e for learning English.</s><s>This may include EAP (
on, Canada so that a student can complete our EAP (
y students.</s><s>I'm TESOL certified to teach EAP (
ents is:</s><s>This series from award-winning EAP (
not apply; however, six credits of college-level EAP (
or at Emory University's Candler School of Theology (
y that ... Continue reading →</s><s>OXFORD EAP (
rted my second year of teaching at BU with the EAP (
hing English on the BU campus in the EAP program (
is article provides a guide to the award-winning EAP (
Edward de Chazal explains the challenges that EAP (
ses, and adjunct professor for E.</s><s>A.</s><s>P (
interests include second language acquisition, EAP (
l is required.</s><s>Students take prerequisite EAP (
onventions.</s><s>Despite the efforts of many EAP (
ernational students at colleges and universities EAP (
survey.</s><s>Theoretical Background</s><s>EAP (

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

SIMONE SARMENTO, ROZANE REBECHI, MARINE LAÍSA MATTE (ORG.)

), which prepares students at tertiary level for further a
) Program (Level 10 with 80%) and then enter the univ
) which means that I must be knowledgeable in all aca
) author Aylin Graves provides a set of lesson plans to
) coursework taken at Florida SouthWestern State Col
program).</s><s>He is also academic director of UGA
) B1+ INTERMEDIATE - components</s><s>This diss
) program.</s><s>I teach level 2 writing every morning
).</s><s>Classes consist of International students fror
) series from author, Aylin Graves.</s><s>Approaches
) learners face, and what teaching staff and lecturers r
) courses.</s><s>She has spent many hours in the cla
), translation, interpreting, education quality assessme
) courses in reading, listening, writing, and research be
) researchers and practitioners to provide support for r
) ELT (Enhanced Language Training) ESP (English for
) researchers, such as Christison and Krahnke, 1986;

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

SIMONE SARMENTO
ROZANE REBECHI
MARINE LAÍSA MATTE
(ORG.)

Porto Alegre • 2024 • 1ª edição

editora ZOUK

Projeto gráfico e edição: Editora Zouk
Revisão: Simone Sarmento, Rozane Rebechi, Marine Laísa Matte
Imagem da capa: SKELL

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Elaborado por Vagner Rodolfo da Silva - CRB-8/9410

E58
English for Academic purposes [recurso eletrônico] : reflections, description e pedagogy / organizado por Simone Sarmento, Rozane Rebechi, Marine Laisa Matte. - Porto Alegre, RS : Zouk, 2024.
268 p. ; ePUB.

Inclui bibliografia.
ISBN: 978-65-5778-135-7 (Ebook)

1. Linguística. I. Sarmento, Simone. II. Rebechi, Rozane. III. Matte, Marine Laisa. IV. Título.

CDD 410
2024-175 CDU 81'1

Contents

Exploring the complexities of EAP: a collection of voices
Simone Sarmento, Rozane Rebechi and Marine Laísa Matte
7

The role of Corpus Linguistics in EAP
Deise P. Dutra and Tony Berber Sardinha
14

From specialized corpus to the EAP classroom: integrating authentic data into materials design
Ana Eliza Pereira Bocorny, Ana Luiza Freitas and Rozane Rodrigues Rebechi
55

Do-It-Yourself Corpora to Support SHAPE and STEM Research Paper Writing
Paula Tavares Pinto, Luciano Franco da Silva, Talita Serpa and Diva Cardoso de Camargo
97

Creating a local learner corpus: Insights on project design and data analysis from the pilot phase
Sandra Zappa-Hollman, Alfredo Afonso Ferreira, Greta Perris, Simone Sarmento, Marine Laísa Matte and Laura Baumvol
127

The role of genre in academic language use: the case of Critiques and Case Studies in BAWE
Marine Laísa Matte, Deise Amaral and Larissa Goulart
155

Investigating Brazilian English Learners' Use of Academic Collocations: A Corpus-Based Study
Marine Laísa Matte and Simone Sarmento
178

From corpus to classroom: evaluating Web-based tools to teach collocations
Larissa Goulart, Maria Kostromitina and Jennifer Klein
204

Driving forces to adopt EMI: scholars' perceived benefits of English medium of instruction in Brazilian higher education
Laura Baumvol, Lucas Marengo and Simone Sarmento
243

About the authors
263

Exploring the complexities of EAP: a collection of voices

Simone Sarmento (UFRGS)
Rozane Rebechi (UFRGS)
Marine Laísa Matte (UFRGS/IFSul)

In this introduction, we aim to discuss aspects related to English for Academic Purposes (EAP), to highlight the significance of this collection to the broader field of EAP, and to provide a brief overview of the book and its contributions.

EAP refers to the study and use of English in academic settings, with a focus on the development of the language skills necessary to succeed in higher education (Hyland, 2009). This includes the improvement of competencies in academic reading, writing, listening, and speaking, as well as the ability to understand and produce discipline-specific vocabulary and discourse. The field has become increasingly important in recent years, as the demand for English language proficiency continues to grow in academic contexts around the world. As a result, there has been a surge of research and teaching practices focused on language skills and competencies required for academic success, from writing research papers to participating in academic discussions (Biber, 2006).

With the field of EAP being an active area of research, new studies are being published regularly. These studies usually rely on a myriad of methods, since different methodological procedures can be employed to answer research questions related to the use of academic language. Among them, Corpus Linguistics (CL) comes as a highly productive research methodology for investigating the demands of academic communication, including the usage of language. One of the greatest contributions of CL to the field of EAP is that it enables access to large amounts of authentic language data, which can be used to identify and analyze the lexical, grammatical, and discourse features of academic language (Nesi, 2016). As a result, EAP

researchers and instructors can identify the most frequent and relevant language patterns creating targeted language learning materials and activities for students. Thus, students can develop their own academic writing and speaking skills by studying and practicing how to use language patterns and structures that are typical of academic discourse.

Finally, CL can facilitate the identification of patterns among different academic disciplines, enabling instructors to tailor their EAP teaching to the students' specific needs in different fields. For example, the language used in medical research papers is likely to differ from that used in humanities papers, and CL creates opportunities to identify these patterns, allowing instructors to provide targeted support to students based on their individual needs. As we will show below, seven out of the eight chapters in this book use CL to varying degrees, exemplifying the productivity of corpus-based research for the field of EAP.

EAP also encompasses English as a Medium of Instruction (EMI), a relatively new branch of EAP, in which the English language is used as the primary means for delivering academic content and facilitating communication in a multilingual academic environment (Macaro, 2017). EMI in higher education settings refers to the use of English as the primary language of instruction for academic courses or programs in universities and other higher education institutions where the students' first language is not English. The use of EMI in higher education can offer learners several benefits, such as the opportunity to study in an international environment, exposure to English-language academic literature and research, and the development of language skills that can enhance future academic and professional opportunities. However, EMI also poses challenges, such as ensuring that students have sufficient language proficiency to understand the subject matter and instruction and that instructors are able to deliver high-quality instruction in English (Marengo, 2022). Research on EMI seeks to better illuminate the benefits and challenges of using English as a teaching language and to identify effective strategies and best practices for promoting both language and subject learning in EMI settings.

This book, entitled "English for Academic Purposes: Reflections, description & pedagogy", brings together nine chapters (the first being this

introduction) written by a diverse group of scholars and practitioners from different universities that share a common interest in exploring the complexities of academic language and communication. The contributors offer unique perspectives on the possibilities, challenges, and opportunities of researching, teaching, and learning EAP.

This book is a valuable resource for the field of English for Academic Purposes (EAP) for several reasons. First, it provides a diverse range of perspectives on the challenges and opportunities of teaching and learning EAP. As EAP is a broad and complex field, encompassing various academic contexts and language skills, this collaborative effort offers unique insights that can enrich understanding within the field and inspire new approaches to support students in their academic language development.

Second, this book demonstrates the complexity of the field of EAP by presenting a range of different research initiatives. It highlights the numerous factors that can impact language learning and use in academic settings, which can inform the design of effective language teaching and learning materials.

Lastly, this book encourages collaboration and dialogue by bringing together a diverse group of scholars and practitioners. This collaborative approach is intended to foster a sense of community and shared purpose within the field of EAP, leading to the development of new ideas and approaches to teaching and learning. In summary, this book is an important contribution to the field of EAP as it provides a platform for advancing research and practice. We now provide a brief overview of the next chapters.

In the second chapter of this book, Deise Prina Dutra and Tony Berber Sardinha provide a comprehensive overview of English for Specific Purposes (ESP), a field that has experienced considerable growth and development over the past decades. Within ESP, EAP has emerged as a key area of focus, with studies from a CL perspective providing insights into the characteristics of academic speech and writing. This chapter explores the contribution of general, specialized, and learner corpora to EAP research and practice, with a particular focus on how corpus-based approaches have influenced the study of vocabulary and grammar in academic texts. The authors review the major literature on corpus-based research in EAP

and highlight the ways in which multi-dimensional analysis can provide a deeper understanding of the underlying patterns of lexico-grammatical characteristics in academic writing. By examining these patterns, the authors shed light on some of the differences across academic registers that have previously been overlooked in the field.

In recent years, the integration of corpus-based language learning and teaching has gained attention in the field of English for Academic Purposes (EAP). Despite the potential benefits of using corpus data in EAP pedagogy, the application of corpus-based approaches in Brazilian EAP classrooms is still limited. This issue is addressed in the third chapter of this book, authored by Ana Eliza Pereira Bocorny, Ana Luiza Freitas, and Rozane Rebechi. The chapter provides a practical guide for EAP teachers on how to integrate corpus data into materials designed for EAP writing courses. The authors review corpus and genre-based approaches to language learning and teaching, besides describing a framework and principles for the design of EAP materials that combine these pedagogies. The chapter concludes by highlighting the feasibility of the application of genre-based corpus linguistics for both novice and experienced teachers, who can use the step-by-step guide to integrate corpus and genre-based approaches for academic writing in their classrooms. This chapter will be of interest to anyone seeking to enhance their understanding of the potential of corpus-based pedagogy in EAP, particularly novice EAP teachers.

Chapter 4, authored by Paula Tavares Pinto, Luciano Franco da Silva, Talita Serpa, and Diva Cardoso de Camargo, explores the potential of using do-it-yourself corpora to support academic writing and translation in the areas of humanities, science, and math. The authors demonstrate how to quickly compile two specialized corpora in SHAPE (Social Sciences Humanities, Arts for People and Economy) and STEM (Science, Technology, Engineering, and Mathematics) areas with the tool AntCorGen and explore them with Sketch Engine to help researchers write their own research papers. By examining the corpora, readers can identify frequently used adjectives, verbs, and lexical bundles, as well as recurrent academic structures for each research paper section, such as the Introduction, Methodology, Discussion, and Conclusions. The chapter offers practical guidance

for researchers who wish to use corpora to enhance their academic writing skills.

In Chapter 5, Sandra Zappa-Hollman, Alfredo Afonso Ferreira, Greta Perris, Simone Sarmento, Marine Laísa Matte, and Laura Baumvol report on their experiences designing and piloting a local learner corpus for use by instructors, students, and researchers at a Canadian university that offers first-year undergraduate programs for speakers of English as an additional language. This project was motivated by the need for data-driven instruction and research, and the authors present the stages of conducting the project, highlighting the importance of collaborative teamwork, and sharing the results of initial data analysis for pedagogical and research applications.

Chapter 6 focuses on how genre mediates variation in language, indicating that different communicative purposes are expressed through the use of different linguistic features. Marine Laísa Matte, Deise Amaral, and Larissa Goulart analyze the variation of linguistic features associated with academic writing in two genres of university assignments: Case Studies and Critiques from the BAWE (British Academic Written English) corpus. Mann-Whitney U tests indicate that there is variation in the use of features between the two genres, with a higher frequency of features in Critiques. The study reveals that, although the two genres share the same features, their usage is mostly diverse as they serve different communicative objectives. This finding suggests that different genres have specific language requirements, which can influence the way in which authors express their ideas and communicate with their readers.

In the seventh chapter, Marine Laísa Matte and Simone Sarmento explore the role of collocations in EAP. Collocations are words that frequently occur together due to their attraction, and their appropriate use is indispensable for ensuring fluency and accuracy in written communication. In this study, the authors analyze how Brazilian students produce collocations in academic texts written in English. The analysis is based on a list of 125 nodes and their corresponding collocates in a comparison between the Brazilian Academic Written English (BrAWE) corpus and the BAWE corpus. The findings indicate that, overall, the nodes are underused

in BrAWE. The study shows a balance of syntatic structures being used in both corpora. Also, this research also reveals that Brazilian students use a limited variety of collocations when compared to students in BAWE.

In recent years, Web-based Learning Tools (WBLTs) that use CL research have become a popular way of teaching learners how to use collocations. In chapter 8, Larissa Goulart, Maria Kostromitina, and Jennifer Klein evaluate the effectiveness of five WBLTs - FLAX, SKELL, Linggle, Just the Word, and Netspeak - aimed at helping learners of English produce accurate collocations. The evaluation is divided into three parts: research conducted in the development of the WBLT, the WBLTs design and accessibility, and WBLT pedagogical applications. The results of the study show that most of these tools rely on frequency-based collocations and contribute to different types of class activities. The authors finish the chapter by proposing task ideas for using these tools in the English language classroom.

In the last chapter of this collection, Laura Baumvol, Lucas Marengo, and Simone Sarmento explore the concept of EMI. EMI is an approach to teaching and learning in which English is the language of instruction, with the purpose of imparting a diverse range of contents through the medium of the English language, rather than teaching the language itself. This phenomenon is rapidly gaining ground on a global scale and is closely linked to the internationalization and globalization of higher education institutions. This study focuses on EMI practices in Brazil, using data collected through a large-scale questionnaire sent to higher education teachers across all regions and states of the country. The authors investigate whether EMI occurs in the eight different fields of knowledge as classified by Brazilian funding agencies and examine teachers' perceptions of the benefits, or lack thereof, of classes taught in English. The findings of the study indicate that EMI is more widely accepted in the "harder" sciences, such as biological sciences, health sciences, agricultural sciences, and STEM. On the other hand, the fields of the "softer" sciences, including human sciences and linguistics, literature, and arts, appear to be more cautious in adopting EMI in their practices.

We hope all these voices can reverberate, so that new avenues of research and teaching arise and foster dialogue around EAP!

Acknowledgements

This book has been supported by the Graduate Program in Linguistics and Literature at the Federal University of Rio Grande do Sul and CAPES' PROEX. Simone Sarmento holds a CNPq research productivity scholarship level 1D.

References

Biber, D (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins Publishing.

Hyland., K. (2009). *Academic discourse: English in a global context.* A&C Black.

Macaro, E. (2017). English medium instruction: Global views and countries in focus: Introduction to the symposium held at the Department of Education, University of Oxford on Wednesday 4 November 2015. *Language Teaching* 1–18. doi:10.1017/S0261444816000380.

Marengo, L. H. F. (2022). *The role of English language proficiency in Brazilian EMI practices*. [Unpublished master's thesis]. Federal University of Rio Grande do Sul.

Nesi, H. (2016). *Corpus studies in EAP*. K. Hyland, K. & P. Shaw (Eds.), The Routledge Handbook of English for Academic Purposes. (206-217) Routledge.

The role of Corpus Linguistics in EAP

Deise P. Dutra (UFMG)
Tony Berber Sardinha (PUC-SP)

Introduction

Since the 1960s a considerable portion of research about English has been intimately connected with the teaching and learning of English for specific purposes (ESP), a branch of applied linguistics which has evolved, especially from 1990 to 2020, to become "a mature discipline of global importance" (Hyland & Jiang, 2022: 23). For instance, such research has helped teachers and material designers by providing word frequency lists that can support class preparation and textbook writing (e.g., General Service List [GSL] by West, 1953; Academic Word List [AWL] by Coxhead, 2000).

ESP comprises several strands, including, among others, business English, aviation English, English for medical purposes, and English for academic purposes (EAP), which is the focus of this book. Unsurprisingly, studies from a corpus linguistics (CL) perspective have informed EAP practices, providing detailed descriptions of academic speech and writing "from lexical, phraseological, grammatical, and genre perspectives" (Nesi, 2016: 206).

Whether corpus is the backbone of teaching syllabus and reference materials, such as in dictionaries (Sinclair, 1987[1]), grammar books (Biber et al., 1999; Carter & McCarthy, 2006), and textbooks (McCarthy et al., 2014), or is used by teachers and students (Johns, 1991; Crosthwaite et al., 2021), these perspectives lead us to reflect on CL's pedagogical implications for language teaching and learning, especially on EAP. Römer (2010) views

1 Collins COBUILD English language dictionary was the first dictionary-based on corpus.

the pedagogical application of corpus as either indirect, as researchers and materials developers use corpora, or direct, when teachers and students are able to have their hands on corpus data. Researchers and material designers deal with corpora results when writing syllabi, textbooks, and reference materials included in other materials. They are the ones who deal with the data from the corpora and filter the relevant information for the audience and teaching context of the proposed material. Therefore, the pedagogical applications are indirect. Conversely, when teachers use corpora to prepare activities or have their students carry out corpus investigations, they are involved in the direct applications of CL through their teaching and learning experiences. Above all, when EAP teachers and students use corpus tools or have access to materials based on corpus, they have access to real language: "[T]he methodological paradigm of corpus research has a direct influence on what is regarded as reliable knowledge sources. Corpus investigations give primacy to data, that is, they prioritize empirical analyses of language use" (Viana & O'Boyle, 2022: 52).

In this chapter, we discuss how corpora studies relate to EAP, showing how they have impacted this area in different ways. We first review the major literature on corpus-based research into EAP vocabulary. Second, we focus on grammatical complexity corpus-based research and how it has affected and could better contribute to EAP. Finally, we discuss how multidimensional analysis (Biber, 1988) approaches to EAP can provide insights into the underlying patterns of lexico-grammatical characteristics found in academic texts, discussing how these patterns can reveal striking differences across academic registers,[2] some of which have been ignored in the field.

2 "… a register is a variety associated with a particular situation of use (including particular communicative purposes). The description of a register covers three major components: the situational context, the linguistic features, and the functional relationships between the first two components" (Biber & Conrad, 2009: 6).

Vocabulary through the lenses of CL: From lists of individual words to phraseological patterns

Corpus-based research may be motivated by teaching and/or learning issues. One of the areas with a direct connection to pedagogical implications (e.g., syllabus preparation, materials design and classroom tasks) is vocabulary, making corpus-generated frequency lists a valuable contribution to EAP. In this section, we concentrate on how word lists have evolved from general English to academic general English to better cater to EAP learners' needs. The aim is to relate CL contribution to the presented lists without exhaustively reviewing all corpus-generated vocabulary to date. Distinctions will be made between contributions that focus on individual vocabulary and on a phraseological perspective for list compilation.

Since West (1953 as cited in Coxhead, 2000) developed the GSL, a corpus-based 2,000-word family list for English as a Second Language (ESL)/ English as a Foreign Language (EFL) learners, it has been widely used by English language teachers. The GSL was compiled to support the teaching and learning of general English while being used as a reference for other lists, including the new AWL[3] (Coxhead, 2000). Following "the assumption that frequency and coverage are important criteria for selecting vocabulary" (Coxhead, 2000: 215), Coxhead considered these compilation criteria: representativeness (Biber, 1993), organization (subregisters' distribution across subject areas), corpus size (Sinclair, 1991), and word selection. To support EAP programs and students, the AWL was based on an academic register corpus with 28 subject areas distributed in four disciplines: arts, commerce, law, and science. The academic subregisters covered in Coxhead's academic corpus were articles, book chapters, course

3 Other academic lists were made available for teachers, students, and material designers in the 20th century (e.g., University Word List by Xue & Nation, 1984), but the AWL (Coxhead, 2000) was the first one based on a digitally compiled corpus. Xue and Nation (1984) used previously composed lists, mainly put together manually (Campion & Elley, 1971; Ghadessy, 1979; Lynn, 1973; Praninskas, 1972, as cited in Gardner & Davies, 2014).

workbooks, laboratory manuals, and course notes.[4] This corpus included 3.5 million words, yielding a list with 570 word families. The AWL's contribution to EAP is undeniable, and it has been influential "in setting vocabulary goals for language courses, guiding learners in their independent study, and informing course and material designers in selecting texts and developing learning activities" (Coxhead, 2000: 214). Criticisms, however, have been leveled against the AWL, especially due to its use of word families and its relationship to the GSL (Gardner & Davies, 2014). In addition, it has been challenged due to its listing of individual words and its basis not being an updated and larger corpus.

Other corpus-based studies have provided academic vocabulary lists (Ackermann & Chen, 2013; Biber et al., 1999; Biber et al., 2004; Gardner & Davies, 2014; Simpson-Vlach & Ellis, 2010)[5] based on larger corpora than the GSL and AWL and included information on word co-occurrence and phraseology. The recognition of phraseology as a central element of language is not novel in linguistics. Nearly 70 years ago, Firth (1957) claimed that to understand a word, it is necessary to consider the other words it co-occurs with. Sinclair's (1991) groundbreaking work in corpus linguistics using large collections of texts made it possible to find evidence of recurrent patterns of words and constructions, which led him to propose the idiom principle that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (p. 110). He further explored how this pervasive principle is productive in language in phrases such as "*set eyes on*," "*it's not in his nature to*" (Sinclair, 1991: 111), "*hard work*," and "*hard evidence*" (p. 112), defining the term "collocation" as "the occurrence of two or more words within a short space of each other in a text" (p. 170). As Ellis (2008: 9) metaphorically puts it, phraseology is everywhere in language: "Like blood in systemic circulation it flows through

4 https://www.wgtn.ac.nz/lals/resources/academicwordlist/information/corpus

5 Even if some of these publications, such as Biber et al. (1999), did not have a major goal of providing a list to EAP, as they carried out careful corpus-based research, they presented results that can be sources for data-based language materials and classes.

heart and periphery, nourishing all." Therefore, phraseology should be vital to language teaching in general and to EAP in particular.

Biber et al. (1999) introduced a particular kind of phraseological unit, which they termed lexical bundles. Lexical bundles are defined as "the sequences of words that most commonly co-occur in a register" (Biber et al., 1999: 989) and "serve the most important communicative needs of a register" (Biber, 2009: 285). Biber et al. (1999) analyzed their use in both conversation and academic prose, while Biber et al. (2004) showed how these units are used in university classroom teaching and textbooks. After generating a list of four-, five-, and six-word lexical bundles, Biber et al. (1999) analyzed them from a structural perspective (e.g., dependent clause fragment, such as *know what I mean*, and noun phrase of preposition phrase fragments, such as *the end of the*). As investigating the use of lexical bundles can contribute to our understanding of language use, Biber et al. (2004) presented not only structural, but also functional categories of lexical bundles. This frequency-driven study followed specific criteria for bundle inclusion for analysis—namely, a frequency cut-off point of 40 times per million words, a bundle word length of four, and the occurrence of the bundle in at least five different texts. Their corpus of classroom teaching and textbooks includes 2,009,400 words, which is not bigger than Coxhead's (2000) corpus. Nevertheless, Biber et al. (2004) compared their results to the Longman Spoken and Written English Corpus's (Biber et al., 1999) conversation section (7 million words of British and American English) and academic prose section. One of their major contributions was the detailed comparison across four registers (classroom teaching, textbooks, conversation, and academic prose), especially the presentation of a functional categorization of the bundles, which was also used in Biber (2006) to analyze other university registers (e.g., office hours, study groups, service encounters). Bundles were classified into four functions: stance expressions (e.g., *I don't know if*, *it is important to*), discourse organizers (e.g., *if you look at*, *on the other hand*), referential expressions (e.g., *that's one of the*, *as a result of*), and special conversation functions (e.g., *I said to him/her*). Biber et al. (2004) did not claim that their study could generate an academic list, but their results can inform EAP professionals of the most

important lexical bundles that students need to understand in both written and spoken higher education English, which adds a register perspective to our understanding of lexical bundle use.

The Academic Formula List (AFL; Simpson-Vlach & Ellis, 2010) expanded on the functional taxonomy provided by Biber et al. (2004), combining quantitative and qualitative criteria to include three to four n-grams in their list, which is also devoted to English used in the university context. Their methodology involved corpus statistics, linguistic analyses, psycholinguistic processing metrics, and EAP instructors' and language testers' insights, yielding a 435-lexical-bundle list. They used the Michigan Corpus of Academic Spoken English (MICASE) and the oral academic part of the British National Corpus (BNC), in addition to Hyland's 2008 corpus and written BNC files of various academic subjects. As the main purpose of creating a list such as the AFL was pedagogical, it is a valuable resource for EAP practitioners. The fact that they took into consideration professionals' perceptions when selecting the bundles as a refinement of what the quantitative analyses provided added pedagogical reliability to the list. EAP practitioners can use this lexical bundle list to inform class activities that go beyond the three major categories (referential expressions, stance expressions, and discourse markers) identified in Biber et al. (2004) and help learners develop an awareness of specific bundle functions as the AFL includes 18 subcategories, such as referential expressions of tangible framing attributes (e.g., *(as) part of [a/the]*, *the change in*), stance expressions of hedging (e.g., *(more) likely to (be)*, *[it/there] may be*), and discourse-organizing function expressions of metadiscourse and textual reference (e.g., *come back to*, *I'm talking about*). The list distinguishes bundles that are core AFL, meaning both frequent in oral and written academic language (e.g., *[a/the] result of*), and bundles that are more frequent in either spoken (e.g., *in order to get*) or written texts (e.g., *as a consequence*).

Another important contribution to EAP has been Ackermann and Chen's (2013) Academic Collocation List (ACL) because it was based on a large corpus, relied on both human judgment and quantitative analyses, and focused on lexical collocations. They used a written curricular component of the Pearson International Corpus of Academic English (PICAE)

comprising over 25 million words. Although Simpson-Vlach and Ellis (2010) also incorporated EAP practitioners' judgment in selecting the bundles, Ackermann and Chen's (2013: 236) list considered human judgment for both "the selection of lexical items for pedagogical purposes" and "for the refinement for the final listing." Their choice of creating a list with collocations is based on several studies (i.e., Nation, 2001; Nesselhauf, 2003, 2005) that pointed out the relevance of teaching collocations as they "are difficult to learn and retain even with the assistance of dictionaries" (Ackermann & Chen, 2013: 246). Above all, Nation (2001) argued that the frequency of academic collocations may not be enough for learning them implicitly. The ACL comes in handy for EAP practitioners as it includes 2,468 entries categorized in four types: noun combinations (adjective + noun or noun + noun; e.g., *anecdotal evidence*, *assessment process*); verb + noun / adjective combinations (e.g., *gather information*, *seem plausible*); verb + adverb combinations (e.g., *explicitly state*, *grow rapidly*); and adverb + adjective combinations (e.g., *highly controversial, (be) markedly different*). A crucial information of the ACL is the high percentage of occurrence of noun combinations: 74.3% (adjective + noun = 71.8% and noun + noun 2.5%; Ackermann & Chen, 2013: 241), leading the authors to suggest that both implicit and explicit collocation teaching is required to impact learners' understanding and production of academic English with high information load. These results support studies (Biber & Gray, 2010, 2016) that show how compressed academic language is, which is an issue that will be discussed in the next section.

