.51966	15
goal of this work is to	13	74.98371	12
The goal of this	12	69.21573	12
objective of this work is	11	63.44775	11
objective of this work is to	11	63.44775	11
objective of this work	11	63.44775	11
The remainder of this	11	63.44775	11
The remaining of this	9	51.9118	9
is structured as follows	9	51.9118	9
The main objective of	8	46.14382	8
document is organized as	8	46.14382	8
rest of this work	8	46.14382	8
The main goal of	7	40.37584	7
The goal of this work is	7	40.37584	7
The goal of this work	7	40.37584	7
remainder of this work is	7	40.37584	7
remainder of this work	7	40.37584	7
remainder of this work is organized	7	40.37584	7
document is organized as follows	7	40.37584	7
This work is divided	7	40.37584	7
the organization of the	7	40.37584	7
The remainder of this work is	7	40.37584	7
The remainder of this work	7	40.37584	7
best of our knowledge	6	34.60786	6
the best of our knowledge	6	34.60786	6
the best of our	6	34.60786	6

Received: 04/06/2023
Revisions: 24/01/2024
Accepted: 08/02/2024

a better understanding of the	6	34.60786	6
a better understanding of	6	34.60786	6
better understanding of the	6	34.60786	6
main objective of this work	6	34.60786	6
main objective of this work is	6	34.60786	6
main objective of this	6	34.60786	6
The main objective of this	6	34.60786	6
The main objective of this work	6	34.60786	6
This work focuses on	6	34.60786	6
the focus of this	6	34.60786	6
goals of this work	6	34.60786	6
The rest of this work is	6	34.60786	6
The rest of this work	6	34.60786	6
rest of this work is organized	6	34.60786	6
rest of this work is	6	34.60786	6
this document is organized as	6	34.60786	6
this document is organized	6	34.60786	6
are presented in Chapter	6	34.60786	5
presents an overview of	6	34.60786	5
of this document is	6	34.60786	6

About the Authors

Wesley Henrique Acorinti is an instructor in Portuguese as an Additional Language at the Federal University of Health Sciences of Porto Alegre (UFCSPA). He began his BA degree in Language Teaching at the Federal University of Rio Grande do Sul (UFRGS) in 2020, specialising in Portuguese Language, English Language, and Literature. From 2021 to 2023, he worked as an undergraduate research assistant at UFRGS, conducting research in Applied Linguistics with the support of the Foundation for the Support of Research in the State of Rio Grande do Sul (FAPERGS). He is expected to graduate in mid-2024. His primary research interests include English for Specific Purposes, Corpus Linguistics, and Portuguese as an Additional Language.

Contact: wesley.acorinti@ufrgs.br

Ana Eliza Pereira Bocorny is an Associate Professor at the Department of Modern Languages at the Federal University of Rio Grande do Sul (UFRGS), Brazil. She holds a Ph.D. in Language Studies (UFRGS), a Master's in Education (PUCRS) and a specialization in English for Specific Purposes (Lancaster University - UK). Her research interests lie in Applied Linguistics, English for Specific Purposes, English for Academic Purposes and Corpus Linguistics.

Contact: ana.bocorny@gmail.com