Based on the 120-million-word academic subcorpus of the Corpus of Contemporary American English interface (COCA; Davies, 2008), the new Academic Vocabulary List (AVL) (Gardner & Davies, 2014) is an invaluable resource for EAP practitioners as it covers nine major disciplines (i.e., education, history, business and finance, medicine and health, law and political science, humanities, philosophy, religion and psychology, science and technology, and social science). In addition, it has been integrated into the COCA interface, allowing users to download it freely, input their texts, and get information about the word(s) of focus in many different ways. The search tool provides "(i) synonyms, (ii) definitions, (iii) relative frequency

across nine academic disciplines, (iv) the top collocates of the word, which provide useful insights into meaning, usage, and phrasal possibilities, and (v) up to 200 sample concordance lines" (Gardner & Davies, 2014: 325). Above all, this powerful resource, integrated into a user-friendly interface, grants students several possibilities to explore language, which could contribute to a more confident use of academic English. In Almeida et al. (2023), this interface is a tool to guide EAP students to reflect on the importance of collocates and how register affects the choices language users make. They can contrast examples from blogs, web, TV/movie, fiction, news, magazine, spoken, and academic registers. The series of activities proposed in their chapter uses information students can extract from accessing the "word" tool (Figure 1) in COCA to understand in which register certain verbs are more frequently used (e.g., *achieve*) and the noun collocates they often attract. The tasks culminate in focusing on the verb–noun collocations in the academic register. They lead students to fill in the blank of authentic sentences extracted from COCA and, finally, create their own texts using the verbs that more often occur in the academic register together with their appropriate noun collocates.

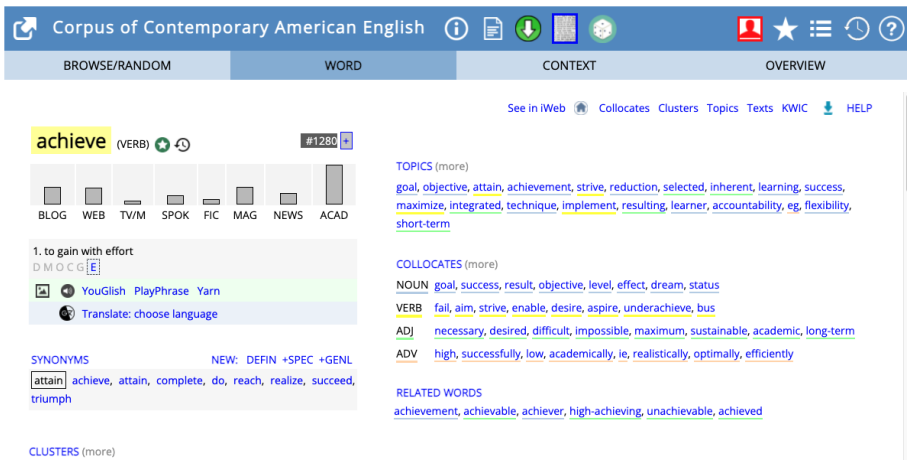


Figure 1. Collocates of *achieve* in COCA.

Other corpus-based studies have also dealt with corpora that allow for the investigation of variation across disciplines, specifically lexical

bundle variation (Cortes, 2013; Hyland, 2008; Lake & Cortes, 2020; Reppen & Olson, 2020). Differentiating between general and specific discipline lexical bundles meets one of EAP's demands to have materials for use in general and specific EAP courses. Hyland (2008), who compiled a corpus of articles, master's degree theses, and doctoral-level dissertations written in four areas (i.e., electrical engineering, biology, business studies, and applied linguistics), discovered that more than 50% of the lexical bundles were not common among the four areas. "The best candidate bundles for a general EAP course are *on the other hand*, *in the case of*, *as well as the*, and *the end of the*" (Hyland, 2008: 13). Taking a similar path as Hyland (2008) to uncover discipline variation, Reppen and Olson (2020) compiled a corpus of more than 25 million words from nine disciplines and 898 texts of textbooks, web pages, and academic articles. They examined more than 700 four-word lexical bundles, identifying cross-disciplinary and discipline-specific bundles:

> The bundles that occurred in four or more disciplines function as discourse frames providing signposts for readers (e.g., *on the other hand*, *the rest of the*; *in the case of*), while the discipline-specific bundles are often content or discipline specific (e.g., *of the interior design*, *role of hotel owners*). (Reppen & Olson, 2020: 172)

Having access to the cross-disciplinary and discipline-specific bundle lists, as presented in Reppen and Olson (2020), can make it easier for EAP instructors to prepare classes that cater to their students' needs. Activities with cross-disciplinary bundles are quite useful in EAP classes with students from various disciplines, and the discipline-specific bundles can certainly be a unique contribution to any EAP classes, especially those that want to boost students' awareness of bundles as they contrast how some of the most frequent bundles vary across disciplines.

As lexical bundles "are an important part of the communicative repertoire of speakers and writers" (Biber et al., 2004: 377), novice writers can be trained to recognize and use bundles, making their oral or written texts easier to understand. Activities in which learners deal with academic cross-disciplinary lexical bundles that work as signposts in writing (Reppen, 2018: 195–196) are of great help to students. Such bundles are

crucial for giving the text appropriate discourse frames, such as presenting "how the text or information is organized (e.g., *at the beginning of, at the end of*), expressing relationships about the information being presented (e.g., *as a result of, in addition to, on the basis of*), showing contrast (e.g., *on the other hand*), and highlighting information or processes (e.g., *it is important to*)." Reppen (2018) presented several activities, including a jigsaw task that can be a great discovery moment for students as, individually or in pairs, they put together words or groups of words (e.g., *at the, of the*) to form the bundles. Once their list is done, they can compare the lexical bundles they formed to a list of academic lexical bundles taken from Biber et al. (1999). Another activity Reppen (2018) suggested is to have students individually look for bundles in academic texts (textbooks or any class readings) and then, in pairs or groups, compare the results to determine if they were all able to identify the same bundles. She also advises to let students work with different texts and have them compare what they found out. Students could also compare the texts they have written for class assignments with the analyzed texts to determine if they used the same bundles that are present in published materials. This last step of the activity would go beyond raising awareness, making it possible for learners to edit their texts, thereby improving the use of lexical bundles in their own written texts.

Along the same phraseological trend as Reppen and Olson (2020) and Reppen (2018), Bocorny and Welp (2021) developed a description of key lexical bundles in the introduction section of physics articles, integrating linguistic analysis, genre awareness, and text production in a way that genre moves and linguistic analyses work hand in hand as the basis of task design. The linguistic description, based on corpus linguistics and genre theory (Swales, 1990), led them to detect key lexical bundles with fixed grammatical words and internal variable slots that are filled with content words (e.g., *the * of * * is/was: the purpose/aim of this paper/present study is/was*). Bocorny and Welp (2021) highlighted that key lexical bundles have clear communicative purposes; therefore, they are worth teaching to the target group on focus (i.e., physicists), who wish to improve their writing skills to be able to successfully publish research articles. Considering that the unique teaching and learning context of EAP warrants that carefully

designed principles be followed, they followed Welp et al.'s, (2019) proposal. First, the target group discipline and students' needs should guide the setting of objectives. Second, text genres should match the objectives and be relevant for the EAP group. Third, authentic texts should be used and "represent the social practices and the genres that are produced in the academic context" (p. 6). Fourth, the use of language should be promoted along with awareness of use. Fifth, tasks should be organized to encourage scaffolding and facilitate learning. Sixth, "tasks should induce relevant interaction among students and texts, students and students and students and teachers" (p. 6). Finally, tasks should generate learning that is meaningful and impacts language usage beyond the classroom. The series of tasks in Bocorny and Welp (2021) is a good example of a sequence that aims to make learners activate knowledge to write the genre they need. They do so by, first, accessing their previous genre knowledge or acquiring new knowledge through observation of the text type. Second, they have several opportunities to see how lexical and phraseological resources are used with specific communicative purposes in the chosen genre. The corpus-based analysis informs the meaningful key bundles and is the basis for this guided language analysis. Finally, they write their own texts, giving and receiving feedback and learning from each other. Consequently, the classroom context may foster scaffolding and meaningful learning opportunities.

In addition to the description of general and specialized corpus, corpus linguists have used learner corpora to conduct systematic description of learner language and to "help to develop new pedagogical tools and classroom practices" (Granger, 1998: 17), which has positively affected the EAP area. The International Corpus of Learner English (ICLE) was the first major learner corpus to compile argumentative essays written in English by university students from 25 mother tongues, totaling 5.5 million words in its third version (Granger et al., 2020).[6] The investigations based on ICLE have contributed to English for general academic purposes (EGAP; Hyland, 2016) as they have covered an array of topics—namely, learners' use of adjective intensification (Lorenz, 1998), adverbial connectors (Altenberg &

6 https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Tapper, 1998), exemplification (Paquot, 2008), and core vocabulary from a phraseological perspective (Granger & Larsson, 2021). In the academic contexts in Brazil, where there is pressure to internationalize higher education (Sarmento et al., 2016), EAP programs have more recently boosted the need for a focus on learners' writing ability in EAP courses.[7] This development has led learner corpus research to flourish with an analysis of discrete categories (Dutra et al., 2017, 2019 on linking adverbials; Matte & Sarmento, 2018) and a great number of linguistic features with the objective of understanding variation in ICLE, especially on the Brazilian learners' subcorpus (Berber Sardinha & Shimazumi, 2021; Delegá-Lúcio, 2013) using the multi-dimensional methodology, which will be further discussed later in this chapter.

Some CL studies have concentrated on academic oral language (Liu & Chen, 2020; Neelly & Cortes, 2009) being good support to EAP instructors, who often need to prepare their students to attend and understand academic lectures. Based on Biber et al.'s (2004) and Nesi and Basturkmen's (2006) lexical bundles' lists, Neely and Cortes (2009) investigated the five most frequent lexical bundles used to introduce new topics in lectures, studying their occurrence in the Michigan Corpus of Academic Spoken English (MICASE) as well as their functions in the academic context. Comparing the use of specific bundles—namely, *if you look at, a little bit of, a little bit about, I want you to,* and *I would like you*—by instructors and students, they were able to, contextually, analyze the specific bundle functions. Neely and Cortes (2009: 29) realized that bundles that are broadly categorized as "discourse markers" or "topic introducer" may play different functions during lectures, such as "*if you look at*, [which is] (…) not always used to introduce a topic in a lecture or student presentation, [but which is] (…) often used to ask students to turn their attention to a new object in the classroom or to imagine or contemplate a topic already under discussion." The authors also presented a series of lesson plans in which students are led to analyze MICASE lecture excerpts to identify lexical bundles used

7 EAP courses in Brazil adopted a greater focus on reading skills in the 20th century (Salager-Meyer et al., 2016).

to introduce new topics, compare such uses with lectures included in textbooks, and detect the specific functions of the bundles. These model lesson plans can serve as inspiration to EAP instructors who are compelled to design materials for their classes, which could also be supported by Liu and Chen's (2020) results in their study on lecture lexical bundle variation across disciplines. In this regard, this article, which is based on an 8.8-million lecture corpus in four disciplines (engineering, science and math, humanities and art, and social sciences), is a valuable source of cross-disciplinary variations in information from a central university register and presentations, allowing for the preparation of activities that could boost learners' listening comprehension skills. Liu and Chen (2020) provided a list of the most frequently used lexical bundles across the four areas, comparing the frequency and the role of the bundles as well as their functions as referential, stance, and discourse-organizer bundles. Among the differences, they highlighted that the engineering, science and math, and social sciences lectures carried more stance lexical bundles than the humanities and arts lectures. The three areas often use bundles, such as *is going to* and *is going to be a*, to give explicit step-by-step guidance in which logical steps, effects, and outcomes can be observed and are crucial for the process. On the other hand, humanities and arts lectures appeared to be "less definite and less clearly defined," enabling students to make connections and come to conclusions in a "distinct style of knowledge construction" (Liu & Chen, 2020: 132). They concluded that, "although the frequency of lexical bundles appearing in disciplines vary considerably, the items used across disciplines are similar" (Liu & Chen, 2020: 133), which can be interpreted by EAP instructors as a warning for working with both cross-disciplinary and discipline-specific bundle activities.

We close this section by bringing to the foreground the notion that lexis and grammar are interconnected and, therefore, their associations are worth studying. This notion is fundamental in corpus linguistics as it "allows researchers to identify and analyze complex 'association patterns'" (Biber et al., 1998: 5). These authors argued that patterns should be investigated in terms of their linguistic associations (how words relate to each other and how grammatical structures are associated). In addition, linguistic features

should be studied from a perspective of non-linguistic associations, such as how registers, dialects, and time periods affect language use. Another perspective would be to explore text or text varieties through the linguistic association patterns of linguistic features, including how patterns co-occur. Our next two sections will present corpus linguistics studies that prioritize the associations mentioned: grammatical complexity with a focus on noun phrases and co-occurrence of linguistic features based on MD analyses. In these sections we will show the centrality of lexico-grammatical features in language, their associations with registers, and contributions to EAP.

Grammatical complexity from the lens of CL and contribution to EAP

In this section, we discuss what grammatical complexity is as well as how it has been studied in first and additional languages and highlight suggestions to EAP programs based on corpus-based studies that deal with such complexity. A widespread interest in language teaching, in both first language (L1) and second language (L2), focuses on writing development, its relation to grammatical complexity, and how to measure it. The T-unit concept of grammatical complexity, defined as "a main clause and all associated dependent clauses" (Biber et al., 2011: 7), has permeated most studies in L1 and L2 in the last century and in the first decade of this century. More specifically, two measures have often been used in investigations on grammatical complexity:

> mean length of T-unit (MLTU), which relies on the overall length in words of the T-unit, averaged across all T-units in a text, and clauses per T-unit (C/TU), which relies on the number of dependent clauses per T-unit, again averaged across all T-units in a text. (Biber et al., 2011: 7)

The common interpretation of these measures was that more complex texts would carry longer words and more dependent clauses. Above all, clausal subordination became synonymous with complex and elaborated L2 written texts, influencing many EAP courses to overemphasize the role of connectors in academic writing.

Despite the popularity of the MLTU and C/TU measures in applied linguistics studies in the 20th century, a few scholars noticed that other measures were called for. Bardovi-Harlig (1992) challenged the T-unit measures as they seemed to not reflect how advanced learners of English were writing. She showed how coordination needed to be accounted for as such measures are frequently used in earlier-stage writings and pointed out that embedding should also be considered as a characteristic of advanced learners. She stated that:

> T-unit analysis artificially divides sentences that were intended to be units by the language learner, imposing uniformity of length and complexity on output that is not present in the original language sample. By treating all conjoined sentences as if they were not conjoined, a T-unit analysis discounts the learner's knowledge of coordination. (Bardovi-Harlig, 1992: 391)

One of her examples, reproduced below, shows that, by simply counting the number of clauses, a T-unit analysis would ignore that the sentence reflects a certain rhetorical sophistication that includes coordination:

> Hundreds of schools were built, and tens of institutions are starting to join in providing technical education to the public. (L1 Arabic) (2 T units/1 sentence). (Bardovi-Harlig, 1992: 391)

Ortega's (2003) review paper, published 11 years after Bardovi-Harlig's warning, confirmed that T-units were still a popular measure in L2 writing studies. In order to understand how studies had been looking at L2 writing syntactic complexity in relation to proficiency, Ortega (2003) analyzed 27 studies: 21 cross-sectional and 6 longitudinal studies. The majority of the reported investigations (i.e., 25) relied on MLTUs. Ortega (2003: 514) was cautious to point out that:

> researchers interested in using syntactic complexity measures as global indices of L2 proficiency may refer to these findings as interpretive landmarks for aiding study design and interpretation of study outcomes in future college level L2 writing research.

She thus recommended that studies focus on developmental prediction and cross-rhetorical transfer.

Biber et al.'s (2011) corpus-based study filled the gap Ortega (2003) identified as they revisited the concept of grammatical complexity in light of a register perspective. This study presented an analysis of 28 features in two different registers, conversation and academic research articles, and concluded that clausal complexity was characteristic of conversation while complexity in research articles was attested by phrasal complexity, such as by nonclausal features frequently embedded in noun phrases. In other words, finite clauses often occur in conversation and function as adverbials and verb complements (e.g., "*I think we better wait. […] he gets mad cause he can't smoke cause we always take non-smoking*"; Biber et al., 2011: 24) while prepositional phrases, attributive adjectives, and noun phrases are commonly found in articles (e.g., *We expected that the use of different transformations would have significant effects on our perceptions of spatial patterns in kelp holdfast assemblages*; Biber et al., 2011: 27). This publication marked a major turning point in grammatical complexity studies demystifying the T-unit and subordination characteristics as the best measures of grammatical complexity. The paper culminated in the presentation of hypothesized developmental English stages for complexity features. These stages are based on their analysis of English as an L1 oral and written texts and are hypothesized as following the same sequence of acquisition in English as an L2 language. They argue that "conversation is acquired first; the grammar of writing is acquired later, and not always successfully" (Biber et al., 2011: 28). Not all native speakers produce academic texts, and the phrasal complexity features detected in research articles, if acquired, would be part of the adult repertoire. Taking into account this rationale, the authors proposed that the hypothesized developmental stages for complexity features include five stages, starting from the production of features, such as "finite complement clauses (*that* and *WH*) controlled by extremely common verbs (e.g., *think*, *know*, *say*)," and continuing to quite complex phrasal embedding: "extensive phrasal embedding in the NP: multiple prepositional phrases as postmodifiers, with levels of embedding," as in

"*The* [*presence of layered* [[*structures*] *at the* [[[*borderline*]] *of cell territories*]]]" (Biber et al., 2011: 31).

In the following paragraphs, we first highlight studies on English as an L1 that were inspired by this expanded notion of grammatical complexity. We then explore how the hypothesized developmental stages influenced studies on English as an L2, taking into consideration the implications for EAP.

Biber and his associates (e.g., Biber, 2006; Biber & Gray, 2010; Biber et al., 2011) have investigated the unique qualities of academic language, culminating in a historical analysis of academic English in Biber and Gray (2016) that revealed how a register can change diachronically to reflect new community practices. In the 18th and 19th centuries, academic scientific papers were most frequently organized around clausal features, and academic research articles were quite similar, linguistically, to fiction; thus, phrasal features were often not found in academic texts of those periods. The authors claimed that, in the 20th century, two major societal changes influenced written texts. First, mass literacy became a reality, increasing readership of any written registers. Many different types of texts, such as fiction books and newspaper articles, had to popularize and were influenced by oral registers. Second, science became much more specialized with the emergence of sub-disciplines, which meant that written scientific texts have increasingly targeted very specific audiences. Biber and Gray (2016) argued that this social force influenced scientific writing in two ways: There is a constant rise in information volume, and texts need to "present more information in an efficient and concise way," leading to "greater 'economy' in written informational texts" (p. 129). In the 20th and 21st centuries, scientific writing has adopted a compressed and dense style, with a high use of phrasal features; when this register is compared with conversation, it becomes clear that clausal embedding is much more frequent in the latter register (Biber et al., 2016), revealing clausal complexity in conversation but not in academic writing. These results from corpus-based studies, unlike investigations using T-unit measures, unveiled a use of phrasal features in academic writing that had not been noticed before.

Along the same lines as Biber et al. (2016) and Biber and Gray (2016), other corpus-based disciplinary and register variation investigations on English as an L1 as well as an L2 have been carried out, uncovering more characteristics of academic discourse that were not known and that can take EAP closer to students' needs. Gray (2013) studied the extent to which discipline as well as the nature of the research (quantitative, qualitative or theoretical) would affect linguistic variation in research articles. The disciplines investigated were physics, biology, applied linguistics, philosophy, history, and political sciences. Some results showed that qualitative history, political science, and applied linguistics text analyses revealed the co-occurrence of similar features (e.g., nouns, time and topic adjectives, tense and aspect markers, communication verbs) whose "focus is on reconstructing an event to serve as the foundation for interpretations and subsequent claims" (Gray, 2013: 168) and characterize contextualized narrative. Quantitative political science and applied linguistics articles showed many fewer narrative features as they also incorporated features that make the text more concise and informative to construct descriptions. Quantitative biology and physics as well as theoretical physics are aligned in their use of several features that convey procedural description, carry heavy information load (e.g., nouns, attributive adjective), and compose the frequent phrasal features. Gray's conclusion was that multiple parameters should be considered to augment the understanding of linguistic variation in research articles. EAP teachers should be aware of discipline variation as well as the nature of the research—be it quantitative, qualitative, or theoretical—as it does influence linguistic variation across and within disciplines.

Considering that complex phrasal structures play a major role in the construction of economic and dense academic scholarly writing, there has been a growing interest in better understanding noun pre-modification (Ang et al., 2017; Dutra et al., 2020; Hutter, 2015). Results from discipline-specific complex noun phrase investigations should provide EAP teachers with information that has received little coverage in popular English textbooks, which "extensively cover finite dependent clausal structures (e.g., relative clauses, conditionals, and complement clauses for reporting speech)" (Biber et al., 2016: 16). Through a detailed description

of complex noun phrases composed of adjectives and/nouns in chemistry and applied linguistics research articles—two distinct disciplines—similarities and differences were uncovered in Dutra et al. (2020). First, high lexical variation in the noun phrases was found, and only 1.7% of adjective pre-modified noun phrases were lexically the same in both corpora. Not surprisingly, these commonly shared noun phrases are not discipline specific. Nonetheless, they play crucial referential roles addressing parts of the article (e.g., *the statistically significant results*) or referring to present or previous studies (e.g., *more recent study*), which make them strong candidates for being easily taught in general EAP classes. Second, they discovered that both disciplines pack a great deal of information as their communities produce noun phrases ranging from two words (e.g., *prosodic nature*) to seven words (e.g., *four identical in-class individual web-based writing tasks*). This result confirms the need to explicitly teach complex noun phrases to EAP learners in these two disciplines. Third, by carefully analyzing the relationship between the elements of long noun phrases, they were able to attest that noun phrase complexity is the result of not only packing premodifiers, but also interrelationships between the elements of the phrase (Dutra et al., 2020). Such a complexity trait was acknowledged by Biber et al. (1999: 600):

> …sequence of words in the premodification can represent a large number of different structural/logical relations, with forms often modifying other premodifiers instead of the head noun. As a result, there is much structural indeterminacy, leading to the possibility of incorrect interpretations.

A good example of how noun phrase complexity can add difficulties to comprehension comes from their chemistry corpus's eight-word noun phrases, most of whose modifiers do not modify the head noun: *low temperature 3He strongly adsorbed gas diffusion experiments* (Figure 2). The head noun (*experiment*) is modified by *gas diffusion* and by *low temperature*, but not by *adsorbed* or *strongly*. The adverb *strongly* modifies *adsorbed*, and this adjective modifies *gas*. Such a noun phrase may not be a barrier in understanding for an expert in the area, but novice writers would

certainly benefit from teaching interventions focused on such a linguistic phenomenon.

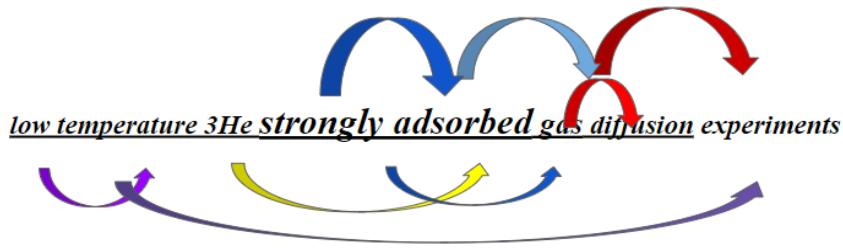


Figure 2. Sample of interrelations of modifiers from a chemistry corpus

Dutra et al. (2020) also noticed that a great deal of applied linguistic complex noun phrases behave quite differently from the chemistry noun phrases since all modifiers refer to the head noun (Figure 3): *writing* modifies *tasks*, the head noun, in the same way that *web-based*, *individual*, *in-class* and *identical* modify *tasks*.

Figure 3. Sample of interrelations of modifiers from an applied linguistics corpus

Presenting the information shown in Figures 2 and 3 in EAP classes should raise learners' awareness of the extent of phrasal complexity in different disciplines. It should also help improve the writing of dense academic texts in higher education in countries such as Brazil where the first language differs from English, in some ways, in how it constructs noun phrases. In other words, long noun phrase structure may pose challenges for many students, especially for the ones whose first language does not

frequently use heavily pre-modified noun phrases, such as for Portuguese speakers (Dutra et al., 2020). Noun phrases are structured in Portuguese, most learners' first language in the country:

> Portuguese allows the use of attributive adjectives but not the use of nouns as pre-nominal modifiers. Consequently, understanding and producing heavily pre-modified [noun phrases] can be an arduous task in a second language, especially in research writing. (Dutra et al., 2020: 209)

It seems undeniable that grammatical complexity should be addressed in EAP in academic writing classrooms, and learner corpus studies can further support the planning and implementation of such interventions so that they are adequate for students' needs. It is not the case that EAP learners do not use complex noun phrases even when they are B2[8], with an intermediate level of proficiency, but the question is which complex noun phrases are used when they produce which type of essay (Queiroz, 2019). Queiroz's study revealed that Brazilian writers use more complex than simple noun phrases, especially those with premodifying adjectives as well as with postmodifying prepositional phrases. The EAP corpus that Queiroz investigated, *CorIFA*[9], included a subcorpus formed from general topic and specific topic essays. Queiroz found that the mean score of complex noun phrases in the specific topic subcorpus was clearly higher than in the general topic essay subcorpus. These complex noun phrases are discipline specific, leading the author to posit that the task type, specific topic essays, promoted the use of more complex noun phrases. This result is relevant for general EAP courses as they should find room for discipline-specific language activities and, above all, should stimulate writings about students' learning and research area.

8 Common European Framework of Reference (CEFR) corresponds to the level of proficiency ranging from A1, beginners, to C2, proficient users of the language.

9 *CorIFA* stands for *Corpus de Inglês para Fins Acadêmicos* (see Dutra et al., 2022 for information on *CorIFA*).

Other learner corpus studies have focused on investigating whether the hypothesized stages proposed in Biber et al. (2011) correspond to real learners' development in their writing skills. Parkinson and Musgrave's (2014) corpus-based study revealed that EAP learners' essays, when compared to the essays of master's degree students in applied linguistics, present significantly more adjectives as premodifiers and fewer prepositional phrases. The more proficient students (i.e., master's degree students) use more nouns as premodifiers and more prepositional phrases as postmodifiers. In other words, more proficient students use more complex noun phrases, as hypothesized.

More recent learner corpus studies have looked at longitudinal data to track learners' development to see if they confirm cross-sectional studies' results (Ansarifar et al., 2018; Parkinson & Musgrave, 2014). Biber et al. (2020) explored a multiple L1 learner corpus compiled from students' disciplinary texts written in English, and Alves (2022) assessed a longitudinal corpus of Brazilian EAP learners who have produced a range of different register assignments (statements of purpose, abstracts, essays, literature reviews, and research articles) in various disciplines. Both studies revealed a decrease of dependent clause complexity features while phrasal complexity feature usage went up as students' proficiency increased, as hypothesized in Biber et al. (2011). However, Alves (2022) found no steady increase of all expected phrasal features along the three moments of corpus compilation, which may be due to the short interval between the terms when students wrote the text (i.e., about 4 months). The author added that a qualitative analysis pointed to an increase in lexical variation, "specifically in the scope of attributive adjectives, linking adverbials, nouns as premodifiers, adjectives in extraposed constructions, and as [preposition phrases'] postmodifiers" (Alves, 2022: 117), and most of them contributed to improvement in textual phrasal complexity. Alves also compared EAP learners' texts across academic divisions (social sciences and education, humanities and arts, physical sciences and engineering, and biological and health sciences), detecting a high use of attributive adjectives in all academic areas as noun modifiers, but a preference for nouns as postmodifiers in social sciences and education texts. These academic divisions include many disciplines,

which means that EAP teachers should consider these results with caution and compare them to discipline-specialized corpora. If they compile or have their students compile small discipline-specialized corpora, according to their students' disciplines, they could lead learners to explore texts written by experts and compare them to their own use of complex noun phrases.

MD Analysis and EAP

Multi-dimensional analysis is a framework used to identify sets of correlated linguistic features shared across many different texts in a corpus. These correlated sets, which are statistically identified through factor analysis, are communicatively interpreted as dimensions, the underlying parameters of variation in language use. In the 1980s, Douglas Biber (1988) developed the multi-dimensional analysis as a tool for analyzing variations in spoken and written language, with the assumption that multiple dimensions shape the texts simultaneously. Such an assumption was in sharp contrast to the literature at the time, which tended to describe registers using a single parameter (e.g., formality, involvement). Multi-dimensional analysis was revolutionary not only because of its emphasis on a multi-faceted approach to text analysis, but also because it was designed as a corpus-based framework at a time when corpus linguistics was in its early stages and the focus of most corpus linguistic studies was the corpus rather than the actual texts in the corpus.

It is beyond the scope of the current chapter to provide a detailed description of the procedures involved in conducting a multi-dimensional analysis (see Almela, Cantos Gómez & Berber Sardinha, 2022; Berber Sardinha, 2000; Berber Sardinha & Veirano Pinto, 2014, 2019; Biber, 1988; Conrad & Biber, 2001; Egbert & Staples, 2019; Friginal & Hardy, 2014; Zuppardi, Veirano Pinto & Berber Sardinha, in prep.). Briefly, however, the basic steps involve: (1) Collecting a corpus that represents a particular register or domain; (2) Tagging the corpus for part-of-speech [10] or for other

10 "Factor analysis identifies sets of features that co-vary …" (Biber 1988: 65)

linguistic characteristics automatically; (3) Counting the linguistic features annotated and norming the counts (e.g. to a rate per thousand words); (4) Entering the counts in a factor analysis, and determining the latent factors in the data; (5) Scoring each text by summing up the counts of the features loading on each factor; (6) Interpreting the factors communicatively by reading samples of texts and assigning a label to each factor that reflects the major communicative properties of the dimension. It is important to note that it is common for dimensions to comprise two 'poles', that is, two different sets of features in complementary distribution in the texts, such that when the features in one pole occur in the text, the features in the other pole are generally absent, and vice-versa. Although these poles are referred to as 'positive' and 'negative', these labels are not evaluative and simply reflect the fact that two complementary sets of features exist in a single dimension. In summary, then, each dimension comprises a set of linguistic features cooccurring in the texts, determined through statistical analysis and interpreted qualitatively by the analyst to reflect its underlying communicative purpose.

The multi-dimensional analysis literature on EAP is vast, encompassing studies conducted on the basis of grammatical structures, lexical units (collocations, lexical bundles), and discourse. Because of its emphasis on cross-text analysis and statistical rigor, multi-dimensional analysis provides rich descriptions that can be of interest to EAP teachers, as these descriptions provide a detailed view of the most used sets of linguistic features in academic registers. It is important to stress that dimensions are sets of correlated linguistic features that frequently occur together in texts because they perform a particular communicative function. As such, multi-dimensional analysis descriptions show how seemingly different features work together to achieve a particular rhetorical purpose and, in this way, can serve as entry points into academic language, thereby enabling EAP curriculum developers to design instructional materials centered around macro functions rather than around individual linguistic features.

In this section, we review multi-dimensional analysis studies that provide an overview of academic language by looking at articles, article sections, reports, textbooks, and campus registers. The first of these studies

was conducted by Gray (2013), who analyzed variation in research articles by academic discipline, using a corpus of 270 research articles comprising three sub-registers (theoretical, qualitative, and quantitative research reports) from six disciplines (philosophy, history, applied linguistics, political science, biology, and physics). The first dimension, labeled "academic involvement and elaboration versus information density," distinguishes between research articles that interact with the reader and present frequent evaluation, argumentation, and interpretation with overt textual signals (the positive pole) and texts that exhibit high-density informational language (the negative pole). The positive pole is marked by such linguistic features as first-person pronouns, predicative adjectives, modals (prediction, possibility, necessity), subordinating conjunctions, adverbial conjuncts, and a range of *that*-complement clauses and *to*-clauses. In contrast, the negative pole comprises nouns, prepositions, passive voice, past tense, a high type–token ratio, and long words. The distribution of the disciplines shows a contrast basically between one single discipline (philosophy), with very high scores on the positive pole, and all the other disciplines, which have either negative scores or scores close to zero on the positive pole. Thus, the involved and elaborated style is very discipline specific whereas the high-information style is more commonly embraced by different disciplines. Yet ample variation exists within each discipline; although most disciplines prefer an information focus rather than an involved, elaborated style, they also allow for both styles. The exception is philosophy, which includes the involved, elaborated style only), and quantitative biology (which includes the high-information style only). The two theoretical disciplines of philosophy and theoretical physics both have texts with positive scores (although theoretical physics includes texts with negative scores, unlike philosophy), suggesting that the involved, elaborated style is generally preferred by theoretical papers.

The second dimension distinguishes between contextualized narration (positive pole) and procedural description (negative pole). Contextualized narration is marked by features such as past tense verbs, third-person pronouns, coordinating conjunctions, *that*- and *to*-complement clauses, long words, a high type–token ratio, and long texts.

Meanwhile, procedural description is marked by nouns, attributive adjectives, and passive voice. The way the disciplines are distributed along the dimensions shows two clusters: one comprising qualitatively oriented disciplines (history, political science, and applied linguistics), with high scores on the positive pole, and the other comprising theoretical and quantitative disciplines, with low positive scores or negative scores. This finding suggests that contextualized narration is a style largely preferred for qualitative reports, whereas procedural description is a common style used in non-qualitative articles.

The third dimension is based on a distinction between a human (positive pole) and non-human focus (negative pole). The positive pole includes such linguistic characteristics as second- and third-person pronouns; mental, cognition, and communication verbs; and *that*- and *to*-complement clauses. The negative pole, in contrast, comprises adjectives (in attributive position), adverbs, and prepositions. Disciplines having a human focus are essentially applied linguistics (qualitative, but to a lesser degree, quantitative) and philosophy whereas all the other disciplines share a non-human focus.

Finally, the fourth dimension identifies academese as a major trait in academic writing, which corresponds to "a concern to overtly represent research as empirical, well-motivated and founded in previous research" (Gray, 2013: 174). Academese is associated with the prevalent use of nominalizations, process nouns, abstract nouns, attributive adjectives, existence verbs, *that*- and *to*-complement clauses, and long words. This is most commonly found in articles from applied linguistics and political science.

Although a research article is generally seen as a single unit in which the internal variation is minimal or of limited relevance, research articles are in fact comprised of several sections, each performing a particular function in the text. For instance, according to Swales (1990), introductions are supposed to establish a territory and a niche (problem) and occupy the niche (present a solution), among other rhetorical moves. In contrast, methods are supposed to lay out the procedures followed by the study and present the data, tools, and other methodological decisions taken by the authors when conducting the study. Given the different rhetorical purposes of the

different research article sections, it is legitimate to expect that variation exists within research articles that reflects the different purposes of the various sections. The variation across the language used in different sections should be of interest to EAP practitioners, especially those concerned with writing instruction, as a detailed description of the most typical language used in different sections could help them better understand and select the teaching points necessary to prepare their students to write efficient article sections.

Dutra and Berber Sardinha (2018, 2021) looked at variation across sections in a corpus of applied linguistics, biology, and chemistry research articles. Each article was segmented into individual sections—namely, abstract, introduction, method, results, discussion, and conclusion. The corpus comprises 900 sections for each discipline, totaling 2.9 million words.

The first dimension, labeled interpretive elaboration, includes third-person pronouns, communication and mental verbs, *that*- and *to*-complement clauses, *wh*-words, infinitives, and nominalizations. This dimension corresponds to a distinction between applied linguistics and the other two disciplines, as all sections from applied linguistics, especially conclusions and discussions, exhibit positive scores on this dimension.

The second dimension, which corresponds to logical argumentation, comprises characteristics such as present tense verbs, adverbs, adjectives in predicative position, adverbial conjuncts, *that*- and *to*-complement clauses, demonstrative pronouns, and prediction modals. The conclusion and discussion sections, mainly from applied linguistics, biology, and chemistry, have higher scores on this dimension.

The third dimension reveals a distinction between informational density (on the positive pole) and procedural narrative and description (on the negative pole). Informational density corresponds to the dense use of long words and adjectives in attributive position whereas procedural narrative and description relies on past tense verbs, agentless passives, long sections, and activity verbs. The variation across sections shows that informational density is more typical of abstracts, conclusions, and introductions whereas procedural narrative and description is more typical of methods and results. Based on the results, the discipline is not a good predictor of

the variation. Rather, the variation is patterned along a combination of discipline and section, with no clear-cut distinctions. For instance, biology conclusions score high on informational density whereas biology methods score high on narrative and description.

In general, all dimensions predict a higher share of the variation when considering discipline and section together rather than when a section alone or discipline alone is considered. This suggests that, because sections can be very discipline specific, care should be taken in EAP to not generalize across disciplines when trying to characterize the language of research article sections. Rather, EAP practitioners should be aware of the section specificities of different disciplines when teaching their students to write academic articles.

Whereas the previous studies reviewed thus far focused on journal articles, the next study looked at student writing in an American university. Hardy and Römer (2013) analyzed the Michigan Corpus of Upper-level Student Papers (MICUSP), which includes samples of written assignments from 16 disciplines, totaling more than 2.6 million words. The samples represent a range of registers, such as argumentative essays, proposals, reports, and research papers, among others.

The first dimension comprises two poles: involved, academic narrative (positive pole) and descriptive, informational discourse (negative pole). The linguistic features that loaded on the positive pole of the first dimension include verbs of different types (mental verbs, private verbs, activity verbs), past tense verbs, *that*-deletion, and first- and third-person pronouns. On the other hand, features loading on the negative pole convey dense quantities of information, such as nominal features like nouns, nominalizations, and adjectives. The disciplines are sharply distinguished on this dimension, with the humanities, arts, and social sciences scoring on the positive pole (particularly philosophy and education) and biological, health, and physical sciences scoring on the negative pole (most markedly physics and biology). The exception is linguistics, which scored in the negative pole.

The second dimension, labeled expression of opinions and mental processes, primarily comprises a large number of stance (both *to*- and

that-stance clauses, controlled by adjectives and verbs) and *that*-complement clauses (controlled by factive, non-factive, verb of likelihood, adjective of likelihood). The disciplines are distributed along this dimension in a similar manner as in the first dimension, with the humanities and social sciences having higher scores on the positive pole (philosophy and education being the top two), thereby being more readily associated with the expression of opinions and mental processes, whereas in the remaining disciplines the expression of opinions and mental processes is much less common (civil engineering and physics as the most marked).

The third dimension corresponds to a distinction between situation-dependent, non-procedural evaluation (positive pole) and procedural discourse (negative pole). The features loading on the positive pole include a range of adverbs (including stance), verbs, pronouns, and *that*-complement clauses controlled by verbs of likelihood. In contrast, the negative pole is based on nouns and passives. The register distribution along the dimension is similar to the previous dimensions, with a split between the humanities on one pole and the remaining sciences on the other. The humanities (e.g., philosophy, English) score highly on the situation-dependent, non-procedural evaluation end of the dimension whereas the natural and exact sciences (physics, mechanical engineering) score highly on the procedural discourse end.

The final dimension, labeled production of possibility, is based on the use of modals (possibility, prediction), stance (*that*-complement clauses controlled by adjectives, *to*-complement clauses controlled by adjectives), infinitives, and verbs in general. Unlike the previous dimensions, the disciplines are not evenly split between the humanities and the remaining sciences. The disciplines most marked by this dimension include human sciences (e.g., philosophy, linguistics), life sciences (nursing, psychology), and education; the least marked include the humanities (history and classical studies), natural sciences (physics), and engineering (mechanical engineering).

As the results of this study indicate, the language used in discipline-specific writing differs sharply, mainly between the humanities and the remaining disciplines. In the humanities, authors prefer language that

is more involved, narrative, opinionated, and situation dependent; in all the remaining disciplines, authors tend to use language that is more informational, less opinionated, and procedural. Yet this divide between the humanities and non-humanities does not apply to the expression of stating possibilities and arguments, where the distinction is much more blurred as each specific discipline has a different attachment to this type of discourse.

Multi-dimensional analysis has been applied to the description of academic English mostly from a grammatical perspective, as the studies discussed thus far have demonstrated. However, multi-dimensional analysis can provide detailed descriptions of academic language from a lexical perspective as well, thereby shedding light on how academic language is patterned for such aspects as collocations (Zuppardi, 2020; Zuppardi & Berber Sardinha, 2020) and discourse (Berber Sardinha, 2021). We next review Zuppardi and Berber Sardinha's (2020) study, which provides a unique view on how collocations cluster in academic writing that can help EAP educators as they prepare their students to handle the large number of collocations needed to master academic English.

Zuppardi and Berber Sardinha (2020) used a novel form of multi-dimensional analysis based on collocations (Berber Sardinha, 2017; Zuppardi, 2020) to analyze a large corpus of academic writing comprising articles and textbooks from seven disciplines: behavioral and cognitive sciences, social and economic sciences, anthropology, political science, psychology, and economics.

The first dimension corresponds to a distinction between collocations referring to human nature, culture, and research methods and collocations related to economics. Collocations in the first group encompass a large number of nominal, adjectival, and verbal collocations formed around nodes such as *literature* (e.g., *literature review*), *culture* (*common culture*), *behavior* (*human behavior*), *human* (*human tendency*), *developmental* (*developmental basis*), *genetic* (*genetic variation*), *highlight* (*highlight the importance*), *review* (*review the evidence*), and *live* (*live alone*). In contrast, the economics collocations include collocations around nodes such as *saving* (*national saving*), *currency* (*foreign currency*), *corporation* (*large corporation*), *fiscal* (*fiscal policy*), *extra* (*extra revenue*), *nominal* (*nominal*

rate), *finance* (*finance and investment*), *purchase* (*purchase bond*), and *borrow* (*borrowing constraints*).The second dimension, which refers to human evolution and society, includes collocations around noun nodes such as *species* (*separate species*), *ape* (*ape behavior*), and *anthropologist* (*cultural anthropologist*); adjective nodes like *ancient* (*ancient remains*), *African* (*African populations*), and *evolutionary* (*evolutionary change*); and verb nodes such as *date* (*date fossils*), *remember* (*remember a discussion*), and *gather* (*gather data*).

The third dimension, interpreted as business and finance, encompasses collocations around nouns like *dollar* (*dollar cost*), *bank* (*bank account*), and *interest* (*interest payments*); adjectives like *net* (*net worth*), *annual* (*annual income*), and *marginal* (*marginal cost)*; and verbs like *sell* (*sell products*), *pay* (*pay dividend*), and *raise* (*raise funds*).

The final dimension, referring to statistical vocabulary, includes collocations with the following nodes: nouns like *error* (*error variance*), *correlation* (*correlation coefficient*), and *population* (*population parameter*); adjectives such as *linear* (*linear model*), *estimated* (*estimated effect*), and *explanatory* (*explanatory variable*); and verbs like *compute* (*compute average*) and *estimate* (*estimate model*).

The dimensions provide a network-like outlook on collocations, unlike the literature in general, which tends to see collocations individually or in small sets. The study demonstrated that collocations are shared systematically across texts. Therefore, a skilled academic writer requires being able to select the most appropriate collocations for the particular topics addressed in the article or textbook. Similarly, the fact that words tend to appear in predictable combinations has consequences for readers as well, as a proficient reader is able to anticipate these collocations in the text. Overall, this study shows that, for the most part, the bulk of the collocations in academic writing is not a set of specialized technical expressions; rather, most collocations can be frequently found in non-academic domains.

Biber (2006) presented a multi-dimensional analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL), which consists of spoken and written registers with which students in American universities need to engage as part of campus life. The first dimension

includes two poles: one corresponding to orality and the other to literacy. The pole corresponding to orality is comprised of linguistic features usually associated with informal spoken language, such as contractions, first-/second-/third-person pronouns, stranded prepositions, *that*-omission, discourse particles, and demonstrative and indefinite pronouns. In addition, this pole includes linguistic features that reflect a non-technical use of language, such as common and relatively common adverbs, verbs in the present tense, lexical bundles initiated by pronouns, verbs, and *wh*-pronouns, all of which reflect the interactive tendency of the dimension. The highest scoring academic registers in this pole include office hours, study groups, classroom management, and classroom teaching. In these registers, the face-to-face interactions between teachers and students are enabled by these linguistic features, which in turn allow for the desired level of informality and interaction in North American university settings.

In the negative pole, the predominant linguistic features are related to the use of specialized nouns, such as abstract nouns, human nouns, and group nouns, as well as *to*-clauses controlled by stance nouns or adjectives. The lexical bundles also reflect this nominal orientation of the dimension, including lexical bundles initiated by prepositions. This dimension pole also includes passive structures, formed with *by*-passive and *by*-less-passive voice structures, and adjectives in an attributive position. All these features—in addition to others not mentioned here—generally refer to nominal structures common in specialized literate language. The academic registers that scored highest on this pole are textbooks and course packs, which make consistent use of the features present in this dimension pole.

Like the first dimension, the second dimension also includes two poles: one corresponding to procedural discourse and the other to content-focused discourse. Procedural discourse is marked mainly by modals (present and future), common verbs of activity and causative verbs, *to*-clauses controlled by verbs, and conditional adverbial clauses. Content-focused discourse, on the other hand, is principally marked by specialized vocabulary, such as rare nouns, rare adjectives, rare verbs, and specialized adjectives. This dimension basically distinguishes between spoken and written registers, with few exceptions. The pole corresponding to

procedural discourse includes spoken registers such as classroom management, office hours, and classroom teaching whereas the pole corresponding to content-based discourse comprises registers such as textbooks and course packs.

The third dimension refers to a reconstructed account of events, distinguishing between language used to report past events (in the positive pole) and to convey concrete information (negative pole). The positive pole is essentially composed of non-specialized vocabulary (common nouns: human and mental, common verbs of communication, and common mental verbs), plus a range of *that*-clauses controlled by communication verbs, likelihood verbs, and stance nouns as well as *that*-omission and past tense verbs. This dimension distinguishes between written and spoken registers, with spoken registers (such as study groups, office hours, lab) occurring mainly in the positive pole and written registers occurring mainly in the negative pole.

The last dimension refers to teacher-centered stance, which relies on adverbial linguistic features such as attitudinal, different adverbial features (certainty and likelihood), conditional adverbial clauses, and *that*-clauses controlled by stance nouns. Unlike the other dimensions, it does not neatly distinguish between written and spoken registers. In the positive pole, the most prominent academic registers are classroom teaching and office hours; in the negative pole, they are study groups and institutional writing.

Conclusion

In this chapter, we presented corpus-based studies and their contributions to EAP. First, we discussed the advances in vocabulary studies as the area moved from lists of individual words to phraseological patterns analysis. Second, grammatical complexity research was considered, showing how CL can point out novel ways of observing linguistic phenomena. Finally, we presented multi-dimensional analysis studies and the insights they have provided into the understanding of lexical-grammatical patterns in academic registers. EAP education can include learning about the registers that students are likely to find in universities, beyond the usual registers

from academia, such as academic articles and dissertations. Corpus linguistics has been an integral part of EAP education, and the continued application of corpus-based language analysis promises to further enrich EAP programs.

References

Ackermann, K. & Chen, Y-H. (2013). Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*, 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Almela, A., Cantos Gómez, P. & Berber Sardinha, T. (2022). Métodos multidimensionales basados en corpus del español. In G. Parodi, P. Cantos Gómez, & L. Howe (Eds.), *The Routledge Handbook of Spanish Corpus Linguistics* (pp. 545-557). Routledge.

Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In: Granger, S. (Ed.). *Learner English on computer.* London: Pearson Education, pp. 80-93.

Almeida, V., Orfanó, B. & Dutra D. (2022). Is there a better choice? Verb-noun combinations in academic writing. In: V. Viana (Ed.). *Teaching English with Corpora: A Resource Book* (pp. 228-231).Abingdon: Routledge. http://dx.doi.org/10.4324/ b22833-47

Alves, J. C. (2022). Grammatical complexity in a learner corpus: assessing students' development through a longitudinal study. Master's Thesis, Universidade Federal de Minas Gerais, Brazil.

Ang, L. H., Tan, K. H. & He, M. (2017). A Corpus-based Collocational Analysis of Noun Premodification Types in Academic Writing. *The Southeast Asian Journal of English Language Studies*, 23(1), 115–131. DOI: 10.17576/3L-2017-2301-09

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

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*, 390–395. doi:10.2307/3587016.

Berber Sardinha, T. (2000). Análise Multidimensional. *DELTA*, *16*(1), 99-127.

Berber Sardinha, T. (2017). Lexical priming and register variation. In M. Pace-Sigge & K. Patterson (Eds.), *Lexical Priming: Applications and Advances*. Amsterdam: John Benjamins. (pp. 190-230). https://doi.org/10.1075/scl.79.08ber

Berber Sardinha, T. (2021). Discourse of academia from a multi-dimensional perspective. In E. Friginal & J. Hardy (Eds.), *The Routledge Handbook of Corpus Approaches to Discourse Analysis* (pp. 298-318). Abingdon: Routledge.

Berber Sardinha, T. & Veirano Pinto, M. (Eds.). (2014). *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. John Benjamins.

Berber Sardinha, T. & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues*. New York: Bloomsbury.

Berber Sardinha, T. & Shimazumi, M. (2021). *Variation in learner writing in English: A multi-dimensional analysis of the new ICLE v.3*. [Paper presentation]. XV Encontro de Linguística de Corpus (ELC). Online.

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

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia, PA: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: multiword patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275-311.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

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

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at.: lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D. & Conrad, S. (2009). *Register, genre and style*. Cambridge. Cambridge.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1), 2–20. doi: 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. (2016). *Grammatical complexity in academic English*: *Linguistic change in writing*. Cambridge: Cambridge University Press.

Biber, D. & Gray, B., Staples, S. (2016). Contrasting the Grammatical Complexities of Conversation and Academic Writing: Implications for EAP Writing Development and Teaching. *Language in Focus Journal*, *2*(1), 1-18. DOI: 10.1515/lifijsal-2016-0001

Biber, D., Reppen, R., Staples, S. & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research*, 6(1), 38-71, 2020. *https://doi.org/10.1075/ijlcr.18007.bib*

Bocorny, A. E. P. & Welp, A. (2021). Desenho de tarefas pedagógicas para o ensino de Inglês para Fins Acadêmicos: conquistas e desafios da Linguística de Corpus. *Revista Estudos da Linguagem, 29*(2), 1529-1638. DOI: 10.17851/2237-2083.29.2.1529-1638

Campion, M. & Elley, W. (1971). *An Academic Word List.* Wellington New Zealand Council for Educational Research.

Carter, R. & McCarthy, M. (2006). *Cambridge grammar of English A comprehensive guide to spoken and written English usage*. Cambridge: Cambridge University Press.

Conrad, S. & Biber, D. (Eds.). (2001). *Variation in English: Multi-Dimensional Studies*. London: Longman.

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. https://doi.org/10.1016/j.jeap.2012.11.002.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213–238.

Crosthwaite, P., Luciana & Wijaya, D. (2021). Exploring language teachers' lesson planning for corpus-based language teaching: a focus on developing TPACK for corpora and DDL, *Computer Assisted Language Learning.* (pp. 1-29). https://doi.org/10.1080/09588221.2021.1995001

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Available online at https://www.english-corpora.org/coca/.

Delegá-Lucio, D. (2013). *A variação entre textos argumentativos e o material didático de inglês: Aplicações da análise multidimensional e do Corpus Internacional de Aprendizes de Inglês (ICLE)*. Doctoral dissertation. Pontifícia Universidade Católica de São Paulo, Brazil.

Dutra, D. P., Orfanò, B. M., Guedes, A. S, Alves, J. C. & Fekete, J. G. (2022). The learner corpus path: a worthwhile methodological challenge. *DELTA*, *38*(2), 1-24. https://doi.org/10.1590/1678-460X202238249731

Dutra, D. & Berber Sardinha, T. (2021). *A multi-dimensional typology of English research article sections*. American Association for Applied Linguistics Conference (AAAL). Online.

Dutra, D. P.; Queiroz, J. M.; Macedo, L. D.; Costa, D.& Mattos, E. (2020). Adjective as nominal premodifiers in Chemistry and Applied Linguistics Corpora. In: Römer, U.; Cortes, V. & Friginal, E. (Eds.). *Advances in Corpus-based Research on Academic Writing Effects of discipline, register, and writer expertise*. Amsterdam: John Benjamins Publishing Company. (pp. 205-226) Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.95.09dut.

Dutra, D. P., Orfanó, B. M. & Almeida, V. C. (2019). Result linking adverbials in learner corpora. *Domínios de Lingu@gem*, *13*(1), 400-431. https://doi.org/10.14393/DL37-v13n1a2019-17

Dutra, D. P. & Berber Sardinha, T. (2018). *A linguistic typology of sections in research articles: a Multi-Dimensional perspective*. [Paper presentation] AZCL Conference, Northern Arizona University, Flagstaff, AZ., USA.

Dutra, D. P.; Queiroz, J. & Alves, J. C. (2017). Adding information in argumentative tests: a learners corpus-based study of additive linking adverbials. *Estudos Anglo Americanos*, *46*(1), 9-32.

Egbert, J. & Staples, S. (2019). Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 125-144). New York: Bloomsbury.

Ellis, N. (2008). Phraseology: the periphery and the heart of language. In F. Meunier, F. & S. Granger (Eds.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: Benjamins, 1-13.

Friginal, E. & Hardy, J. A. (2014). Conducting Multi-Dimensional analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber* (pp. 298-316). Amsterdam & Philadelphia: John Benjamins.

Firth, J. R. (1957). *Papers in linguistics: 1934–1951*. London, England: Oxford University Press.

Gardner, D. & Davies, D. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.https://doi.org/10.1093/applin/amt015

Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: a new approach, *English Teaching Forum 17*, 24–7.

Granger, S., Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of thing *Journal of English for Academic Purposes*, *52* https://doi.org/10.1016/j.jeap.2021.100999.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020). The International Corpus of Learner English, Version 3. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), Learner English on Computer (pp. 3–18). Harlow: Longman.

Gray, B. (2013). More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora, 8*, 153-181.

Hardy, J. & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 8*, 183-207.

Hutter, Jo-Anne. (2015). *A Corpus Based Analysis of Noun Modification in Empirical Research Articles in Applied Linguistics*. Master's Thesis, Portland State University.

Hyland, K. (2008). "As can be seen: Lexical bundles and disciplinary variation", *English for Specific Purposes* 27, 4–21. doi:10.1016/j.esp.2007.06.001

Hyland, K. (2016). General and specific EAP. K. Hyland, K.; & P. Shaw, (Eds.). *The Routledge Handbook of English for academic purposes.* New York: Routledge. (pp. 17-29).

Hyland, K. & Jiang, F. (2021). Delivering relevance: The emergence of ESP as a discipline. *Journal of English for Academic Purposes*, 64, 13-25 https://doi.org/10.1016/j.esp.2021.06.002

Johns, T. (1991). Should you be persuaded - two samples of data-driven learning materials. T. Johns, P. King, P. (eds) *Classroom Concordancing. ELR Journal, 4*, 1-16.

Lake, W. M. & Cortes, V. (2020). Lexical bundles as reflections of disciplinary norms in Spanish and English literary criticism, history, and psychology research. In Romer, U., Cortes, V. & Friginal, E. *Advances in Corpus-based Research on Academic Writing Effects of discipline, register, and writer expertise* (pp 95-183). Amsterdam: John Benjamins Publishing Company.

Liu, C.-Y. & Chen, H.-J. H. (2020). Analyzing the functions of lexical bundles in undergraduate academic lectures for pedagogical use. *English for Specific Purposes, 58*, 122-137 https://doi.org/10.1016/j.esp.2019.12.003

Lorentz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on Computer* (pp. 53–66). Harlow: Longman.

Lynn, R. W. (1973). Preparing word lists: a suggested method. *RELC Journal 4*, 25–32.

Matte, M. L. & Sarmento, S. (2018). A corpus-based study of connectors in student academic writing. *English for Specific Purposes World*, *20*(55), 1-21.

McCarthy, M., McCarten, J., & Sandiford, H. (2014). Touchstone 1. Cambridge: Cambridge University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Neely, E. & Cortes, V. (2009). *A little bit about*: analyzing and teaching lexical bundles in academic lectures. *Language Value*, *1*(1) 17–38.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistic*s, *24*(2), 223–242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Nesi, H. (2016). Corpus studies in EAP. K. Hyland, K.; & P. Shaw, (Eds.). *The Routledge Handbook of English for academic purposes* (pp. 2016-217). New York: Routledge.

Nesi, H. & Basturkmen, H. (2006). "Lexical bundles and discourse signalling in academic lectures". *International Journal of Corpus Linguistics, 11*(3), 283- 304.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*, 492–518. doi:10.1093/applin/24.4.492.

Paquot, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier, F. & S. Granger (Eds.). *Phraseology in Foreign Language Learning and Teaching* (pp. 101-119). Amsterdam & Philadelphia: Benjamins.

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

Praninskas, J. (1972). *American University Word List*. Longman.

Queiroz, J. (2019). *The grammatical complexity of English noun phrases in Brazilian learners' academic writing: a corpus-based study*. MA thesis - Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Reppen, R. (2018). Teaching lexical bundles: Which ones and how? In E. Hinkel (Ed.). *Teaching essential units of language: Beyond single word vocabulary* (pp. 186-200). Routledge. https://doi.org/10.4324/9781351067737

Reppen, R. & Olson, S. B. (2020). Lexical bundles across disciplines. In U. Römer, V. Cortes & E. Friginal. *Advances in Corpus-based Research on Academic Writing*: *Effects of discipline, register, and writer expertise* (pp. 169-182). Amsterdam: John Benjamins.

Römer, U. (2010). Using general and specialized corpora in English language teaching: past, present and future. M. C. Campoy-Cubillo, B. Belles-Fortuno, & M. L. Gea-Valor, (Eds.). *Corpus-Based Approaches to English Language Teaching* (pp. 18-35). London: Continuum.

Salager-Meyer, F., de Segura, G. M. L. & Ramos, R. C. G. (2016). EAP in Latin America. In K. Hyland, & P. Shaw, (Eds.), *The Routledge Handbook of English for academic purposes* (pp. 109-124). New York: Routledge.

Sarmento, S.; Dutra, D. P.; Barbosa, M. V. & Moraes Filho, W. B. (2016) IsF e Internacionalização: da teoria à prática. In S. Sarmento, D. M. de Abreu-e-Lima.; W. B. Moraes

Filho. (Org.). *Do Inglês sem Fronteiras ao Idiomas sem Fronteiras: a construção de uma política linguística para a internacionalização* (pp. 77-100). Belo Horizonte: Editora UFMG.

Simpson-Vlach, R. & Ellis, N.C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512.

Sinclair, J. (1987). C*ollins COBUILD English language dictionary*. London: Collins.

Sinclair, J. (1991). *Corpus, concordance and collocation*. Oxford: Oxford University Press.

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

Viana, V; O'Boyle, A. (2022). *Corpus Linguistics for English for Academic Purposes* (Routledge Corpus Linguistics Guides) Abingdon: Taylor and Francis. Kindle Edition.

Welp, A., Didio, Á. & Finkler, B. (2019). Questões contemporâneas no cinema e na literatura: o desenho de uma sequência didática para o ensino de inglês como língua adicional. *Brazilian English Language Teaching Journal*, *10*(2), 1-25, DOI: https:// doi.org/10.15448/2178-3640.2019.2.3586

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

Xue, G. & Nation, P. (1984). A university word list. *Language Learning and Communication 3*, 215–29.

Zuppardi, M. C. (2020). *Collocation dimensions in academic English.* PhD dissertation. Pontifícia Universidade Católica de São Paulo, São Paulo.

Zuppardi, M. C. & Berber Sardinha, T. (2020). A multi-dimensional view of collocations in academic writing. U. Römer, V. Cortes, & E. Friginal, (Eds.), *Advances in Corpus-based Research on Academic Writing. Effects of Discipline, Register, and Writer Expertise* (pp. 334–353). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.95.14zup

Zuppardi, M. C., Veirano Pinto, M. & Berber Sardinha, T. (in prep.). Multi-Dimensional Analysis. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (2nd ed.). Hoboken, NJ: Wiley.

From specialized corpus to the EAP classroom: integrating authentic data into materials design

Ana Eliza Pereira Bocorny (UFRGS)
Ana Luiza Freitas (UFCSPA)
Rozane Rodrigues Rebechi (UFRGS)

Introduction

Almost two decades ago, Sinclair (2004a) anticipated that corpus-based language teaching would revolutionize language pedagogy. After all, relying on empirical evidence enables the design of pedagogical applications based on authentic input, providing teachers and researchers with an actual perspective of how language works. Today, the positive impact of corpus-based approaches to additional language learning and teaching is undeniable (Boulton & Cobb, 2017; Boulton, 2021; Karlsen, 2021; Anthony, 2022a; O'Keeffe, 2022).

Despite the importance of corpus linguistics as a means of identifying authentic language use and the fact that many studies (Flowerdew, 2009, 2013, 2014; Gray et al., 2020; Charles & Frankenberg-Garcia, 2021) suggest integrating corpus data into English for Academic Purposes[1] (EAP) pedagogy, the use of authentic data in language classrooms around the world is still incipient (Kavanagh, 2021; Poole, 2020; Pérez-Paredes, 2019). Moreover, according to Römer (2006: 122), "there is still a strong resistance towards corpora from the side of students, teachers, and materials writers."

1 The term English for Academic Purposes (EAP) refers to the English which is needed to study or conduct research in the academic context. Although it is often associated with non-native speakers of the language, EAP has extended also to native speakers who are faced with writing essays, presenting papers, reading articles, etc. (Charles, 2013).

Previous studies have suggested that "lack of time, group sizes, and technological obstacles" (Kavanagh, 2021: 2) could be standing in the way between corpus data and the language classroom. Poole (2020: 1) reports that although teachers embrace the use of corpus, they also reveal "emergent tensions regarding the use of ready-made corpus activities and the key affordances of discovery, authenticity, and autonomy often forwarded in support of corpus pedagogy." Breyer (2011: 207) claims that the lack of "(classroom) user-friendly concordancing software" was mentioned by teachers as one of the hurdles to the smooth adoption of corpora as language learning input. Other reasons identified by Mukherjee (2004: 243) had to do with the fact that not enough teachers were acquainted with "the basic foundations, implications, and applications of Corpus Linguistics."

Ranging from the context of graduate and undergraduate students from the Federal University of Rio Grande do Sul (UFRGS), this contribution arose from the needs of Brazilian pre-service and in-service EAP novice teachers when designing EAP writing course materials with corpus data at the Center of Languages for Academic Purposes (CLA)[2]. After being introduced to corpus linguistics principles and methods, these novice teachers were asked to design a Pedagogical Unit (PU), i.e., a set of learning activities sequenced together to promote advances in learning, for a given EAP course where selected language features would be taught within the context of a given academic genre. Those teachers were then asked to extract and analyze said language data and integrate it into their EAP materials.

Having this said, the aim of this chapter is twofold: (i) help EAP teachers better understand corpus linguistics methods for the extraction of language data from specialized corpora and (ii) show how said language data can be used in the design of EAP writing course materials through a pedagogical model that combines corpus and genre-based approaches.

The first section – 'Combining corpus and genre-based approaches' - reviews the literature on corpus and genre-based approaches to language learning and teaching and on pedagogical models that combine

2 CLA website: https://www.ufrgs.br/cla/

both approaches. Section 2 – 'The design of EAP materials' - describes the framework suggested in the study for designing EAP materials and presents a step-by-step guide on extracting and integrating corpus data into materials used for EAP writing courses. Finally, we finish the chapter with some final considerations and suggestions for further studies.

Combining corpus and genre-based approaches

Corpus Linguistics

According to Sinclair (1991: 171), "a corpus corresponds to a collection of natural texts chosen to characterize a state or variety of language". For Biber and Conrad (1999: 4), the notion of corpus is naturally approached from the perspective of register: "a collection of spoken or written texts, organized by the register and codified for other discursive considerations, comprises a corpus." McEnery and Hardie (2012: 1) define corpus linguistics as "an area which focuses upon a set of procedures, or methods, for studying language." As such, it can be applied to different areas.

Two central concepts are pillars of the field: the empiricist approach and the view of language as a probabilistic system. The empiricist system is based on the fact that knowledge originates from data organized in the form of a corpus. The view of language as a probabilistic system stems from the epistemological basis of the field, according to which linguistic traits do not happen randomly. Nevertheless, it is possible to point out and quantify patterns of regularity, highlighting a correlation between such traits and the situational contexts of use. From these patterns, it can be recognized that a language is not limited to empty spaces arbitrarily filled. Instead, the linguistic environment acts on the co-selection of lexical items. Within a linguistic environment, a given item prefers another one. This way, language is seen as a non-arbitrarily motivated and functional system of potential choices. These aspects refer to the issue of usage patterns and, therefore, to the idiomatic principle postulated by Sinclair (1991).

Let us take an example from the corpus used to extract linguistic data in this text. 'The aim of this study' is a sequence whose continuity is limited

by a word within the verb category 'be' followed by the preposition 'to', confirming a preference of academic textual genres/records (Hyland, 2008; Biber & Conrad, 1999) for a greater incidence of this association of words. Thus, the phrase above is expected to precede 'is not' or 'was to'.

Although the literature proposes many definitions for what constitutes a corpus (such as Atkins et al., 1992; Francis, 1992; Kennedy, 1998; McEnery et al., 2006), the consensus is that it should comprise:

1. Authentic Linguistic Data;
2. Readable Computer Segments;
3. Specially Organized Language Portions;
4. Texts Capable of Representing a Particular Language or Variety of Language.

For this chapter, a corpus is roughly understood as a set of machine-readable texts compiled with the aim to provide answers to specific research questions (McEnery & Hardie, 2012). To achieve these goals, a corpus should be built under well-defined criteria.

Corpus-Based Pedagogy

Since John Sinclair's seminal work on corpus research led to the use of corpus-based approaches (Sinclair, 1987, 1991, 2004b), corpus linguistics has always been connected with language teaching. Contributions such as Gavioli (2005), O'Keeffe et al. (2007), Aijmer (2009), Flowerdew (2012), and Cotos (2014), among others, all followed the principles of adopting empirical data to boost language learning. Hence corpus-based pedagogy is the application of corpus linguistics's foundations to facilitate the teaching and learning of additional languages springing from authentic occurrences of language.

Among the advantages of adopting corpora for language teaching are the possibilities of explaining the differences in the uses of words and linguistic forms, among other traits, based on the probability of occurrence in specific contexts (Biber et al., 1998), as intuition alone could not explain these facts (Sinclair, 1991). As pointed out by Shepherd (2009: 152), the analytical enterprise "cannot depend on the researcher's intuitions, since

human beings tend to recognize what is not typical more often than what is standardized". Corpora, therefore, are used to generate empirical knowledge about languages. Besides, using corpora for pedagogical purposes can disclose solutions to language queries that have not been dealt with otherwise. Furthermore, the use of corpora can highlight frequency patterns of words and language structures, and such patterns can be used to teach and create or improve teaching materials.

The most common tools used in corpus analysis for pedagogical purposes are concordancing programs, understood as text search engines with sorting functions, as will be demonstrated in the 'Step-by-step guide' to 'The design of EAP materials' below. Currently, among the most popular concordancing programs are WordSmith Tools (Scott, 2020), Sketch Engine (Kilgarriff et al., 2004), and AntConc 4.1 (Anthony, 2022b). As they are queried, these tools enable users to get in contact with "a collection of the occurrences of a word-form, each in its textual environment" (Sinclair, 1991: 32).

By using corpora for teaching purposes, users are empowered, as this approach holds the potential to foster autonomous and personalized learning (Boulton & Cobb, 2017; McEnery & Wilson, 1997). That happens because, on the one hand, the adoption of corpora encourages discoveries. Corpora can be employed, for example, to have students explore patterns of specific language features that stand out from the concordance lines. On the other hand, exploring language corpora by employing software enables learners within the same class to focus on different language features. Furthermore, corpus-based pedagogy can lead learners themselves to draw conclusions about language use and its principles.

Data Driven Learning (DDL)

As Boulton (2021: 9) affirms, "Data-driven learning (DDL) typically involves language learners consulting corpus data, either directly or via prepared materials, to answer questions about language." Some alleged benefits of using DDL are that it stimulates learners' autonomy and increases language awareness (Boulton, 2007). As for teachers, the use of DDL allows

for a change of roles from a lecturer to "a co-ordinator of student-initiated research" (Johns, 1991: 3). Nevertheless, the change of roles mentioned by Johns (1991) does not come without challenges, such as learning how to compile and extract language data from a corpus or how to include the language data extracted into the materials designed for EAP courses in a meaningful and contextualized way. Besides, employing DDL implies choosing which approach to be used, whether direct DDL, through hands-on activities (where you teach your learners how to look for information in the corpus) or indirect DDL, an approach through which you (teacher) previously extract the language data yourself and include them into pedagogical units.

Corpus processing systems like Sketch Engine (Kilgarriff et al., 2004), WordSmith Tools (Scott, 2020), AntConc (Anthony, 2022b), and #LancsBox v6 (Brezina et al., 2020) can be of great help. They usually offer varied resources to extract language features, such as lists of words, keywords, and n-grams. In Sketch Engine (SE), it is also possible to use Corpus Query Language (CQL) to create special search syntaxes or queries to look for more complex grammatical and lexical patterns (see 'Description of the EAP writing course', Table 5, for examples of language features and ways to retrieve them from the corpus using CQL queries). The smart search option available in #LancsBox v6 (henceforth, LancsBox) software package is another option for extracting more complex language patterns. Pérez-Llantada (2022), for example, uses the LancsBox smart search option to retrieve passive voice forms from four corpora.

To cater to the challenges mentioned above, in this chapter we provide EAP teachers with a step-by-step guide on retrieving and integrating corpus data into materials designed for EAP writing courses through indirect DDL. At this moment, we chose to focus on indirect DDL because we considered its simplicity an asset to encourage novice EAP teachers in their pursuits of work with corpus-based pedagogy.

Genre, Genre-Analysis, Move-Analysis and Genre-Based Pedagogy

Bhatia (1993: 13) defines genre as "a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs". For Swales (1990, 1994), these characteristics are organized from models that shape the structure of the text and guide specialists of the discursive communities in terms of content and style choices. While guiding members, these models are, at the same time, delimited by their motivations regarding the schematic formatting of the manuscript.

When Swales (1990) introduced criteria for defining the academic genre, he also established an organizational description of the conventions for introducing academic articles, which would become widespread. The structure, known as the Create a Research Space (CARS) model, comprises the description of the segments[3] that perform specific functions in the text, called rhetorical moves.

Next, we present the CARS model, as adapted from Swales (1990: 141), set into three moves that cover specific steps:

1. Move 1 – Establish the Territory
Step 1: Establish the importance of research and/or
Step 2: Make generalizations about the topic and/or
Step 3: Review the literature

2. Move 2 – Establish the Niche
Step 1a: Counterargue or
Step 1b: Indicate gap(s) in already established knowledge or
Step 1c: Raise questions or
Step 1d: Continue the tradition

3 Various labels have been used to refer to the information units observed from this format: moves and steps (Swales, 1990), moves and sub-moves (Santos, 1999), moves and subfunctions (Motta-Roth, 1995), moves and strategies (Araújo, 1999) and rhetorical units (Meurer, 1997).

3. Move 3 – Occupy the Niche
Step 1a: Outline the goals or
Step 1b: Submit the survey or
Step 2: Present the main results or
Step 3: Indicate the structure of the article.

The models for the rhetorical structure of genres are not prescriptions but classifications for didactic purposes. Therefore, as mentioned above, they are subject to variations that derive from the characteristics of the different research areas. According to Biber and Conrad (2009), academic texts do not encompass universal characteristics, but may vary situationally, given their publication conditions. However, the traits we recognize as the most constant show us what is most relevant and conventional to the user's discursive community in question. Likewise, such traits indicate what should be prioritized, as this investigation aims to highlight.

Genre pedagogy, genre-based pedagogy, and genre-based approach are some of the names given to the framework comprised of a set of assumptions, strategies, and practices for EAP teaching and learning that have as a premise the need to communicate a message to a particular audience in an appropriate way using discourse genres (for example, research papers, webinars, abstracts).

Swales's (1990: 9) genre pedagogy, as described in his seminal book *Genre Analysis: English in academic and research settings*, "rests on a pragmatic concern to help people, both non-native and native speakers, to develop their academic, communicative competence". It is essential to mention that, even though genre pedagogy has its origins in academic settings, the approach is used to teach different discourse genres.

Pedagogical Models Combining Corpus and Genre-Based Approaches

According to Charles (2020), even though corpus methods and genre analysis share a close connection, applications of such approaches for teaching purposes are not so frequent in practice. In said applications, both the target genre and the language features to be taught play a fundamental role. While the target genre serves as the starting point and the context

within which language features are built-in, the language data extracted from the corpus reveal patterns that are conventionally used by experts of the discourse community of a given discipline. Therefore, the language features to be taught should be selected according to their relevance to the chosen genre and students' needs.

As reported by Moreno and Swales (2018), the identification of linguistic features characterizing the various rhetorical moves of different genres for pedagogical purposes has been reported in many studies as the main aim of move analysis (for example, Cortes, 2013; Cotos et al., 2017; Kanoksilapatham, 2005; Le & Harrington, 2015; Swales, 1981). Moreno and Swales (2018: 41) highlight that filling the "function-form gap" involves "establishing the most salient types of text items, or patterns, occurring in a specific rhetorical context in an RA, or any other genre, that may lead a competent reader to interpret a given communicative function in a highly predictable manner". Few research methodologies and pedagogical models, though, have managed to converge these two analytic paradigms: the top-down, which involves investigations into "the rhetorical composition of texts through Swalesian (1981, 2004) move analysis", and the bottom-up, which refers to "investigations into the linguistic characteristics of texts through analysis of lexical, phraseological, grammatical, and lexico-grammatical patterns of use" (Gray et al., 2020: 261). Charles (2007: 289), for example, suggested reconciling top-down (discourse analysis) and bottom-up (corpus investigation) approaches as she presents EAP writing materials designed through "a pedagogic approach which combines discourse analysis with corpus investigation".

As the pedagogical model described above sets the scene for the EAP teaching and learning framework to be suggested in this chapter, it is essential to remember that another gap needs to be filled: the one between corpus linguistics and teaching practice. It is also noteworthy that initial decisions should be made in EAP course planning and materials design. An essential first step is to carry out a needs analysis in order to know the students' background (e.g., their language proficiency level, their background knowledge in the discipline they work with), their learning preferences (e.g., using inductive or deductive methods), as well as what they expect

and need from the course[4]. Also, decisions about which genre (e.g., oral presentation, research article), section (e.g., abstract, introduction, methodology, results), discipline (e.g., Nursing, Physics, Applied Linguistics), and language skill(s) (e.g., reading, listening, writing, speaking) the EAP course will focus on, need to be made. Information about the course to be taught and its target audience allows for defining clear and achievable learning objectives based on the learners' prior knowledge, skills, needs, preferences, and expectations. The choice of an appropriate methodology, the selection and design of materials, the feedback between learners and teachers, and the construction of knowledge that will be a consequence of this process are essential elements for designing and implementing EAP courses. It is always important to remember that course and materials design are not linear processes. Figure 1 shows an interplay between actions and procedures involved in implementing an EAP course, being the design of materials one of them:

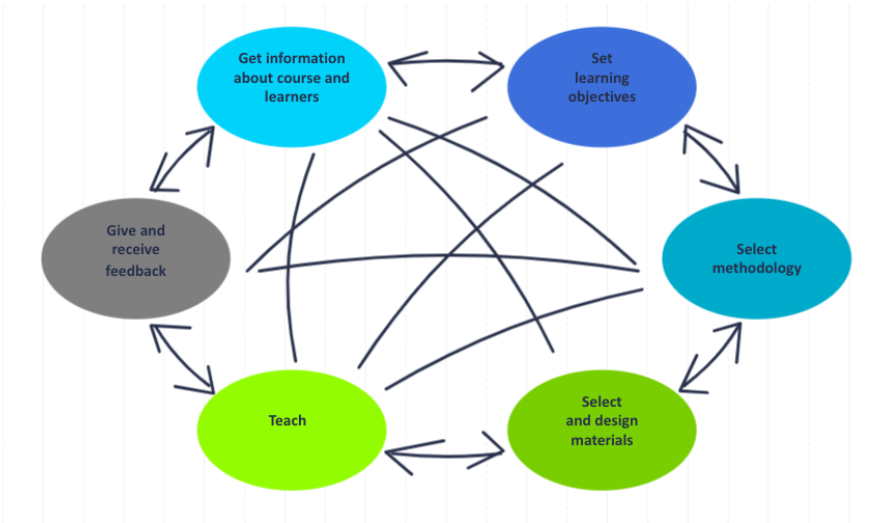


Figure 1. Stages involved in the process of designing and implementing an EAP course

4 See Viana et al. (2018) for a detailed overview of types of information that can be gathered in a needs analysis, the likely sources to be examined and methods that can be employed.

The design of EAP materials

Framework

Schneuwly and Dolz (2004: 51) define didactic sequences[5] as "a sequence of teaching modules, organized together to improve a given language practice." The authors advocate for having genres as the basis for organizing didactic sequences. With the genre as a starting point, the process of knowledge construction is scaffolded by tasks, activities, and exercises[6] designed according to specific guiding principles (Bocorny & Welp, 2021: 1601-1602), ultimately achieving pre-established learning objectives within a specific time frame.

For the design of activities with online corpora, Reppen (2010: 43) suggests a checklist with general guidelines;

- Have a clear idea of the point that you want to teach;
- Select the corpus that is the best resource for your lesson;
- Explore the corpus completely for the point you want to teach;
- Make sure that your directions are complete and easy to follow;
- Make sure that your examples focus on the point that you are teaching;
- Provide a variety of ways for interacting with the materials;
- Use a variety of exercises types;
- If you are using computers, *always* have an alternative plan or activity in the event of computer glitches.

In coursebooks, a pedagogical unit can be the focus of one or more classes, and its structure tends to be the same throughout the book. Table 1 shows the structure of the pedagogical unit and the section titles used in the EAP writing course presented as an example in this chapter:

5 In this study, the terms 'didactic sequences' and 'pedagogical units' are considered equivalent in meaning.

6 In this study, the term 'task' is used as a didactic plan to produce a communicative response from participants, comprising one or more sets of activities. The terms 'activity' and 'exercise', in turn, are considered equivalent in meaning, and, for this reason, they are used interchangeably in the sense of segments that make up a task.

PEDAGOGICAL UNIT STRUCTURE	SECTION TITLES OF A PEDAGOGICAL UNIT
Context of use, purpose and definition	1) Activate previous knowledge
Characteristics of the genre	2) Learn about key characteristics
Rhetorical structure	3) Find the parts
Language features	4) Know important language features
Production of genre	5) Analyze examples
	6) Write the first draft
	7) Get feedback
	8) Write the final draft

Table 1. Pedagogical unit structure for an EAP writing course

Welp et al. (2019: 6) list guiding principles to orient teachers in planning and designing general English teaching materials. Those principles were adapted by Bocorny and Welp (2021: 1601-1602) to guide the design of EAP materials:

1. Learning objectives should be established based on the knowledge area and academic needs of the group of learners the tasks are aimed at;
2. Target genres should be academically relevant and coherent with the established learning objectives;
3. Selected texts should be authentic and representative of social practices and genres that circulate in the academic context;
4. Tasks should offer the learners opportunities to use the language proper to the texts produced in the learners' domain and raise awareness on such use in a contextualized way;
5. Tasks dealing with linguistic resources should take into account the frequency of lexical and discursive items present in academic texts in the learners' area of knowledge;
6. Tasks' order and statements should be organized in a way to promote progress and scaffold learning;
7. Tasks should provoke relevant interactions between learners and texts, learners and learners and learners and teacher;
8. Task performance should provide meaningful learning opportunities and achieve results beyond the classroom.

Specifically, when it comes to the design of EAP materials within a framework that combines corpus and genre-based pedagogies, two elements are key: knowing the rhetorical structure of the target genre and identifying language features that are relevant to the genre that is being taught, considering the learners' prior knowledge, skills, needs, and expectations (see 'Corpus Linguistics' and 'Genre, Genre-Analysis, Move-Analysis and Genre-Based Pedagogy' above for details on both elements). In particular, it is vital to identify the language features used to realize the functions expressed in genre moves and steps. Moreno and Swales (2018: 40) mentioned that "A widely shared aspiration of move analysts has been to identify the linguistic features characterizing the various RA moves not only in English but also across languages."

A checklist for planning and designing EAP materials within a corpus and genre-based framework is proposed in the next section having in mind these two major elements, along with the guidelines suggested by Reppen (2010) and the principles put forward by Welp et al. (2019) and used by Bocorny and Welp (2021).

Step-by-step guide

This section is organized as a guide to be used by novice EAP teachers when designing materials within the proposed pedagogical model that combines corpus and genre-based approaches. We use the first five guiding principles suggested by Welp et al. (2019) and adapted by Bocorny and Welp (2021) as a checklist to be followed. Next, we provide brief explanations and describe some associated actions for each of the five first principles. Finally, examples of the proposed actions are presented, considering an EAP writing course for producing Health Sciences structured abstracts.

Description of the EAP writing course

As can be seen in Table 2, structured abstracts are the target genre of the course, which is aimed at upper-intermediate (B2, C1) Health Sciences graduate students and researchers. The course is to be taught online with a

total of 16 hours divided into 8 hours of synchronous activities and 8 hours of asynchronous activities:

Name of the course	Written production of structured abstracts in the area of Health Sciences
Target genre	Structured abstracts
Target section	All sections
Students level of proficiency	Upper-intermediate (B2, C1)
Students level of education	Tertiary level (graduate students)
Course modality	Online
Length of the course	4 week course (16 hours: 8 hours of synchronous activities and 8 hours of asynchronous activities)

Table 2. Description of the EAP writing course

PRINCIPLE 1. Learning objectives should be established based on the knowledge area and academic needs of the group of learners the tasks are aimed at

EXPLANATION: A learning objective is a description of what the learner should be able to do upon successful completion of an educational step (for example, course, task, exercise/activity) over a period of time. Clearly defined learning objectives specify the knowledge, skills, and/or attitudes the learner will gain from the educational step so that such aspects can be assessed later on.

EXAMPLE: As can be seen in Table 3, there are two types of learning objectives for the course described: (i) the course learning goal, which is the outcome that is expected after its successful conclusion (being able to produce a structured abstract in the area of Health Sciences to be submitted to a journal in the area) and (ii) the learning goal of each class. The fruitful accomplishment of each of these goals is verifiable through implementing pedagogical tasks:

Learning objective of the course	By the end of this course, participants should be able to produce a structured abstract in the area of Health Sciences to be submitted to a journal in the area.
Learning objective of class 1	By the end of this class, participants should be able to understand what a structured abstract is and in which contexts it is used in the area of Health Sciences.
Learning objective of class 2	By the end of this class, participants should be able to recognize the rhetorical structure of a structured abstract in the area of Health Sciences.
Learning objective of class 3	By the end of this class, participants should be able to use language features relevant to producing a structured abstract in the area of Health Sciences.
Learning objective of class 4	By the end of this class, participants should be able to produce the first draft of a structured abstract in the area of Health Sciences.

Table 3. Learning objectives for course and classes

PRINCIPLE 2. The target genres should be academically relevant and coherent with the established learning objectives

EXPLANATION: The target genre is the one that is going to be worked with along the course. As it has already been mentioned (see 'Framework'), within the framework proposed, two elements are central: knowing the rhetorical structure of the target genre and identifying relevant language features. Many patterns representing the rhetorical structure of academic genres can be found in the literature. Can et al. (2016: 4), for example, present the rhetorical structure of abstracts within Applied Linguistics, as shown in Figure 2:

Abstract Moves (Pho [20])	Function/Description	Question Asked	Move Labels along with Abbreviations in the Present Study
Situating the research	setting the scene for the current research	What is known in the field?	introduction (I)
Presenting the research	stating the purpose of the study, research questions and hypotheses	What is the study about?	purpose (P)
Describing the methodology	describing the materials, subjects, variables, procedures, etc.	How was the research done?	methods (M)
Summarizing the findings	reporting the main findings of the study	What did the researcher find?	results (R)
Discussing the research			
(a)	interpreting the results/findings and/or giving recommendations	What do the results mean?	discussion (D-a)
(b)	no discussions or recommendations		pseudo-discussion (D-b)

Figure 2. Rhetorical structure of Applied Linguistics abstracts. From Can et al. (2016: 4)

The rhetorical structure of a given genre can also be obtained by using: (i) text structure analyzers like AntMover (Anthony, 2003); (ii) rhetorical tagging or rhetorical move-step coding (Bondi, 2022; Berdanier, 2019; Gray et al., 2020; Yoon & Casal, 2020a; 2020b; Geluso, 2019) or, concerning structured abstracts, (iii) the section headings, as suggested by Freitas and Bocorny (2021).

EXAMPLE: The target genre of the course described is structured abstracts, that is, abstracts that "describe a study using specific content headings rather than paragraph format" (Stevenson & Harrison, 2009: 1). Figure 3 exemplifies the rhetorical structure aimed at in a writing course for structured abstracts in health sciences:

ABSTRACT

Introduction: The Brazilian Ministry of Health had planned face-to-face workshops for professional training about the Clinical Protocols and Therapeutic Guidelines for Comprehensive Care for People with Sexually Transmitted Infections for the year 2020. Due to the COVID-19 pandemic, the workshops were cancelled, and a new strategy was adopted: virtual meetings, called Webinars — Clinical Protocols and Therapeutic Guidelines for Comprehensive Care for People with Sexually Transmitted Infections 2020. **Objective:** To report the experience at the Ministry of Health in online training about the clinical protocol and therapeutic guidelines for comprehensive care for people sexually transmitted infections for health professionals in 2020. **Methods**: The webinars were held in partnership with the Brazilian Society of Sexually Transmitted Diseases and the Pan American Health Organization. Each chapter of the Clinical Protocols and Therapeutic Guidelines for Comprehensive Care for People with Sexually Transmitted Infections — 2020 was converted into a webinar, with the participation of at least three experts, two speakers, and a moderator. **Results**: In total, 16 webinars were presented, covering topics such as sexually transmitted infections surveillance, prevention, diagnosis, treatment, public policies, and sexual violence. The initiative had more than 77,000 hits, with an average of 4,900 hits per webinar and the topic "syphilis" being the most accessed. The event reached all 27 federative units of Brazil, as well as 27 other countries. About 500 questions were received from the audience and answered during the sessions and/or through a document published later on by the Ministry of Health. **Conclusion**: Given the high number of hits and inquiries received, we can conclude that health professionals remained engaged in the topic of sexually transmitted infections during the pandemic. This experience shows the great potential of innovative methods for distance learning to promote continuing education, including a series of webinars aimed at strengthening the fight against sexually transmitted infections.

Keywords: Sexually transmitted infections. Education, continuing. Professional training. Clinical protocols. Education, distance.

Figure 3. Example of a structured abstract in Health Sciences. From Gaspar et al. (2022: 2)

The example of the rhetorical structure frequency distribution shown in Figure 4 was extracted from three corpora of structured abstracts in the area of Epidemiology using the section headings, as suggested by Freitas and Bocorny (2021). To obtain the rhetorical structure shown in Figure 4, the following CQL was used in Sketch Engine: <s> []{1,3} [word=":"]:

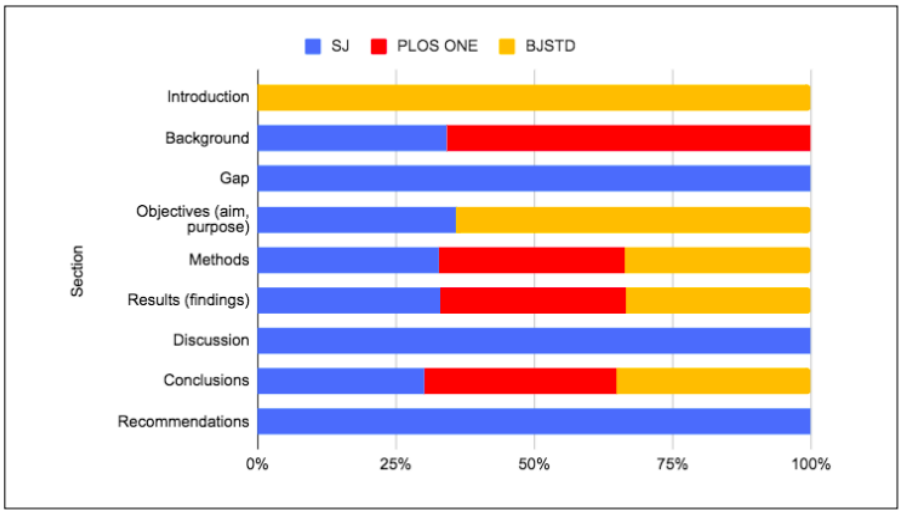


Figure 4. Rhetorical structure of Epidemiology structured abstracts. From Freitas and Bocorny (2021: 3)

As seen in Figure 4, the section headings in all the three corpora are Methods, Results/Findings, and Conclusions, and in two corpora, Background and Objectives (aim, purpose). The procedure for identifying SECTION HEADINGS used in this study is described below.

PROCEDURE 1:

1) Go to Sketch Engine
2) Select the corpus you want to work with
3) Go to Concordance
4) Select Advanced
5) Click on CQL
6) Paste the CQL <s> []{1,3} [word=":"]
7) Click on GO

The results from **PROCEDURE 1** are shown in Figure 5. These headings can be categorized into families representing the sections of the structured abstracts of the discipline under study:

136	doc#3620 ıonitoring the cancer burden in this at-risk population.</s><s>	Aims:	This retrospective case-control study was aimed at identifying
137	doc#3632 ɔd lymphoma risk and medical radiation associations.</s><s>	Aims:	The European Senior Program (ESP) aims to avoid waiting lis
138	doc#3779 vices, which is amplified by the COVID-19 pandemic.</s><s>	Aims:	We studied the pattern of spatial association between post-acı
139	doc#3810 roke among different CKD stages are not well known.</s><s>	Aims:	We aimed to investigate whether the severity of CKD would im
140	doc#3829 ər from those observed in day-to-day clinical practice.</s><s>	Aims:	To compare the risk of stroke/systemic embolism (S/SE) and r
141	doc#3914 ı become more prevalent with increasing prematurity.</s><s>	Aims:	To investigate the association between PCOS and extremely p
142	doc#4036 inal fusion surgery without additional adverse effects.</s><s>	Aims:	The aims of this study were to examine the prevalence of hosp
143	doc#2011 and weight loss and thus worsening the quality of life.</s><s>	Aims and methods:	Our aim was to find correlations from a multicentre database r
144	doc#2776 ɔlate with the degree of liver fibrosis in these patients.</s><s>	Aims and methods:	To investigate the accuracy of noninvasive scoring systems in
145	doc#2966 ɔ.</s><s>GPS data were collected for all drug shops.</s><s>	Analysis:	Quantitative data were analyzed using SPSS for descriptive st
146	doc#0 ɔd link between inflammation and cancer progression.</s><s>	Author Summary:	Cancer progression has been depicted as a linear process, du
147	doc#1 led design of solid tumor immunotherapy in the clinic.</s><s>	Author Summary:	Among the many potential drugs explored within the scope of
148	doc#2 ı to effectively utilize existing drugs for new purposes.</s><s>	Author Summary:	The combination of distinct drugs in combinatorial therapy can
149	doc#3 tegies that best inhibit diverse tumor cell populations.</s><s>	Author Summary:	Immunologic surveillance is a function of the immune system v
150	doc#4 ɹs disease, computer viruses, or ecological networks.</s><s>	Author Summary:	WHO/CDC recommendations prioritize influenza vaccinations
151	doc#5 ɗied, for both influenza and other infectious diseases.</s><s>	Author Summary:	The spread of infectious diseases can be inhibited by both vac
152	doc#6 ly assessed and compared with previous pandemics.</s><s>	Author Summary:	The ever-increasing availability of timely, large-scale clinical ep
153	doc#7 ınt on spatial interactions between metastatic lesions.</s><s>	Author Summary:	We used mathematical modelling to formalize the standard the
154	doc#4038 but may also pave the way for interventional studies.</s><s>	Author Summary:	Malaria remains a major source of morbidity and mortality thro
155	doc#3468 n patients with N1 HNC when combined with surgery.</s><s>	Backgroud:	Resection is still the only potentially curative treatment for pati
156	doc#3479 women living with advanced breast cancer in Ghana.</s><s>	Background & aim:	In clinical practice, transarterial chemoembolization (TACE) ha

Figure 5. Section heading of the structured abstracts being studied

PRINCIPLE 3. The selected texts should be authentic and representative of social practices and genres that circulate in the academic context

EXPLANATION: An authentic and representative sample of texts to extract language data to inform materials design can be obtained in existing freely-available corpora (for example, COCA[7], MICUSP[8], CODISSAE[9]). However, suppose you want to design a pedagogical unit of a genre (or section of a genre) that is not available in the existing freely-available corpora. In that case, you can compile your corpus using tools like AntCorGen (Anthony, 2022b)[10] or Sketch Engine (Kilgarriff, 2004)[11]. AntCorGen, for example, is very useful for designing tasks and exercises for discipline and section-specific EAP writing courses on research articles or abstracts, that is, EAP courses that focus on one of the sections of research articles within a particular discipline. Now, suppose you want to work with a more specific genre within a particular area. In that case, you may have to compile your corpus manually and upload it to a tool that will enable language data extraction.

EXAMPLE: Three corpora were compiled for the course on the **Written Production of Health Sciences Structured Abstracts**. As described by Freitas and Bocorny (2021), the corpora comprise abstracts from Epidemiology articles published in peer-reviewed indexed journals between 2003 and 2021. Their characteristics are represented in Table 4:

7 https://www.english-corpora.org/coca/

8 http://micusp.elicorpora.info/

9 https://drive.google.com/drive/folders/145ZFPOUuCwvTWFirM-lqG1vGbD-1g7p7o?usp=sharing

10 https://www.laurenceanthony.net/software/antcorgen/

11 https://www.sketchengine.eu/blog/build-a-corpus-from-the-web/

Domain	Corpus	Words with repetition (tokens)	Words without repetition (types)	Texts	Average words per abstract
Epidemiology	SJC	662,747	21,087	1,915	346
Epidemiology	PLOS ONE	1,000.003	43,066	4,330	230
Epidemiology	BJSTD	83,261	9,010	360	231

Table 4. Numbers of corpora used in the study. From Freitas and Bocorny (2021: 2)

PRINCIPLE 4. The tasks should offer the learners opportunities to use the language proper to the texts produced in the learners' domain and promote reflections on such use in a contextualized way

EXPLANATION: After compiling the corpus that will be used to inform the design of tasks and exercises within a pedagogical unit, it is time to choose a language feature (or language features) that will be focused on. Said language feature needs to be proper and relevant to the texts produced in the learners' knowledge area. The decision on which language features to focus on in EAP courses can challenge novice EAP teachers. Some of these features have been addressed in different studies as relevant for producing academic genres. Swales and Feak (2009), for example, mention tenses (past tense x simple present tense), passive voice, metadiscoursal expressions, lexical bundles, 'that' clauses, reporting verbs, pronouns (I, we). Kanoksilapatham (2005) refers to passive constructions, past tense, 'that' clauses, and metatextual devices. Table 5 provides examples of language features and ways of retrieving them from corpora using SE CQL queries. It is important to emphasize that the previous identification of language features elicited by learners as relevant also works as a compass needle pointing to what to focus on.

Language feature to be analyzed	Way to extract language feature using SE CQL queries
Sentence voice	**Passive voice:** []{1,5} [tag="VBD.*" | tag="VBG" | tag="VBN" | tag="VBP" | tag="VBZ"] [tag="VVN"] **Passive voice in each section of a structured abstract:** <s> []{1,3} [word=":"] []{1,5} [tag="VBD.*" | tag="VBG" | tag="VBN" | tag="VBP" | tag="VBZ"] [tag="VVN"] **Obs:** It is possible to FILTER the results obtained in the previous search by section heading or specific words (for example, the word 'by') to obtain concordance lines with **passive voice in section CONCLUSION of a structured abstract followed by the word 'by'. See Appendix 5 for results.**
Pronouns (I, we)	**Pronouns in each section of a structured abstract:** <s> []{1,3} [word=":"] [lemma="we" | lemma="I"]
Lexical Bundles	**Lexical bundles in each section of a structured abstract** <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4} **Obs:** In this case, the word 'study' can be replaced by any of the collocation nodes identified in the wordlist (see Figure 11)

Table 5. Some language features and ways of retrieving them from corpora using SE CQL queries.

Some of these language features are easier to extract and analyze. Imagine that one of your students wants to know whether to use 'I' or 'we'[12] when writing structured abstracts. Simply checking the wordlist for pronouns will show that, in our study corpus, 'we' occurs 3,345 times per million words (pmw) while 'I' occurs 95 times (pmw). If your students want to know which pronoun is more conventional in the different sections of structured abstracts in initial position, after the section heading (for example, 'CONCLUSION: We concluded that'), it is possible to use the CQL <s> []{1,3} [word=":"] [lemma="we" | lemma="I"]. All the 1,037 concordance

12 Previous research has explored the role of personal pronouns in academic writing (Henderson & Barr, 2010; Martínez, 2005; Hyland, 2002). According to Hyland (2002), a solid authorial identity that refers to authors taking 'ownership' for their work has to do with the use of self-reference in active voice constructions (where personal pronouns are used) as opposed to the anonymity of passive forms.

lines obtained with this query show section headings followed by the pronoun 'we'. This information could orient an exercise on authorial identity (see footnote 11) and on the use of pronouns in a course on writing structured abstracts.

EXAMPLE: For the course on **Written Production of Structured Abstracts in Health Sciences**, the language feature selected was Lexical Frames (LFs), that is, discontinuous sequences of words forming a structure around variable slots (Gray & Biber, 2013). According to Gray and Biber (2013), written academic discourse relies primarily on LFs. For this reason, that language feature has great pedagogical importance in written academic genres.

PRINCIPLE 5. Tasks dealing with linguistic resources should take into account the frequency of lexical and discursive items present in academic texts in the learner's area of knowledge

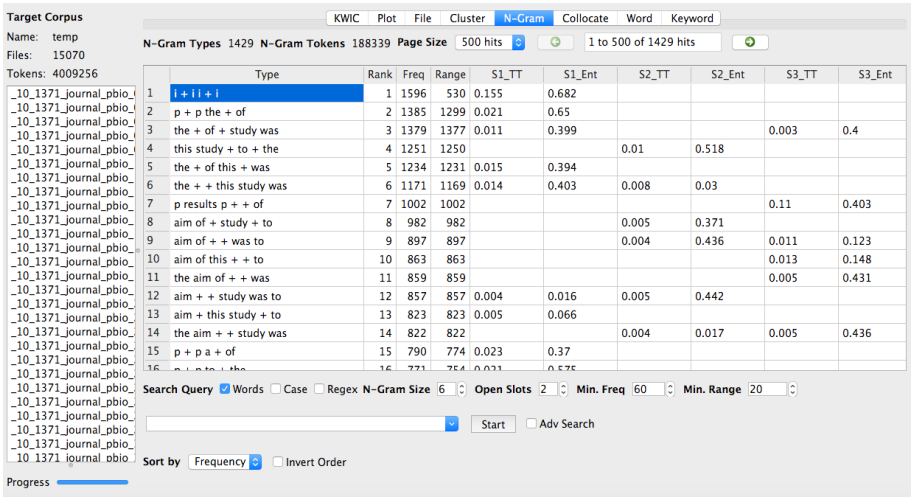
EXPLANATION: The lexical and discursive items selected as language features should be conventional. In other words, they should reveal the language used by the expert discourse community of a given discipline.

EXAMPLE: Learning about tools that can facilitate the teacher's access to linguistic data obtained from corpora might help bridge the gap between corpus linguistics and language teaching (Cheng, 2010). Different methodologies (for example, bundles-to-frames approach and fully inductive approach[13]) and tools (for example, AntGram 0.0.3 (Anthony, 2017), AntConc 4.1 (Anthony, 2022b)[14], WordSmith Tools 8.0 (Scott, 2000), KfNgram 1.3.1

13 Bundles-to-frames approach (Biber, 2009; Römer, 2010) and fully inductive approach (Gray & Biber, 2013) are methodological procedures for identifying LFs in a corpus. While, according to Gray and Biber (2013), the former starts by finding the most frequent continuous lexical sequences in a register and then analyzes the sequences to determine if they are associated with discontinuous lexical frames with variable slots, the latter "directly identifies the full set of discontinuous sequences in a corpus" (Gray & Biber, 2013: 111).

14 The use of different versions of AntConc implies the impossibility of extracting certain data related to Lexical Frames.

(Fletcher, 2012)) have been suggested for the extraction of LFs. AntConc 4.1 is, in our opinion, the most user-friendly tool for extracting LFs. Figure 6 shows the LFs extracted from the corpus of Health Sciences RA structured abstracts with AntConc 4.1 (Anthony, 2022b). The criteria used for the extraction was: n-gram size = 6, open slots = 2, minimum frequency = 60, minimum range = 20.



	Type	Rank	Freq	Range	S1_TT	S1_Ent	S2_TT	S2_Ent	S3_TT	S3_Ent
1	i + i i + i	1	1596	530	0.155	0.682				
2	p + p the + of	2	1385	1299	0.021	0.65				
3	the + of + study was	3	1379	1377	0.011	0.399			0.003	0.4
4	this study + to + the	4	1251	1250			0.01	0.518		
5	the + of this + was	5	1234	1231	0.015	0.394				
6	the + + this study was	6	1171	1169	0.014	0.403	0.008	0.03		
7	p results p + + of	7	1002	1002					0.11	0.403
8	aim of + study + to	8	982	982			0.005	0.371		
9	aim of + + was to	9	897	897			0.004	0.436	0.011	0.123
10	aim of this + + to	10	863	863					0.013	0.148
11	the aim of + + was	11	859	859					0.005	0.431
12	aim + + study was to	12	857	857	0.004	0.016	0.005	0.442		
13	aim + this study + to	13	823	823	0.005	0.066				
14	the aim + + study was	14	822	822			0.004	0.017	0.005	0.436
15	p + p a + of	15	790	774	0.023	0.37				

Figure 6. LFs extracted with AntConc 4.1 described in PROCEDURE 2. From Anthony (2022b)

PROCEDURE 2:

1) Open AntConc 4.1
2) Upload the corpus you want to work with
3) Click on N-Gram
4) Select the extraction criteria (in this extraction we used n-gram size = 6, open slots = 2, minimum frequency = 60, minimum range = 20).
5) Click on START

The results show the most recurrent LFs in this corpus. It is possible to see that the most frequent units are those that linguistically express the rhetorical function 'presenting the aim of the study'. If you double-click on

one of the LFs (for example, 'this study + to + the'), you can see the unit in context, as shown in Figure 7:

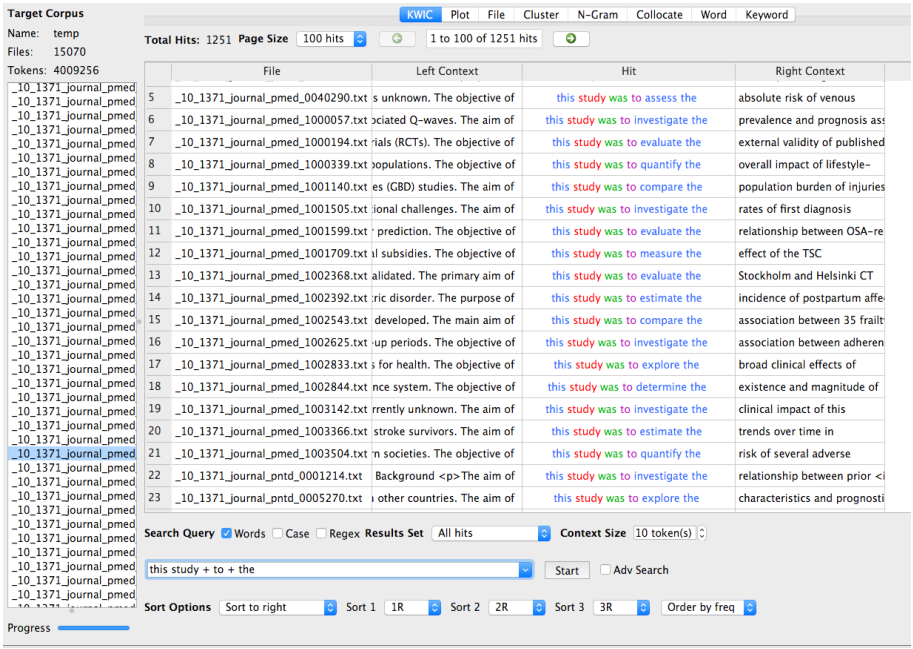
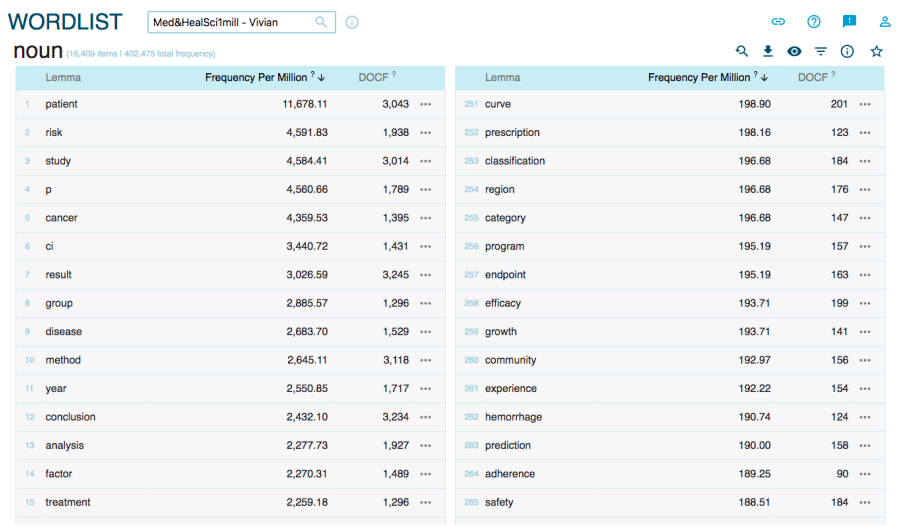


Figure 7. LF 'this study + to + the' in context. From Anthony (2022a)

The LFs extracted with AntConc 4.1 can 'inspire' the creation of a CQL that could be used in SE to identify the LFs used in the different sections of the structured abstracts. For example, the LF 'the + of + study was' can lead to the following CQL [lemma="the"] [tag="N.*"] [lemma="of"] [lemma="this"] [lemma="study"] [tag="VB.*"] [lemma="to"] [tag="V.*"]. To extract the LF in different sections of structured abstracts, this CQL should contain <s> []{1,3} [word=":"]. Hence, the CQL becomes: <s> [] {1,3} [word=":"] [lemma="the"] [tag="N.*"] [lemma="of"] [lemma="this"] [lemma="study"] [tag="VB.*"] [lemma="to"] [tag="V.*"].

Another way of identifying recurrent LFs in sections of structured abstracts is by having collocation nodes as a starting point. Following Flowerdew (2013), Freitas and Bocorny (2021) used a combination of lexical and phraseological elements to extract LFs from Epidemiology RA structured abstracts. A list of frequent noun collocation nodes was used

"as a starting point for collocation look-ups" (Frankenberg-Garcia et al., 2021: 208). As can be seen in Figure 8, the five most frequent nouns in the Epidemiology PLOS ONE study corpus were 'patient', 'risk', 'study', 'cancer', and 'result'. Collocation nodes could also be found in other word classes, like verbs, adjectives, adverbs, and prepositions:



WORDLIST Med&HealSci1mill - Vivian

noun (16,409 items | 402,475 total frequency)

	Lemma	Frequency Per Million	DOCF
1	patient	11,678.11	3,043
2	risk	4,591.83	1,938
3	study	4,584.41	3,014
4	p	4,560.66	1,789
5	cancer	4,359.53	1,395
6	ci	3,440.72	1,431
7	result	3,026.59	3,245
8	group	2,885.57	1,296
9	disease	2,683.70	1,529
10	method	2,645.11	3,118
11	year	2,550.85	1,717
12	conclusion	2,432.10	3,234
13	analysis	2,277.73	1,927
14	factor	2,270.31	1,489
15	treatment	2,259.18	1,296

	Lemma	Frequency Per Million	DOCF
251	curve	198.90	201
252	prescription	198.16	123
253	classification	196.68	184
254	region	196.68	176
255	category	196.68	147
256	program	195.19	157
257	endpoint	195.19	163
258	efficacy	193.71	199
259	growth	193.71	141
260	community	192.97	156
261	experience	192.22	154
262	hemorrhage	190.74	124
263	prediction	190.00	158
264	adherence	189.25	90
265	safety	188.51	184

Figure 8. Noun wordlist for the Health Sciences PLOS ONE study corpus. From Kilgarriff et al. (2004)

Using Sketch Engine and searching for concordance lines with the lemma 'study' as a noun, it is possible to retrieve language data that could be easily integrated into exercises to be used in the course **Written Production of Structured Abstracts in the Area of Health Sciences**. Figure 9 shows the results:

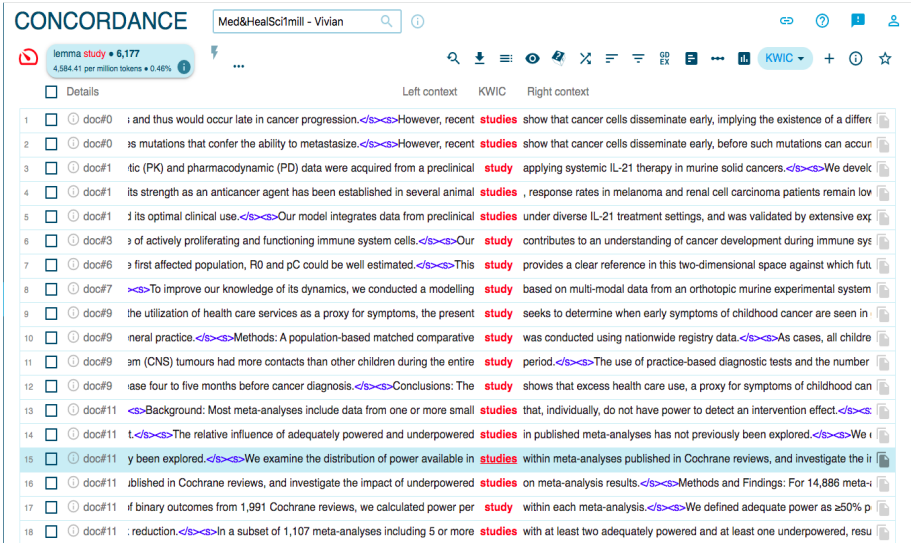


Figure 9. Concordance lines with the lemma 'study' as a noun. From Kilgarriff et al. (2004)

PROCEDURE 3:

1) Open Sketch Engine
2) Select the corpus you want to work with
3) Choose Concordance
4) Select Advanced
5) Click on lemma, in Query type
6) Click on noun, in Part of speech
7) Write 'study' (or any other recurrent collocation node) under Lemma
8) Press GO

Figure 10 illustrates the search for 'study':

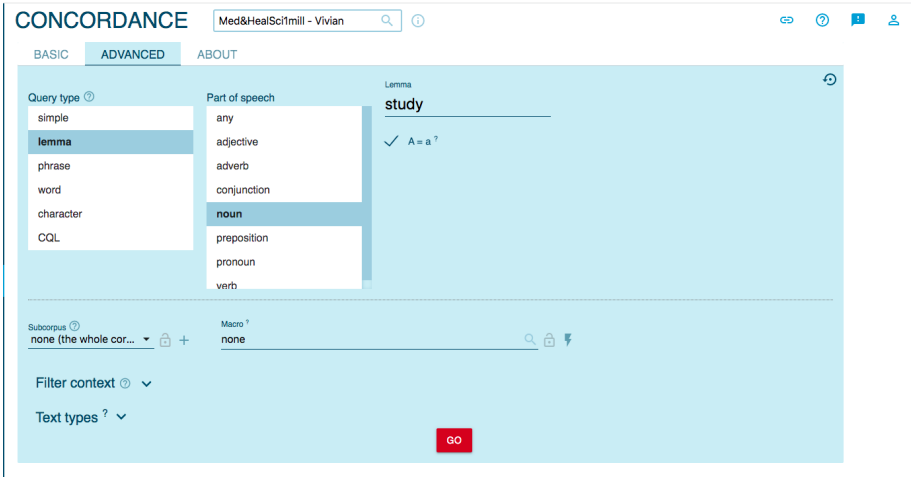


Figure 10. SE interface for PROCEDURE 3. From Kilgarriff et al. (2004)

The results obtained with PROCEDURE 3 can be filtered for each structured abstract recurrent section heading: (METHODS, RESULTS/ FINDINGS, CONCLUSIONS BACKGROUND, and OBJECTIVES/AIM/ PURPOSE). For example, Figure 11 shows the filtered results of concordance lines with the lemma 'study' for the section CONCLUSIONS:

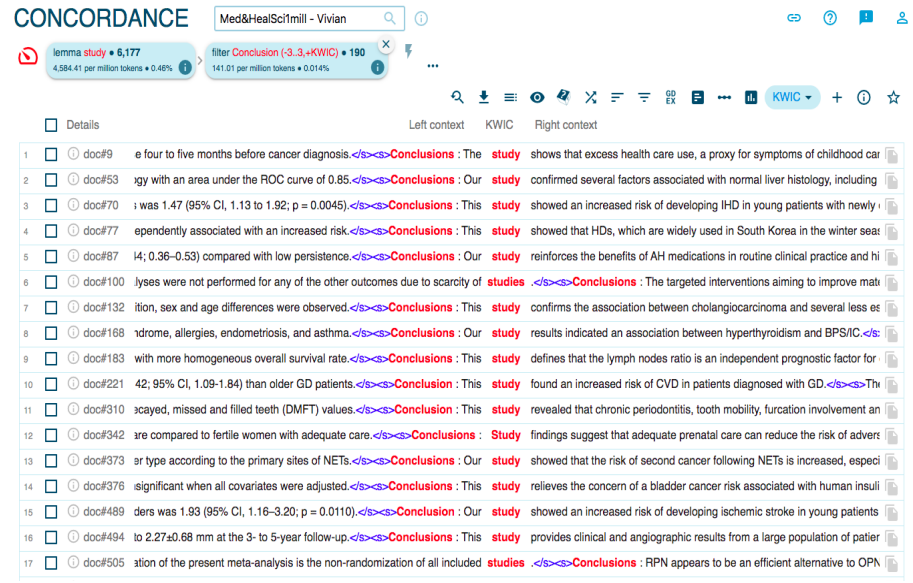


Figure 11. Filtered results of concordance lines with the lemma 'study' for the section CONCLUSIONS. From Kilgarriff et al. (2004)

PROCEDURE 4 presents the steps for filtering data:

1) Use the results obtained with PROCEDURE 3 (search for the lemma 'study', as a noun)
2) Click on the Filter icon, as shown in Figure 12:

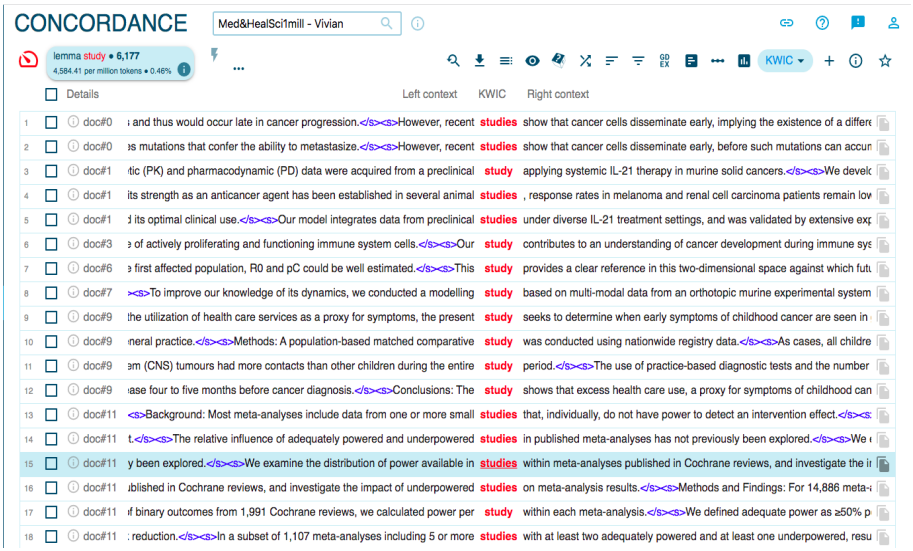


Figure 12. Filtering data in SE. From Kilgarriff et al. (2004)

3) Select Advanced
4) Click on lemma, in Query type
5) Click on noun, in Part of speech
6) Write 'Conclusion', under Lemma
7) Press GO

Figure 13 illustrates the search:

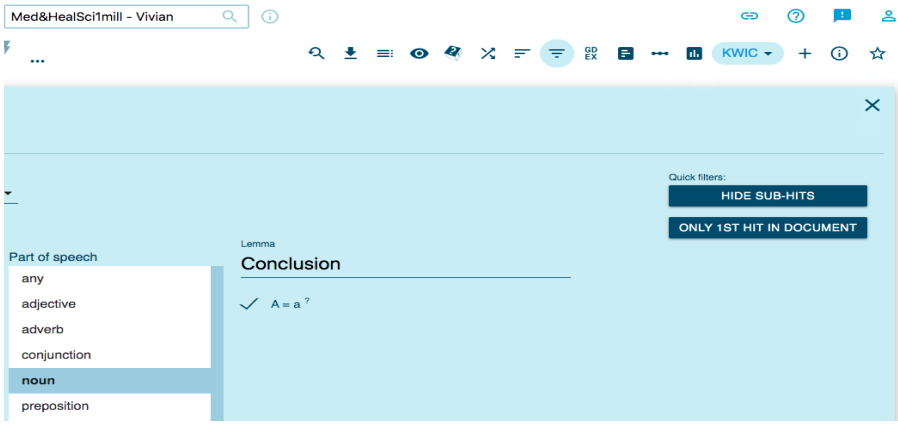


Figure 13. SE interface for PROCEDURE 4. From Kilgarriff et al. (2004)

If you want to organize the results obtained with PROCEDURE 4, you can click on the icon SORT (to the left of the FILTER icon). The results obtained are shown in Figure 14:

Figure 14. Sorting data in SE. From Kilgarriff et al. (2004)

A more direct way of finding recurrent LBs (and afterwards the LFs) in the sections of structured abstracts is to use Corpus Query Language (CQL) syntaxes. The CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4}, for example, extracts all the collocations that occur in the sections of

the structured abstracts that have 'study' as a collocation node. In this case, the collocation node 'study' can be replaced by any of the collocation nodes identified in the wordlists extracted from the corpus. Figure 15 shows the results when using this CQL:

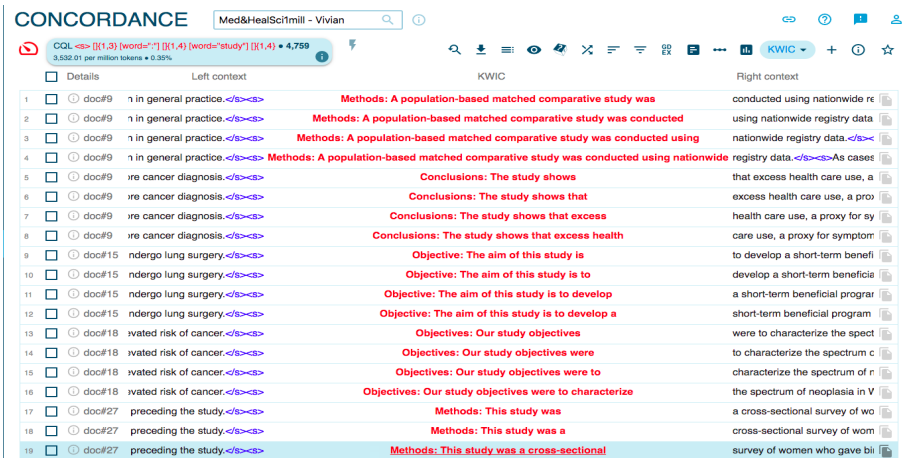


Figure 15. Results for the CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4}. From Kilgarriff et al. (2004)

PROCEDURE 5:

1) Open Sketch Engine
2) Select the corpus you want to work with
3) Go to Concordance
4) Select Advanced
5) Click on CQL, in Query type
6) Paste the CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4} under CQL
7) Press GO
8) Click on KWIC (to organize the results alphabetically)

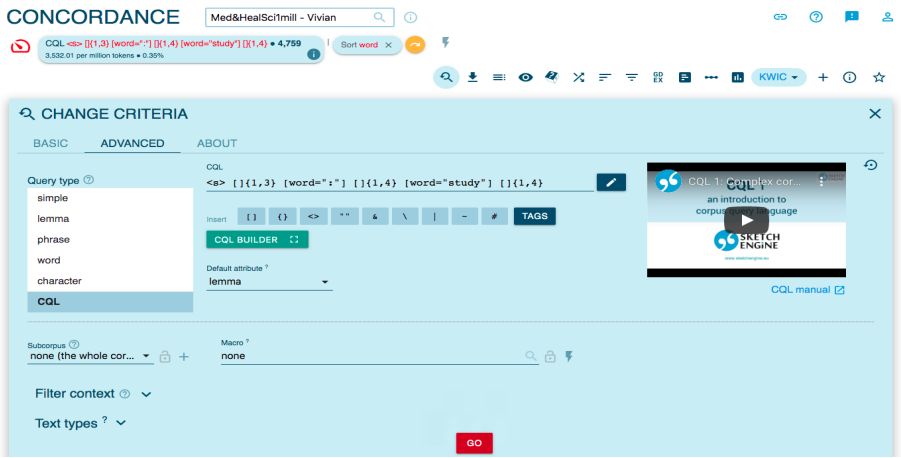


Figure 16. SE interface for extracting LBs from sections of the structured abstracts using CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4}. From Kilgarriff et al. (2004)

The results in Figure 16 indicate that collocations with 'study' occur across sections of these structured abstracts. These results can also be filtered for each section identified as part of the rhetorical structure of the abstracts under study. For example, as shown in Figure 17, collocations with the word 'study' occur 440 times in the section CONCLUSION in the corpus of Health Sciences:

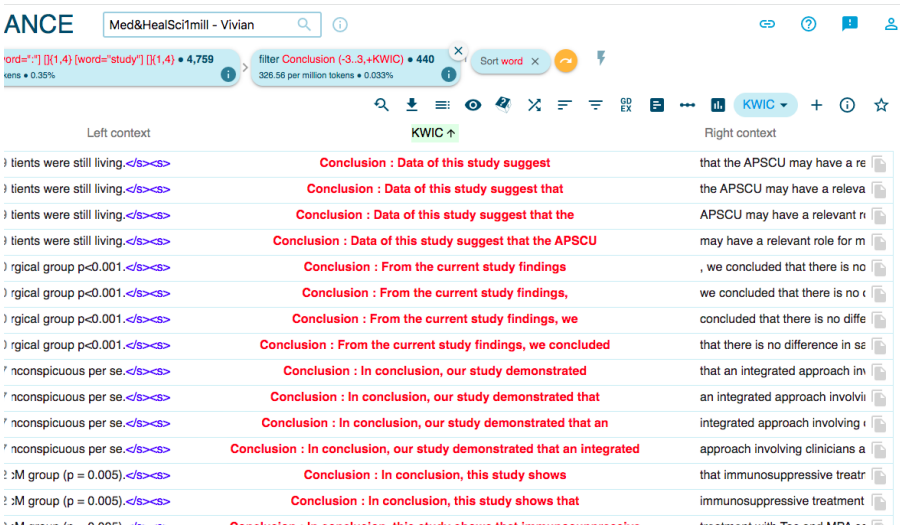


Figure 17. Collocations with the word 'study' filtered for the section CONCLUSION. From Kilgarriff et al. (2004)

The collocations extracted with the node 'study' filtered for the section CONCLUSIONS show different LFs that can be used in exercises. An example is the LF shown in Table 6, below:

*	*	**study**	*	that
-	The		showed (68x)	
The results of (25x)	Our		shows (48x)	
	This		suggests (54x)	
			suggested (6x)	
			indicates (24x)	
			indicated (8x)	

Table 6. LF with the node 'study'

As can be seen in Table 6, the LF ***(The, Our, This) study *(show(ed), suggests, indicates)** is a chunk of language that can be taught as an option to be used at the beginning of the section CONCLUSION(S) in structured abstracts in Health Sciences. 'The results of' precedes some of the sentences where this LF occurs. 'Showed' is the most recurrent slot filler after the collocation node 'study'. The procedure of filtering, shown in Figure 12, can be done with the other sections of structured abstracts to identify LFs to be

included in exercises with the LFs that are recurrent in different sections of structured abstracts.

Concluding remarks

As aforementioned, this chapter drew from the needs of Brazilian pre-service and in-service EAP novice teachers, graduate and undergraduate students from the Federal University of Rio Grande do Sul (UFRGS), all teachers at CLA (Center of Languages for Academic Purposes). While the COVID-19 pandemic obliged us to stay home for two years and two months, we held weekly online pedagogical meetings. During these meetings, we reported and reflected upon our online classroom experiences, to find solutions to problems that we had never faced before. Moreover, we discussed language learning and teaching theories. Finally, we planned courses and classes. However, above all, we tried to figure out how corpus linguistics and genre studies could guide us to design materials to help our students, the Brazilian academic community, to write more conventional academic texts. The insights that came up from these meetings guided the writing of this chapter.

During this period, we identified that novice EAP teachers were not confident using corpus linguistics to inform their teaching practice, even though this approach has been proved effective by many scholars. With this gap in mind, we created a framework drawing on the principles proposed by Welp et al. (2019) and adapted by Bocorny and Welp (2021) to design EAP materials combining corpus and genre-based pedagogies. In this chapter, we introduced a step-by-step guide to help teachers to retrieve and integrate corpus data into materials designed for EAP writing courses through indirect DDL. Moreover, we provided explanations and descriptions of actions for each of the five first principles. Besides exemplifying those actions, we had in mind an EAP writing course for producing Health Sciences structured abstracts.

The COVID-19 pandemic is now over (or so we believe), and we are back to on-site classes. Nevertheless, we are glad to say that we genuinely believe we have all become more skilled and knowledgeable teachers.

Although we had a particular group of teachers in mind to produce this study, we believe that the insights it led to can be generalized. Even so, further studies could focus on work with a more significant sample of teachers, both from the secondary and tertiary levels. Above all, we expect this contribution will help to bridge the gap between corpus linguistics and EAP materials design.

Acknowledgement

For the shared route and mutual support, we wish to thank the CLA teachers for sticking together and supporting each other in a genuine case of scaffolding.

References

Aijmer, K. (2009). (Ed.) *Corpora and language teaching*. John Benjamins Publishing.

Anthony, L. (2003). *AntMover* (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2017). *AntGram* (Version 0.0.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2019). *AntCorGen* (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anthony, L. (2022a). International Perspectives on Corpus Technology for Language Learning - The University of Queensland Seminar Series). Addressing the challenges of data-driven learning through corpus tool design: An introduction to AntConc 4 [Video]. https://languages-cultures.uq.edu.au/event/session/8171

Anthony, L. (2022b). *AntConc* (Version 4.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Araújo, A. D. (1999). Uma análise de organização discursiva de resumos na área de Educação. Revista do GELNE – *Grupo de Estudos Linguísticos do Nordeste, 1*, 26-30.

Atkins, S., Clear, J. & Oslter, N. (1992). Corpus design criteria**.** *Literary and Linguistic Computing*, *7*, 1-16.

Berdanier, C. G. (2019). Genre maps as a method to visualize engineering writing and argumentation patterns. *Journal of Engineering Education*, *108*(3), 377-393.

Bhatia, V. K. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14 (3)*, 275-311.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D. & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. OKSEFJELL (Eds.). *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-190). BRILL.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Bocorny, A. E. P. & Welp, A. K. D. S. (2021). O desenho de tarefas pedagógicas para o ensino de inglês para fins acadêmicos: conquistas e desafios da Linguística de Corpus. *Revista de estudos da linguagem, 29(2),* 1589-1638.

Bondi, M. (2022). Comparable corpora in cross-cultural genre studies: Tools for the analysis of CSR reports. *Corpus Linguistics and Translation Tools for Digital Humanities: Research Methods and Applications*, 37-63.

Boulton, A. (2007). But where's the proof? The need for empirical evidence for data-driven learning. In *Proceedings of the BAAL Annual Conference 2007,* p. 13-16.

Boulton, A. (2021). Research in data-driven learning. *Beyond Concordance Lines: Corpora in language education*, *102*, 9-34.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language learning*, *67*(2), 348-393.

Brezina, V., Weill-Tessier, P. & McEnery, T. (2020). #LancsBox 5.x and 6.x [software]. http://corpora.lancs.ac.uk/lancsbox

Breyer, Y. (2011). *Corpora in Language Teaching and Learning. Potential, Evaluation, Challenges.* Peter Lang.

Can, S., Karabacak, E. & Qin, J. (2016). Structure of moves in research article abstracts in applied linguistics. *Publications*, *4(3)*, 23.

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*(4), 289-302.

Charles, M. (2013). English for academic purposes. *The handbook of English for specific purposes*, 137-153.

Charles, M. (2020). Combining genre analysis and corpus consultation in class: Using do-it-yourself corpora to explore the literature review. *Approaches to Specialized Genres*, 243-258.

Charles, M. & Frankenberg-Garcia, A. (2021). Introduction: Dichotomies and debates in corpora and ESP/EAP writing. In M. Charles, A. & Frankenberg-Garcia (Eds.). *Corpora in ESP/EAP Writing Instruction* (pp. 1-10). Routledge.

Cheng, W. (2010). What can a corpus tell us about language teaching?. In M. McCarthy & A. O'Keeffe (Eds.) *The Routledge handbook of corpus linguistics* (pp. 319-332). Routledge.

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.

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. New York, NY: Palgrave Macmillan.

Cotos, E., Haufman, S. & Link, S. (2017). A move/step model for methods sections: Demonstrating Rigour and Credibility. *English for Specific Purposes, 46*, 90-106.

Fletcher, W. H. (2012). *kfNgram* (Version 1.3.1). Retrieved from http://kwicfinder.com/kfNgram/kfNgramHelp.html

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International journal of corpus linguistics*, *14*(3), 393-417.

Flowerdew, L. (2012). *Corpus and Language Education*. Basingstoke: Palgrave Macmillan.

Flowerdew, L. (2013). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching*, *48*(1), 99-116.

Flowerdew, L. (2014). Corpus-based analyses in EAP. In J. Flowerdew & C. Candlin (Eds.). *Academic discourse* (pp. 105-124). Routledge.

Francis, W. N. (1992). Language Corpora BC. In Svartvik, J. [ed.] *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, Stockholm. Berlin/ New York, p. 17-32.

Frankenberg-Garcia, A., Lew, R., Rees, G. Roberts, J.C., Sharma, N. & Butcher, P. (2021). *ColloCaid* (around 30 thousand academic English collocations and examples of collocations in context curated from corpora of expert academic English), open access at http://www.collocaid.uk/

Freitas, A. L. P. & Bocorny, A. E. P. (2021). How to write medical abstracts? The rhetorical structure and phrases used in Epidemiology. *Brazilian Journal of Sexually Transmitted Diseases*, *33*, 1-6.

Gavioli, L (2005). *Exploring corpora for ESP learning* (pp. 1-176). Amsterdam: John Benjamins.

Gaspar, P. C., Santos, A. S. D. dos., Santana, L. B., Aragón, M. G., Machado, N. M. da S., López, M. A. A., Passos, M. R. L., Pereira, G. F. M. & Miranda, A. E. (2022). The fight against sexually transmitted infections cannot stop in the COVID-19 era: a brazilian experience in online training for sexually transmitted infections guidelines. *Brazilian Journal of Sexually Transmitted Diseases*, *34*. https://doi.org/10.5327/DST-2177-8264-20223404

Geluso, J. (2019). *Frequency, semantic, and functional characteristics of discontinuous formulaic language: A learner corpus study*. Master's dissertation, Iowa State University.

Gray, B. & Biber, D. (2013). Lexical frames in academic prose and conversation. *International journal of corpus linguistics*, *18*(1), 109-136.

Gray, B., Cotos, E. & Smith, J. (2020). Combining rhetorical move analysis with multi-dimensional analysis: Research writing across disciplines. *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*, 137-168.

Henderson, A. & Barr, R. (2010). Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research*, *2*(2), 245-264.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, *34*(8), 1091-1112.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–61.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *English Language Research Journal*, *4*, 27–45.

Karlsen, P. H. (2021). *Teaching and Learning English through Corpus-based Approaches in Norwegian Secondary Schools: Identifying Obstacles and a Way Forward*. Doctoral thesis. Inland Norway University of Applied Sciences.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for specific purposes, 24*(3), 269-292.

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), pp.1–35. DOI: http://doi.org/10.35360/njes.522

Kennedy, G. (1998). *An introduction to Corpus Linguistics*. New York: Longman.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004) The sketch engine. *Proceedings of the 11th EURALEX International Congress*: 105-116.

Le, T. N. P. & Harrington, M. (2015). Phraseology used to comment on results in the discussion section of applied linguistics quantitative research articles. *English for Specific Purposes*, *39*, 45-61.

Martínez, I. A. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of second language writing*, *14*(3), 174-190.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-Based Language Studies*. USA/Canada: Routledge.

McEnery, T. & Wilson, A. (1997). Teaching and Language Corpora. *ReCALL*, *9*(1), 5-14.

Meurer, J. L. (1997). Esboço de um modelo de produção de textos. *In* J. L. Meurer & D. Motta-Roth (Eds.). *Parâmetros de textualização*. (pp. 14- 27). Santa Maria: Editora da UFSM.

Moreno, A. I. & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, *50*, 40-63.

Motta-Roth, D. (1995). *Rhetorical Features and Disciplinary Cultures: A Genre-Based Study of Academic Book Review in Linguistics, Chemistry and Economics*. Doctoral thesis. Universidade Federal de Santa Catarina.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor & T. A. Upton, T. A. (Eds.). A*pplied Corpus Linguistics*, 239-250. Amsterdam: Rodopi.

O'Keeffe, A. (2022). *Data-driven learning and second language acquisition – it's time to connect.* [Video]. School of Languages and Cultures. https://languages-cultures.uq.edu.au/event/session/7987

O'Keeffe, A., McCarthy, M. & Carter, R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Pérez-Llantada, C. (2022). Online Data Articles: The Language of Intersubjective Stance in a Rhetorical Hybrid. *Written Communication*. https://doi.org/10.1177/07410883221087486

Pérez-Paredes, P. (2019). The pedagogic advantage of teenage corpora for secondary school learners. *Data-Driven Learning for the Next Generation*, 67-87.

Poole, R. (2020). "Corpus can be tricky": revisiting teacher attitudes towards corpus-aided language learning and teaching. *Computer Assisted Language Learning*, 1–22. doi:10.1080/09588221.2020.1825

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.

Römer, U. (2006). Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. *Zeitschrift für Anglistik und Amerikanistik, 54*(2), 121-134. https://doi.org/10.1515/zaa-2006-0204

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*, *3*(1), 95-119.

Santos, A. R. (1999). *Metodologia científica: a construção do conhecimento*. Rio de Janeiro: DP & A Editora.

Schneuwly, J. & Dolz, B. (2004). *Gêneros orais e escritos na escola.* Campinas: Mercado de Letras.

Scott, M. (2020). *WordSmith Tools version 8*. Stroud: Lexical Analysis Software.

Shepherd, T. M. (2009). O Estatuto da Linguística de *Corpus:* Metodologia ou Área da Linguística? *Matraga*, *16*(24), 150-172.

Sinclair, J. (1987). *Collins Cobuild English Language Dictionary: Helping Learners with Real English.* Heinle ELT.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. (2004a). 'Introduction'. *In* J. M. Sinclair (Ed.) *How to Use Corpora in Language Teaching* (pp. 1-13). Amsterdam and Philadelphia: John Benjamins.

Sinclair, J. (2004b). *Trust the text: Language, Corpus and Discourse.* London/ New York: Routledge.

Stevenson, H. A. & Harrison, J. E. (2009). Structured abstracts: Do they improve citation retrieval from dental journals?. *Journal of orthodontics*, *36*(1), 52-60.

Swales, J. (1981). *Aspects of article introductions.* Birmingham: Language Studies Unit. University of Aston [Aston ESP Research Reports 1].

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

Swales, J. (1994). The writing of research articles introduction. *Written Communication*, *4*(2), 175-191.

Swales, J. (2004). *Research genres: Exploration and applications.* Cambridge: Cambridge University Press.

Swales, J. & Feak, C. B. (2009). *Abstracts and the writing of abstracts* (Vol. 2). University of Michigan Press ELT.

Viana, V., Bocorny, A. & Sarmento, S. (2018). *Teaching English for Specific Purposes. ELT Development Series*. TESOL Press.

Welp, A. K., Didio, Á. R. & Finkler, B. (2019). Questões contemporâneas no cinema e na literatura: o desenho de uma sequência didática para o ensino de inglês como língua adicional. *BELT-Brazilian English Language Teaching Journal*, *10*(2), e35861-e35861.

Yoon, J. & Casal, J. E. (2020a). P-frames and rhetorical moves in applied linguistics conference abstracts. *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*, *95*, 282-305.

Yoon, J. & Casal, J. E. (2020b). Rhetorical structure, sequence, and variation: A step-driven move analysis of applied linguistics conference abstracts. *International Journal of Applied Linguistics*, *30*(3), 462-478.

Appendix I - Checklist for planning and designing an EAP course using a framework that combines corpus and genre-based pedagogies

Information about learners	Know learners' language proficiency level
	Know learner's level of instruction or position (e.g. undergraduate, graduate master, graduate doctor's, professor)
	Know discipline learner works with
	Know learners' needs
	Know learners' wants
	Know learners' expectations
Information about the course	Select the target genre
	Select the target section (may not apply)
	Select the target skill(s)
	Know how many and which disciplines (multiple or single) you will be working with
Planning the course	Set learning objectives
	Select methodology and approach
Select materials	Find existing materials
Design materials that are corpus-based, genre (section) and discipline specific	Find the target-genre rhetorical structure in the literature or describe it
	Decide which language features are worth working within the academic context in which the target genre is used and considering all the previously collected information
	Compile a genre (section) and discipline specific corpus
	Extract language data from the corpus
	Use said language data to design tasks, exercises, activities within the context of the target genre

Appendix II - Example of completed checklist for the course **Written Production of Health Sciences Structured Abstracts**

Information about learners	Language proficiency level	B2, C1
	Learner level of instruction or position (e.g. undergraduate, graduate master, graduate doctor's, professors)	Graduate students
	Discipline, specialty learners works with	Health sciences
Information about the course	Target genre	Structured abstracts
	Target section (may not apply)	Background and objectives, method, results, conclusion
	Target skill(s)	Written production
	Discipline (multiple or single)	Single discipline
Rhetorical structure of the target genre	Found in the literature or described by the teacher	Described by the teacher
Language feature(s) worth working within the context of the target genre	Lexical Frames	The first LF after the section name
Methodology	Combination of corpus and genre-based approaches	

Do-It-Yourself Corpora to Support SHAPE and STEM Research Paper Writing

Paula Tavares Pinto (Unesp)
Luciano Franco da Silva (Unesp)
Talita Serpa (Unesp)
Diva Cardoso de Camargo (Unesp)

Introduction

Writing research papers in English may be a challenge for newcomer authors at the beginning of their academic careers. For those who are non-native speakers of English and did not have the chance to use academic English with frequency it may be even harder. Most of the time these researchers are used to reading scientific papers, but do not have much experience in writing them.

Some of the scholars who have studied academic writing in depth are Swales and Feak (2004, 2009), Hyland (2004, 2014), Lee and Swales (2006), and Flowerdew (2010). Even though these authors have widely described the features of academic writing, there are some characteristics that may still not be as salient for novice researchers such as the use of academic collocations and lexical bundles. Some authors use word combinations that do not sound natural to their scientific community and this may impair their article acceptance. Some of the scholars who have pointed out the academic issues found in research papers of non-native speakers of English are Charles (2012), Howarth (2013), Chang and Swales (2014), Karpenko-Seccombe, (2020), Tavares-Pinto et al. (2021) and Pinto et al. (2021).

In this context, corpus linguistics has played an important role in providing a range of writing tools to help researchers from different fields to find language patterns in academic discourse that are recognized by their

peers. This happens because authors will rely on large collections of academic texts, hereafter, corpora, which can show them how their research community generally writes and the specific terminology and frequent patterns that can be rapidly identified and retrieved for writing purposes. This methodological approach can be used in different areas, such as Mathematics, Humanities and Biological areas. In order to do that, authors can use pre-compiled specialized corpora or compile their own collection of research papers published in high impact journals and use them as a Do-it-Yourself corpus (Vantarola, 2002; Maia, 2002; Frankenberg-Garcia et al., 2019; Carvalho et al., 2021).

According to Berber Sardinha (2010: 304), these linguistic patterns will show how co-occurring combinations are vital to the written discourse and how things are "said" and "organized" when structuring language. To the author, corpus linguistics

> […] shows that language is used in a patterned way (that is, in a way recognized as 'expected' or 'typical' by its users), with correlations between usage and context - different contexts are expressed in different ways, with their own usage probabilities, often quite specifically adjusted […] to the social, situational, speaking, historical period context. etc. […]. Therefore, through the use of corpora in teaching, we can bring this system to students more clearly than with contributions from other linguistic theories and methodologies. The nature of knowledge of a language changes with corpora research. 'Knowing a language' implies knowing how to say and write according to the conventions of specific varieties of the language (a specific genre or register in a given context); for this, it is necessary to know the lexicogrammar of the necessary and desired choices for that specific situation. In order to use lexicogrammar efficiently, it is necessary to know the probabilities of those choices, that is, the frequencies of the elements, their combinations and their frequencies[1] (Berber Sardinha, 2010: 304).

1 Original text: […] mostra que a linguagem é usada de modo padronizado (isto é, de modo reconhecido como 'esperado' ou 'típico' por seus usuários), com correlações entre uso e contexto - contextos diferentes são expressos de maneiras distintas, com

By using corpora, the writer will be able to observe the useful information according to his or her specific needs and will develop an autonomous process of learning that will lead him or her to mastering the academic English based on his interpretation of his or her peers' writing.

This chapter will bring a discussion on how specialized corpora can be explored by researchers who want to compile their own language database to help them write different sections of their own research papers. We will illustrate our proposal by taking examples from SHAPE disciplines, which involve Social Sciences Humanities, Arts for People and Economy, as well as STEM disciplines, which involve Science, Technology, Engineering, and Mathematics.

The next sections of this chapter are divided into the following topics: 2. Corpus Linguistics and Academic Writing; 3. AntCorGen for the compilation of SHAPE and STEM areas; 4. Analyses with Sketch Engine; 5. Building your Research paper with SHAPE Plos and STEM Plos corpus; 6. Discussion and 7. Final considerations.

---

suas próprias probabilidades de uso, muitas vezes ajustadas de modo bastante específico […] ao contexto social, situacional, falante, período histórico, etc. […] Assim, por meio de uso de *corpora* no ensino, podemos trazer aos alunos esse sistema de modo mais claro do que com aportes de outras teorias e metodologias da linguística. A natureza do conhecimento de uma língua se altera com a pesquisa em *corpora*. 'Saber uma língua' implica conhecer como dizer e escrever segundo as convenções de variedades específicas da língua (um gênero ou registro específico em um contexto determinado); para isso, é preciso conhecer a lexicogramática das escolhas necessárias e desejadas para aquela situação específica. Para usar a lexicogramática com eficiência, é necessário conhecer as probabilidades daquelas escolhas, isto é, as frequências dos elementos, suas combinatórias e as frequências destes (Berber Sardinha, 2010: 304).

Corpus Linguistics and Academic Writing

Corpora will help the authors in developing their pragmatic competencies such as the intercultural competency which will, according to Hurtado Albir (2001), help them in recognizing the contextual norms of a given text. Varantola (2003) also points out that the "proficiency" will depend on competence and practical skills that are combined to favor the cultural and linguistic decision-making process.

As we elevate corpora to the status of teaching and informational material, we allow the writer to concentrate on numerous possibilities of language variation and specialized language which will be discussed in this chapter. By using a bottom-up approach, the writers will also be able to observe different texts within the academic genre, depending on the kind of text they will be writing.

The use of corpus linguistics has been advocated by several scholars. In the case of Brazilian academia, Berber Sardinha (2003) had pointed out that university students and scholars should be able to have access to basic tools and infrastructure in order to explore corpora in class. Almost 20 years later, we have seen this advance in academia since more and more researchers have been using corpus linguistics tools to help them write their own texts. This possibility has recently been used at the São Paulo State University (Unesp) and at the Federal University of Rio Grande do Sul (UFRGS) where 127 researchers and English for Academic Purposes teachers worked in partnership to learn how to use corpora tools to write their own research papers and produce EAP teaching materials. The experience was described in detail in two publications by the British Council (Frankenberg-Garcia et al., 2019; Frankenberg-Garcia, 2020) and by Carvalho et al. (2021).

During the course, junior and senior researchers were introduced to corpus techniques and tools and were able to compile their own study corpora from high impact journals in their respective fields/disciplines. In this course, they learned how to use Sketch Engine (Kilgarriff, 2014) to explore academic language and see how key terms were used in specific contexts. Researchers and EAP teachers could help each other by analysing recurrent

language features and typical terminology in their DIY study corpora. There were specialists of Engineering, Agricultural Sciences, Humanities, Social Sciences and Health, among others. According to Carvalho et al. (2021):

> results showed that although scholars were familiar with the terminology of their own areas, the tool pointed out other possibilities of word combinations they had difficulty with, such as verbal collocations and the most common patterns of academic English if compared to Portuguese. At the same time, the English teachers who were participating in the workshops were inspired by the terminology and language to develop teaching activities for their own EAP students (Carvalho et al., 2021: 79).

To add the usefulness of corpora as a learning and translating material, Zanettin et al. (2003: 2) have stated that "(…) competent use of corpora and corpus analysis tools will enable students to become better language professionals in a working environment where computational facilities for processing text have the rule rather than the exception".

It is important to mention that writers will need to be trained on how to better explore corpora with appropriate tools; and they will also need to know how to interpret the information generated by those tools. By doing so, these writers will be using the *Data Driven Learning* approach (Johns, 1991), which shows concordance lines that will be displayed on the screen to the reader.

According to Varantola (2003), the use of corpora will provide two sets of skills to the writers related to: i) *corpus compilation* - criteria for corpus compilation, strategies to find relevant language pattern, access to reliable corpora, recognition of corpus compilation tools, integration of text processing and corpora processing tools; ii) *use of corpus information* - skills for deduction based on corpus information, use of pre-compiled corpora for translation retrieval, corpus assessment for translation decisions, new correlated skills for corpus management.

Regarding the search for specialized terminology, Bowker (1999) states that corpora will make it possible terminologists and language users to become aware of specificities of technical and scientific language.

The researcher points out that translators, when dealing with specialized texts, will be able to interact with the lexicon and terminology in different areas if using corpus tools and collections of specific texts. Therefore, when focusing on academic, technical and scientific writing, corpora may help researchers to compile glossaries that can be used in present and future works. Pearson (1996) supports this idea stating that corpus will enable the observation of domains and subdomains in the same areas. Also, Maia (2000) points out the importance of deepening the use of corpora for specific purposes and collection of vocabulary and observation of complex language when preparing teaching materials.

Formulaic language and its contributions to language studies

Corpora studies have shown that many language patterns are so recurrent among language users that they could be classified as pre-fabricated structures. The recurrence of pre-fabricated expressions in the language is explained by Sinclair (1991), through his idiom principle, in which he proposes that speakers do not simply choose random words to perform certain language functions, in fact, they seem to routinely use the same set of language combinations instead of creating new ones.

In this chapter, we chose the term formulaic language, coined by Wray (2002) to refer to the different types of semi-preconstructed language combinations. Sinclair (1991) and Wray (2002) argue that the human brain optimizes the processing of large amounts of data, through the repeated use of conventionalized language structures, which in turn reduces cognitive demands of on-line processing during language production and prevent speakers from becoming overloaded by decoding phrases and combinations they have never heard before. Because of this double advantage, the proper use of formulaic language is one of the central aspects for teaching and learning any language (Schmitt & Carter, 2004; Wray, 2002; Wray & Perkins, 2000). However, it is not so simple to define what is or is not formulaic in language (Granger & Paquot, 2008; Schmitt & Carter, 2004; Siyanova-Chanturia, 2015; Wray, 2002). Depending on the theoretical-methodological criteria, one can find dozens of terms for similar lexical

combinations, such as idioms (e.g. kick the bucket), collocations (e.g. fast food), lexical bundles (e.g. if you look at), among many others.

Nevertheless, Conrad et al. (2004) explain that, regardless of the name adopted, there are some characteristics that tend to be especially recurrent in the identification of formulaic language, such as fixedness; idiomaticity; frequency; length of sequence; completeness in syntax, semantics, or pragmatics; and intuitive recognition by the speaker of a language community.

The authors also explain that different types of formulaic language are identified depending on the priority these features receive. In other words, if the focus is idioms, the researcher is expected to prioritize certain characteristics that would not be interesting to identify collocations or lexical bundles, for example.

In the present study, we use the frequency-driven approach to find the most recurrent combinations in two DIY-study-corpora, which enables the semi-automatic extraction of massive amounts of linguistic data from a corpus, based on external criteria set by the researcher. Studies of recurrent combinations tend to converge towards similar goals, as evidenced by Conrad et al. (2004: 58):

> Our research questions in this approach are exploratory. We ask whether there are multi-word sequences that are used with high frequency in texts, whether different registers tend to use different sets of these sequences, and, if so, to what extent the bundles fulfill discourse functions and thus play an important part in the communicative repertoire of speakers and writers.

By exploring DIY corpora, researchers will also have a better view of discourse variation in different academic areas, therefore, when applying to academic writing, many studies have presented evidence of disciplinary variation based on corpus analyses (Bondi & Sanz, 2014; Gray, 2015; Hyland, 2012; Römer et al., 2020).

According to Becher and Trowler (2001), scientific knowledge is created from different disciplinary communities or tribes with particular interests, literature and conventions that shape how researchers see the world and interpret reality. Similarly, the concept of discipline is presented by

Hyland (2004, 2012) as a human institution where the creation of knowledge and use of language are influenced by personal and interpersonal factors from its members, as well as by institutional and sociocultural norms of the community in which they are part of. Considering that these investigations into disciplinary discourse have a great relevance, a member of a disciplinary community or novice research needs not only to demonstrate technical and theoretical competence in his field, but also know the linguistic conventions that create and maintain the cultural identity of its members (Becher & Trowler, 2001; Hyland, 2004, 2008).

The next section will discuss how two corpora were compiled in SHAPE and STEM disciplines.

AntCorGen for the compilation of SHAPE and STEM areas

AntCorGen (Anthony, 2019) is a tool used to quickly compile specialized corpora with research papers from the PLOS one platform. A tutorial of this tool was recorded by its creator in a short video[2]. Below we will talk about the compilation of SHAPE Plos and STEM Plos and their exploration for academic writing.

SHAPE

As previously mentioned, SHAPE disciplines stand for Social Sciences Humanities, Arts for People and Economy. All these disciplines and subareas can be found at PLOS, which is a nonprofit, open access multi-disciplinary publisher[3]. All areas of SHAPE can be easily accessed in AntCorGen and the researcher can choose the parts of research papers he wants to analyse. Since we wanted to have mostly written material, we selected the articles' abstracts, introduction, materials & methods, results & discussion and conclusions, as we can see in the figure below:

2 AntCorGen tutorial <https://www.youtube.com/watch?v=WrsIzE9to4o>. Access: Oct. 30th, 2021.

3 PLOS available at <https://plos.org/about/> Access: October 27th, 2021.

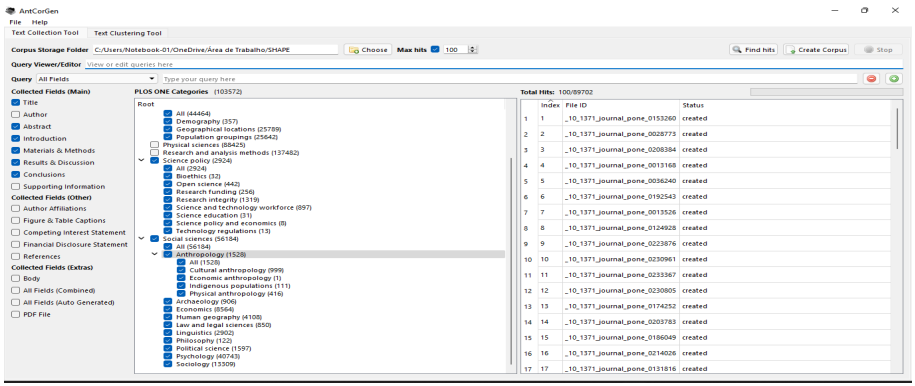


Figure 1. AntCorGen screen with part of SHAPE disciplines selected

We called this corpus SHAPE Plos and, since it was compiled for describing the process in this chapter, we set the maximum of 100 articles, but it is possible to have a much larger study corpus if desired. After this compilation we had a study corpus of 445,291 words to be explored.

STEM

STEM disciplines are related to both Biology and Hard Sciences. Although the figure below seems to have only Biology and Life Sciences, the actual list of disciplines selected was longer and we could include areas such as Math and Computer Sciences as well. In the same way, we selected 100 articles for STEM Plos corpus as shown in the figure below:

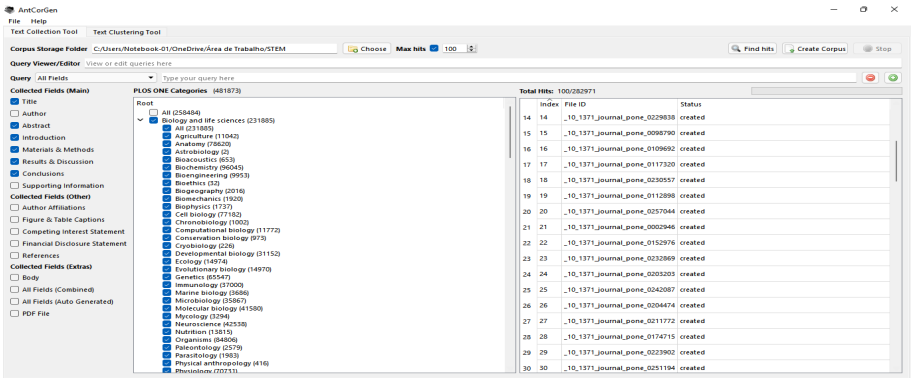


Figure 2. AntCorGen screen with part of STEM disciplines selected

After this compilation, we had a specialized corpus of STEM disciplines with a total number of 297,255 words to be observed and compared to the results from SHAPE Plos.

Analyses with Sketch Engine

We uploaded both corpora, SHAPE and STEM, to Sketch Engine (Kilgarriff, 2014) so we would be able to observe the frequent adjectives and verbs in each broad area and see the similarities and differences between them. We could also generate concordance lines with search words, terms and phrases that can be used by researchers to explore and observe how international researchers in their area have been writing different sections of their research papers.

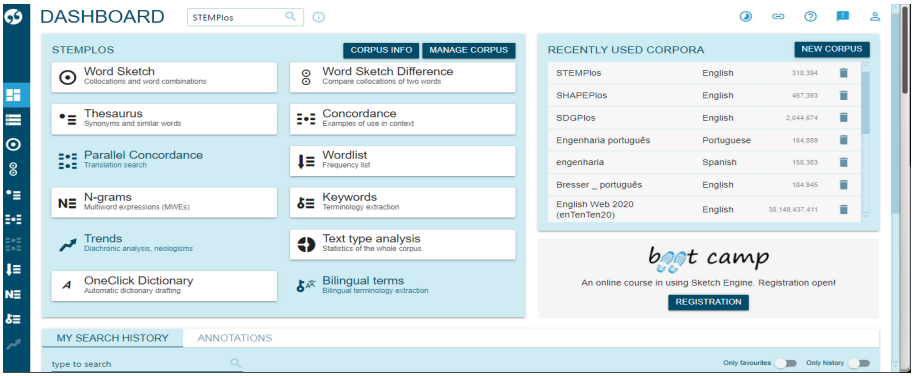


Figure 3. Sketch Engine dashboard with its tools and STEM Plos and SHAPE Plos as main corpora

The main tools we have been using to explore both corpora are *Wordlist*, *Concordance* and *Word Sketch*. *Wordlist* will list all words from a study corpus in order of frequency, from the most frequent to the least; *Concordance* will generate concordance lines with a search word in context that can be expanded if the researcher wishes to do so; and *WordSketch* will show a search word with co-occurrent categories such as modifiers, verbs with the search word as object or subject, and frequent adjectives and adverbs used with it. We are going to describe these searches in the next section.

Adjectives and verbs in SHAPE Plos and STEM Plos

During the Academic Masterclasses (Frankenberg-Garcia et al., 2019), one of the main searches was on how to use appropriate adjectives to bring more emphasis to the articles; therefore, instructors taught researchers and EAP teachers how to look for adjectives within the corpus wordlist. Below we bring the list of the twenty most frequent adjectives and modifiers used by international researchers in SHAPE Plos corpus.

SHAPE Plos				STEM Plos			
Item	**Freq.**	**Item**	**Freq.**	**Item**	**Freq.**	**Item**	**Freq.**
high	1138	large	439	such	549	standard	251
other	953	**linguistic**	404	different	509	open	242
such	786	small	369	other	508	many	230
social	785	**female**	337	available	443	low	228
different	699	likely	327	new	318	**mobile**	220
low	646	**sound**	326	high	280	several	203
more	599	**religious**	306	large	273	**medical**	192
significant	523	similar	304	same	271	**multiple**	186
same	494	**cultural**	298	more	258	specific	185
first	451	**lexical**	298	standard	251	good	183

Table 1. Twenty most first adjectives and modifiers in SHAPE Plos and STEM Plos corpora

The adjectives and modifiers present in the lists of SHAPE disciplines, on the left columns, and STEM disciplines, on the right columns in Table 1 show specific adjectives and modifiers to each area in bold and common adjectives and modifiers to both areas which have been underlined. Authors can choose specific items as shown in the examples that we are going to present, where we see concordance lines with "social" and "linguistic" from SHAPE Plos and "standard" and "mobile" from STEM Plos. The underlined words are the ones being modified by the search words:

1. Although a great deal of attention has been paid to how conspiracy theories circulate on **social** media, and the deleterious effect that they, and their factual counterpart conspiracies, have on political institutions, there has been little computational work done on describing their narrative structures. [SHAPE Plos]

2. Public pension insurance has become a major form of **social** protection around the world.[SHAPE Plos]

3. Feature stability, time and tempo of change, and the role of genealogy versus a reality in creating **linguistic** diversity are important issues in current computational research on linguistic typology. [SHAPE Plos]

4. The database is pre-prepared for statistical and phylogenetic analyses and contains both linguistic typological data from languages spanning over four millennia, and **linguistic** metadata concerning geographic location, time period, and reliability of sources. [SHAPE Plos]

We did the same search for adjectives in STEM Plos, which can be seen below:

5. Other implemented functions are focused on the quality control of the fitted **standard** curve: detection of outliers, estimation of the confidence or prediction interval, and estimation of summary statistics. [STEM Plos]

6. On the other hand, if there are not enough dilutions of the **standard** samples, the extra standard sample using the background information will influence data similarly to an outlier due to the fact that the standard points have not reached the lower asymptote.[STEM Plos]

7. Our data allow a fleeting glimpse into the future, where **mobile** health will not replace the doctor-patient relationship, but will hopefully help to establish more effective and efficient treatment and accelerate e-health strategies. [STEM Plos]

8. During investigations of crimes involving mobile devices, there is usually some accumulation or retention of data on the device that will need to be identified, preserved, analyzed and presented in a court of law–a process known as digital or **mobile** forensics (also known as cyber forensics). [STEM Plos]

Some excerpts turned a noun into a complex expression, such as "social media", "standard curve", and "mobile forensics", others simply qualified a noun, such as "social protection", "linguistic diversity' and "standard

curve". More than examples of context, the search for adjectives and modifiers will bring several combinations that are frequently used with a search word which can bring strength to a text and can be used as an inspiration to authors in different areas.

Some adjectives were used in both SHAPE Plos and STEM Plos, such as "high". One of the advantages of using corpus tools is that you can quickly search for examples in both corpora, which will show how the same adjective is being used in different articles, as it is shown below:

9. Although, productivity is maximized by the combination of high wages and low labor input, **high** productivity cities show invariably high wages and high levels of employment relative to their size expectation.[SHAPE Plos]

10. Both logistic regression and PSM models revealed that early marriage decreased the chances of completing the first cycle of **high** school. [SHAPE Plos]

11. We also find that the effect of ICT use on economic growth is higher in **high** income group rather than other groups. [STEM Plos]

12. The framework employs features of centralized monitoring, **high** availability and on demand access services of computational clouds for computational offloading. [STEM Plos]

In examples 9, 11 and 12 we see the adjective "high" used to intensify the nouns that are accompanied by it, the only exception being "high school", that is a complex term.

If an author wants to find other adjectives that can be used as a synonym or that are present in the same contexts, such as antonyms, he can use the Thesaurus option in Sketch Engine. We looked for options in SHAPE Plos and found, in order of frequency, "great", "large", "overall" and "positive", at the same time, we also looked for other options for "high" in STEM Plos and found "overall", "maximum", "large" and "great".

In the same Academic Masterclasses we encouraged researchers and postgraduate students of SHAPE and STEM disciplines to do a search on the most frequent verbs that had been used in their areas of research. In a similar way, we show a list of the most frequent verbs in Table 2:

SHAPE Plos				STEM Plos			
Item	**Freq.**	**Item**	**Freq.**	**Item**	**Freq.**	**Item**	**Freq.**
be	15020	provide	449	be	10879	develop	349
have	2865	suggest	436	use	2057	make	331
use	1563	compare	424	have	1637	propose	296
do	919	report	417	provide	592	create	280
show	910	associate	409	include	486	present	256
include	720	consider	405	show	480	follow	255
find	688	indicate	375	base	394	define	250
see	588	follow	369	do	373	find	243
give	483	make	369	require	368	describe	242
base	465	increase	345	allow	368	identify	237

Table 2. Twenty most frequent verbs and modifiers in SHAPE Plos and STEM Plos corpora

When we compare the twenty most frequent verbs in SHAPE Plos and in STEM Plos, we find eleven verbs used in both areas, some of them are "show", "include" and "provide", which are academic content verbs that are common to all areas, but will be used in specific contexts, as we can see in the examples 1 to 6:

1. Results **show** that the amount of fine does not impact tax payments, whereas participants' beliefs regarding tax authority's power significantly shape compliance decisions. [SHAPE Plos]

2. Detail results that **show** how tally the simulation results and the analytical results in both abstract and graphical forms and some scientific justifications for these have been documented and discussed. [STEM Plos]

3. These effects **include** stress regularity and stress consistency, both of which have been especially important in studies of word recognition and reading aloud in Italian. [SHAPE Plos]

4. A systematic framework and associated workflow **include** cloud service filtration, solution generation, evaluation, and selection of public cloud services. [STEM Plos]

5. It would be difficult to **provide** a comprehensive explanation for this result. [SHAPE Plos]

6. The evaluation of all network breakups can **provide** transportation planners and administrators with plenty of data for further statistical analyses. [STEM Plos]

When we look for verbs that are specific to any of both areas, we find only one verb that could be considered from SHAPE areas, which is the verb to "see". To illustrate that use, we bring some concordance lines in examples 7 to 9:

7. We also **see** internal fluctuations in the use of this style during this campaign. [SHAPE Plos]

8. Those who believe that their own religious group is something special tend to **see** extremism as an opportunity to assert their own group interests. [SHAPE Plos]

9. Other than education, for social participation we **see** that disability characteristics, motivation, and knowledge of the system are important for explaining the education gradient. [SHAPE Plos]

We also find verbs in both columns that, although not being present in both columns, could be used in SHAPE and STEM, such as "compare" and "associate", from SHAPE list, and "allow" and "propose", which are part of the STEM list.

Clusters, which are recurrent groups of words, can also help an author quickly identify features of academic writing. In the next subsection we present this discussion.

Clusters in SHAPE Plos and STEM Plos corpora

Another tool from Sketch Engine that can be used in search of clusters or lexical bundles is the one called *n-gram*. Table 3, below, shows the twenty most recurrent *n-grams* in the introduction section from the papers in the study corpora SHAPE Plos and STEM Plos. The *n-grams* in bold indicate that the sequence was recurrent in both corpora.

SHAPE PLOS	Normalized frequency (x100.000)	STEM PLOS	Normalized frequency (x100.000)
the number of	69	**the number of**	58
as well as	61	**in order to**	48
number of children	35	**as well as**	46
more likely to	34	**based on the**	36
based on the	31	**one of the**	30
in terms of	29	**the use of**	30
in order to	26	can be used	27
the effect of	26	the accuracy of	23
in this study	24	**due to the**	22
the relationship between	21	be used to	21
due to the	21	according to the	21
one of the	20	in this paper	21
there is a	20	of the data	20
the fact that	19	**on the other**	19
on the other	19	the development of	19

a number of	18	a set of	19
the distribution of	17	that can be	18
the present study	17	**on the other hand**	18
on the other hand	17	in addition to	17
the use of	17	Part of the	16

Table 3. Clusters in SHAPE Plos and STEM Plos corpora

As it can be seen, the sequence *as well as* was the most recurrent in the introduction section in both study subcorpora and it was commonly used to structure the discourse by adding new elements to the text, as shown in examples 1 and 2:

1. Violent and delinquent behaviour patterns, **as well as** associated attitudes, can also manifest themselves in various forms of extremism. [SHAPE Plos]

2. Healthcare provision via wearable devices brought changes in treatment and examination of patients, **as well as** research and development in different areas. [STEM Plos]

Other recurring elements of textual cohesion in the introductory section in both study corpora were the n-grams *on the one hand* and *on the other hand*, illustrated below, used to express contrast between the ideas and elements in the text, as shown in examples 3 and 4:

3. Loneliness at work is such a possible mediator: **on the one hand** there is a potential association between working temporarily and loneliness at work, on the other hand there are indications of a negative association between loneliness at work and job satisfaction. [SHAPE Plos]

4. Pharmacokinetics is the study of what the body does to a drug including processes from drug absorption to excretion. **On the other hand**, pharmacodynamics focuses on the effects of drugs on organisms. [STEM Plos]

Another discourse function expressed by the extracted *n-grams* was the limitation of research conditions expressed by the clusters. Although

this function was found in both subcorpora, the data indicates that the authors of SHAPE and STEM domains use different sequences for this function such as *in terms of*, *the relationship between* and *based on the*, as illustrated in examples 5 to 9:

5. In this study, we aim to refine the analysis **in terms of** the Liberal versus the Individual views [SHAPE Plos]

6. The latter type of news effects has been studied mainly **in terms of** news on the internet, rather than television. [SHAPE Plos]

7.In the present research, we investigate **the relationship between** linguistic cohesion and real-world action in times of social conflict and unrest. [SHAPE Plos]

8. We thus introduce a simple but practical measure evaluating network disintegration **based on the** overall number of people isolated from the primary network. [STEM Plos]

9**. Based on the** employed cryptographic mechanism, Lu et al. [6] distinguished the privacy-preserving authentication scheme of VANETs into five categories. [STEM Plos]

In the previous sections we discussed how the search for content words and lexical bundles can help writers use a more specific and elaborated language in their articles. In the following sections, we will discuss how researchers may access academic phrases by carrying out a search in concordance lines that will help them write different sections of their research papers.

Building your Research paper with SHAPE Plos and STEM Plos corpus

If a researcher wants to have examples of research papers in SHAPE and STEM disciplines, they can search for common expressions in the corpus. In our case, we have divided both subcorpora into research sections that are usually found in research articles. Based on Karpenko-Seccombe (2020), we are going to discuss how researchers can use their own specialized corpora for writing their research papers. The search we are going to

propose is similar to what is found in the Manchester Phrasebank (Morley, 2014), where it is possible to observe frequent phrases in different parts of an article. However, different from the Manchester Phrasebank, where phrases of all areas may be seen, the advantage of the search in a specialized corpus is that the researcher will be able to read more contexts about their own areas.

Researchers who read concordance lines can do it similarly to reading a dictionary, where they will find several examples of a search word or expression and they will select the one that better suits their own texts. Therefore, there will be a combination of a fast search aided by the tool, and human selection of the best examples which will be done by researchers.

In the following sections, authors will find useful strategies for searching for contexts in the sections of introduction, materials and methods, discussion and conclusion.

Writing the Introduction Section

According to Swales and Feak (2009), a research paper introduction typically contains three main steps or *moves*: a) establishing the area of research, where the authors will show the importance of a field and introduce previous research in the area; b) establishing a gap in the knowledge or problem to be solved, and c) presenting the paper, i.e., identifying objectives, introducing expected outcomes and describing the structure of the work. In order to explore introductions in SHAPE Plos and STEM Plos corpora, we searched for concordance lines with the query phrase "this paper" and selected some of the lines to be used as examples here:

1. **This paper** attempts to fill the gap of existing research concerning the link between public pension and fertility. [SHAPE Plos]

2. In **this paper**, we perform a comprehensive survey of the worldwide linguistic landscape as emerging from mining the Twitter microblogging platform. [SHAPE Plos]

3. In **this paper**, we are interested in measuring linguistic regularities both at the level of word structure and at the level of word order. [SHAPE Plos]

4. **This paper** explores the ways abortion attitudes intersect with causal beliefs about gender categories, within the unique social context of a national referendum held to legalise abortion in the Republic of Ireland. [SHAPE Plos]

5. In **this paper**, we introduce a novel mobile application called "Medikamentenplan" ("Medication Plan"), which was developed to support medication compliance and vital sign documentation. [STEM Plos]

6. In **this paper**, we propose a concise, improved and effective privacy framework for wearable device manufacturers, as well as application developers, capable of providing greater privacy and security to the wearable device owners. [STEM Plos]

7. **This paper** innovatively proposes countermeasures to improve the innovation of e-commerce practitioners in rural areas. [STEM Plos]

8. The objective of **this paper** is to outline our approach of establishing and implementing this IT infrastructure. [STEM Plos]

We can see that authors from SHAPE and STEM use similar strategies to introduce their research papers. In 1, 4 and 7, authors used the structure *This paper + [adverb] + verb (infinitive)*. In examples 2, 3, 5 and 6, authors opted to use *In this paper + we + verb (infinitive)*. Finally, in example 8, the author preferred to introduce his paper by using the structure *The objective of this paper is + to + verb (infinitive)*.

We can see a pattern in the previous examples that can be used in a more confident way by researchers of SHAPE and STEM.

Writing the Materials and Methods Section

According to McCombes (2019), in the methodology section the authors will explain what they did and how they did it. By doing so, other researchers will be able to evaluate the reliability and the validity of a research. In this section, authors will discuss the type of research they carried out, and how they collected and analysed the data. They will also include the tools and materials of the research. This section is usually written in the past tense.

Similarly to the previous section, by consulting concordance lines a researcher will have access to the writing of different authors who have described their methods in SHAPE and STEM disciplines. Below you will find eight concordance lines describing the methods and methodological procedures that can be used as examples to writing this section:

1. Recent advances in data-driven **methods** of embedding words and phrases into a multidimensional vector space such that their Euclidean distances have correlations with their semantic similarity have made it possible to assign a quantitative measure to the similarity metric. [SHAPE Plos]

2. This **method** provides a second ranking of headwords including non-named entities. [SHAPE Plos]

3. The **methods** compared are: Cysouw and colleagues consider the consistency of the cross-linguistic distribution of an individual feature with the pattern generated by multiple features, and they propose three quantifications of this measure based on Mantel's correlation, a coherence and a rank method (…) [SHAPE Plos]

4. There are several well-established **methods** for combining significance (p-value) and effect size information from independent tests of the same null hypothesis, especially developed for meta-analyzes, such as:Fisher's classic method [45], and the more recentZ-transform [46], but a priori they are not appropriate to our case due to the mentioned non-independence. [SHAPE Plos]

5. The above-described method resulted in better recognition of confluent colonies than **methods** employing binary thresholding and segmentation (using, e.g., watershed separation), which we tried as alternatives.
[STEM Plos]

6. In this study we aimed at reproducing the results from 11 PLOS ONE papers dealing with statistical **methods** for longitudinal data. [STEM Plos]

7. In this section, we introduce our experimental **methods**, which include definitions, attack strategies and benchmark networks. [STEM Plos]

8. The most common issue was that papers did not provide enough detail about the **methods** used (e.g. model type was mentioned but no detailed model specifications, for details see Table 4). [STEM Plos]

In examples 1, 4, 6 and 7 we see the structure *modifier/adjective + methods* which can show a range of possibilities for a reader to select the one that can be used in their own text. Examples 2 and 3 show the structure *This method + verb* in the active voice, bringing "method" as the agent of an action. In examples 5 and 8 we have *method + present/past participle*. We have three different ways of describing our methodology which are used in both SHAPE and STEM, that can be used by other researchers.

Writing the Discussion and Conclusions Sections

In "Discussion" and "Conclusions sections", authors will talk about their achievements and will conclude by: a) highlighting the significance of the results; b) comparing their results with previous research; c) emphasising the novelty and contribution of their research or d) suggesting treating results with caution. One way of knowing how researchers write their Discussion and Conclusions sections is by searching for the keyword "Results", in both subcorpora, as we have done below:

1. The **results** show that, in the pre-period of 2010, women in the NRPS group have more children and are more likely to have a second child than those without NRPS coverage, while there is no significant difference between treatment and control groups in the post-period of 2014. [SHAPE Plos]

2. These **results** suggest that a post-treatment effect on women's fertility outcomes may occur when they had participated in the pension scheme. [SHAPE Plos]

3. The **results** demonstrate a noteworthy extension of the common support between the treated and control groups, implying that the overall distributions of the conditional probability to participate in the NRPS are similar between the two groups. [SHAPE Plos]

4. The **results** show that while some variables are significantly different between the unmatched treated and control group, the differences between the two groups for all covariates are no longer significant after matching. [SHAPE Plos]

5. The **results** for both Chromeleon and HappyTools show a higher percentage of Fab-glycosylation in ACPA samples than IgG samples, with the values reported by ThermoFisher Chromeleon and HappyTools showing a significant correlation (Fig 3 and S5–S7 Tables). [STEM Plos]

6. The **results** of the present study can be compared directly to our previous study that focused on the accuracy of the GPS60 for the detection of bouts of walking and resting [15]. [STEM Plos]

7. The algorithm then omits all **results** related to the combinations of links containing at least one of the marked links. [STEM Plos]

8. The **results** show that the relative Aps reported by HappyTools are comparable to both Waters Empower and ThermoFisher Chromeleon (Fig 2 and S3 Table). [STEM Plos]

Most of the examples shown above follow the structure presented in Swales and Feak (2009), which are *The results show/ suggest/ demonstrate + that*. In examples 5 and 6 we see another structure, which is *The results for/of + object + verb*. Finally, example 7 brings "results" as an object of a sentence.

Discussion

In this chapter we discussed the advantages of compiling specialized corpora in the areas of SHAPE and STEM, which can be explored by researchers in different areas with the aid of corpus tools, such as Sketch Engine. By using this set of tools, researchers can quickly access the specific terminology in their own areas as well as select the lexicon that will best suit their own writing. By searching for specific vocabulary with WordList, WordSketch and Concordance lines, it is possible to observe frequent adjectives to each area, such as "social" in SHAPE and "standard" in STEM, as well as observe that "high" is one of the most frequent adjectives in both areas, however, it is used in specific contexts for each area such as in "high productivity cities" in SHAPE and "high income" in STEM. On the other hand, verbs did not show very specific use since the lists of frequent verbs are very similar in SHAPE and STEM. The only verb that was present in the list of twenty most frequent ones in SHAPE that was not frequent in STEM is the verb to "see".

Several of the most recurrent *n-grams* found in the introduction sections are text-oriented (Hyland, 2008: 13), which means they are concerned with the organization of the text and its meaning as a message or argument. Some examples of text-oriented n-grams are: *as well as, in addition to, on the other* [*hand*]; these sequences are important to signal logical relationship between the ideas presented and maintaining logical cohesion.

Following the findings of Swales and Feak (2004, 2009) and Karpenko-Seccombe (2020), another important aspect discussed in this paper was the use of similar language structures in each research section. The examples previously presented show how authors keep the same way for introducing their papers (*This paper aims …*), writing their methodology (*method + present/past participle*), discussing their findings and conclusions (*The results show/ suggest/ demonstrate + that*). Taking that into account we can infer that these structures provide safe ground to non-native speakers of English and novice researchers to "walk on" and to use in their own research papers in order to be accepted by their discourse

communities which will include peer reviewers and internationally recognized researchers.

The last aspect we would like to mention is that although we have used Sketch Engine to explore SHAPE and STEM corpora to write our own paper, there are similar tools that can be used by researchers, such as AntConc (Anthony, 2005) and LexTutor (Cobb, n.d.).

Final Considerations

In this chapter we presented an overview on how to compile specialized corpora in SHAPE and STEM with the AntCorGen tool and how researchers can use those corpora to access the academic language used by their peers. By doing so, researchers will confirm or refute ways of presenting their studies according to each research paper section, as well as the best way of describing their methodological approach, and call attention to their studies contribution. We hope this chapter may inspire research teams to start building their own language database that can be used by future members and can be constantly updated.

References

Anthony, L. (2005, July). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In IPCC 2005. *Proceedings. International Professional Communication Conference*, 2005. (pp. 729-737). IEEE.

Anthony, L. (2019). AntCorGen (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Becher, T. & Trowler, P. R. (2001). *Academic Tribes and Territories* (2nd ed.). SRHE.

Berber Sardinha, T. (2003). Uso de corpora na formação de tradutores. *Delta: documentação de estudos em lingüística teórica e aplicada*, *19*, 43-70.

Berber Sardinha, T. (2010). Como usar a linguística de corpus no ensino de língua estrangeira–por uma linguística de corpus educacional brasileira. *Corpora no ensino de línguas estrangeiras*, 293-348.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405.

Bondi, M. & Sanz, R. L. (2014). *Abstracts in Academic Discourse: Variation and Change* (1st ed.). Peter Lang.

Bowker, L. (1999). Exploring the potential of corpora for raising language awareness in student translators. *Language awareness*, *8*(3-4), 160-173.

Carvalho, C. T., Laranja, L. A. N. & Pinto, P. T. (2021). DIY Corpora: o que são e para quem são?. *Tradterm*, *37*(1), 64-87. https://doi.org/10.11606/issn.2317-9511.v37p64-87

Chang, Y. Y. & Swales, J. M. (2014). Informal elements in English academic writing: threats or opportunities for advanced non-native speakers?. In *Writing: Texts, processes and practices* (pp. 145-167). Routledge.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*(2), 93-102.

Cobb, T. (n.d.). *Range for texts v.3* [computer program]. Retrieved from <http://www.lextutor.ca/> at 19 november, 2021.

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe; M. McCarthy (Eds.) *The Routledge handbook of corpus linguistics*, 444-457.

Frankenberg-Garcia, A., Bocorny, A.E.P., Tavares-Pinto, P. & Sarmento, S. (2019) Supporting the Internationalization of Brazilian Research. *Workshops delivered at the Federal University of Rio Grande do Sul and at São Paulo State University*, Porto Alegre and São José do Rio Preto, April-June 2019.

Frankenberg-Garcia, A. (2020). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, v. 53, n. 1, 29-43.

Granger, S., Paquot, M. (2008). Disentangling the phraseological web. *Phraseology: An interdisciplinary perspective*, 27-49.

Gray, B. (2015). On the complexity of academic writing: Disciplinary variation and structural complexity. In V. Cortes & E. Csomay (Eds.), *Corpus-based Research in Applied Linguistics : Studies in Honor of Doug Biber* (1st ed., pp. 49–78). John Benjamins Publishing Company.

Howarth, P. A. (2013). *Phraseology in English academic writing*. Max Niemeyer Verlag.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press.

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

Hyland, K. (2012). Disciplinary Differences: Language Variation in Academic Discourses. In K. Hyland & M. Bondi (Eds.), *Academic Discourse Across Disciplines* (1st ed., pp. 17–45). Peter Lang**.**

Hyland, K. (2014). *Disciplinary discourses: Writer stance in research articles*. Routledge. In H. Candlin, & K. Hyland (Eds.), Writing: Texts, Processes and Practices Hyland, K. (2014). Disciplinary discourses: Writer stance in research articles. In H. Candlin, & K. Hyland (Eds.), Writing: Texts, Processes and Practices. Routledge. pp. 99-121.

Hurtado Albir, A. (2001). *Traducción y traductología. Introducción a la traductología.* Cátedra.

Johns, T. F. (1991). Should You Be Persuaded: Two Examples of Data-Driven Learning Materials. *English Language Research Journal*, No. 4, 1-16.

Karpenko-Seccombe, T. (2020). *Academic writing with corpora: A resource book for data-driven learning*. Routledge.

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

Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, *25*(1), 56-75.

Maia, B. (2000). Making corpora – a learning process. In: Bernardini, S. & Zanettin, F (eds). *I corpora nella didattica della traduzione.* Bologna: CLUEB. 47-6.

Maia, B. (2002). Do-it-yourself, disposable, specialised mini corpora–where next? Reflections on teaching translation and terminology through corpora. *Cadernos de Tradução*, *1*(9), 221-235.

McCombes, S. How to write a research methodology (2019). Available at <https://www.scribbr.com/dissertation/methodology/> Access on November 9th, 2020.

Morley, J. (2014). *Academic phrasebank*. Manchester: University of Manchester.

Pearson, J. (1996). Electronic texts and concordances in the translation classroom. *TEANGA: The Irish Yearbook of Applied Linguistics*, *16*, 85-95.

Pinto, P. T., de Camargo, D. C., Serpa, T. & da Silva, L. F. (2021) Analysing the behaviour of academic collocations in a corpus of research-papers: a data-driven study/ Analisando o comportamento de colocações acadêmicas em um corpus de artigos científicos: um estudo dirigido por dados. *Revista de Estudos da Linguagem*, *29*(2), 1229-1252.

Römer, U., Cortes, V. & Friginal, E. (2020). *Advances in corpus-based research on academic writing: Effects on discipline, register, and writer expertise* (1st ed.). John Benjamins Publishing Company.

Schmitt, N. & Carter, R. (2004). Formulaic sequences in action. *Formulaic sequences: Acquisition, processing and use*, 1-22.

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* Oxford: Oxford.

Siyanova-Chanturia, A. (2015). On the 'holistic'nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2), 285-301.

Swales, J. M. & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). Ann Arbor, MI: University of Michigan Press.

Swales, J. M. & Feak, C. B. (2009). *Abstracts and the writing of abstracts* (Vol. 2). University of Michigan Press ELT.

Tavares-Pinto, P., Rees, G. & Frankenberg-Garcia, A. (2021). Identifying collocation issues in English L2 research article writing. Charles, Maggie; Frankenberg-Garcia, Ana. *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis. 01ed. London: Rouledge*, 01-20.

Varantola, K. (2003). Translators and disposable corpora. *Corpora in translator education*, 55-70.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press, 110 Midland Ave., Port Chester, NY 10573-4930 (45 British pounds).

Wray, A. & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1-28.

Zanettin, F., Bernardini, S. & Stewart, D. (ed.). (2003) *Corpora in translator education*. Manchester: St. Jerome.

Creating a local learner corpus: Insights on project design and data analysis from the pilot phase

Sandra Zappa-Hollman (UBC-CA)
Alfredo Afonso Ferreira (UBC-CA)
Greta Perris (UBC-CA)
Simone Sarmento (UFRGS)
Marine Laísa Matte (UFRGS)
Laura Baumvol (UBC-CA)

Introduction

A learner corpus (LC) is a principled collection of texts produced by additional language learners. These texts are collected and systematically organized electronically to allow for a range of teaching and research applications. Learner corpora have been typically created by academics or publishers for so-called "delayed pedagogical use" (i.e., not necessarily for the immediate benefits of those students sharing their writing samples), as well as for research purposes; that is, for contributing to theorization in additional language acquisition and applied linguistics through identification of patterns in learner language. More recently, however, a growing number of LCs have been created locally by researcher-practitioners for "immediate pedagogical use" in their specific institutional contexts (Granger, 2009, 2015), leading to data-driven enhancements in curriculum development, teaching, and learning.

The LC project we report on here was designed to systematically collect and access large samples of our students' writing for relatively immediate pedagogical application. Over time, this resource is meant to better track writing development within and across student cohorts and identify patterns of variation at larger scales such as across disciplines, language background of learners, and instructional programs. This scope of interest

across teaching and research is indicative of the close relationship between them in data-based learning. In addition to helping us systematize access to student texts for research purposes, our LC is also designed to inform curriculum development and instructional practice.

When we embarked on this project, our team represented a range of expertise and background knowledge that enabled us to envision the overall objectives and structure of our LC. Yet it was evident early on that creating a successful local LC would require effort and a steep learning curve. This chapter reports on some of the key choices we made as we designed and implemented the pilot phase of the LC project. Some challenges we had to overcome and important considerations we made in relation to technological and logistical aspects. And to illustrate the potential benefits to teaching and research in our context of even a small dataset from the pilot phase of the project, we also present the results of an analysis of comparative discourse in student expository writing. We close the chapter with reflections synthesizing what we have learned from the pilot phase and outline on our following steps.

The VanCor Project

The Vantage College Corpus of Student Texts Across Disciplines (henceforth, VanCor) project that to create a systematic and searchable online repository of student written assignments. VanCor is conceived as a resource for faculty at Vantage College (VC) in The University of British Columbia (UBC) have easy access to written assignments that students engage in across a range of disciplines in first year programs. VanCor has the potential to be relevant for research, data-driven curriculum development, instructional materials development, and program evaluation purposes.

Institutional Context

Launched in 2014, VC is a unit at UBC that offers first year programming for international English as an additional language (EAL) speakers whose proficiency is slightly below the university's English language admission standards for direct entry. At the time of data collection, three program options were available: first-year Bachelor of Arts, first-year Bachelor of Engineering, and first-year Bachelor of Science. Program faculty include a team of English For Academic Purposes (EAP) instructors who work with disciplinary faculty seconded to VC from their respective departments in Arts, Engineering, and Science.

VC offers instructional programming tailored to support of students' transition into the second-year of their bachelor's degree at UBC. VC programs are characterized by a cohort-based model and standard timetables, providing a coordinated curriculum that includes content-focused and language-focused[1] credit-bearing courses. Thus, alongside their program-specific courses, students receive general EAP and discipline-specific English instruction. After successfully finishing their first year at VC, students continue as second-year students in their respective faculties. The program expands the usual two academic terms of first year to three academic terms, totaling 11 months of instruction. This time extension accommodates the required disciplinary courses in the respective programs of study as well as VC-specific programming aimed at scaffolding students' linguistic, cognitive, and skills development as apprentice multilingual scholars.

The custom-designed programming includes an introductory research methods course with an application component that engages students in a small group research project they eventually present at an annual

1 VC uses an integrated language and content approach which views the learning of language and subject area knowledge as inseparable and mutually constitutive. We use "content-focused" and "language-focused" as shorthand to refer to what otherwise are also referred to in the literature as "subject, or disciplinary" courses versus "language" courses. Yet we view both types of courses as involving both content as well as language. To try and foreground this relationship between language and content, we classify these as courses that place an emphasis or focus in either of the two, based on what most course learning outcomes stipulate.

student-led capstone conference. To explicitly support their academic (general as well as discipline-specific) language and literacy, students complete academic English courses informed by Systemic Functional Linguistics as well as have access to on-demand academic English support via writing consultations.[2]

These multiple, relatively uncommon, aspects of the programs at VC make it an attractive context for researching learners' language characteristics, use, and development. In what follows, we recount the genesis of the international collaboration that led to the VanCor project.

International Collaboration: A Brief History

The VanCor project brings together researchers and educators from UBC in Canada, and the Federal University of Rio Grande do Sul (UFRGS) in Brazil. The genesis of this project was in late 2019, over conversations amongst Simone, Alfredo, Laura, and Sandra, about ways to collaborate around a project of mutual interest. Since one of the mandates of VC is to serve as a living lab for pedagogical and research innovation, designing a research project with the goal of supporting activities such as curriculum development and design of student tasks seemed most fit and appealing. Given Simone's expertise in LC development and the desire from VC members to create an institutional learner corpus, our group decided to embark on a project, seeing the potential benefits of the international collaboration. By early 2020, we had obtained competitive funding via a Social Sciences and Humanities Research Council (SSHRC) institutional grant. This funding supported the hiring of our two graduate research assistants from UBC. In what follows, we provide an overview of the project sequence and key stages.

2 For further details on the Vantage program, see Zappa-Hollman & Fox (2021), Ferreira & Zappa-Hollman (2019), and Zappa-Hollman (2018), as well as the Vantage College website: https://vantagecollege.ubc.ca/program-overview

Project Timeline

The pilot phase involved four stages (Fig. 1).

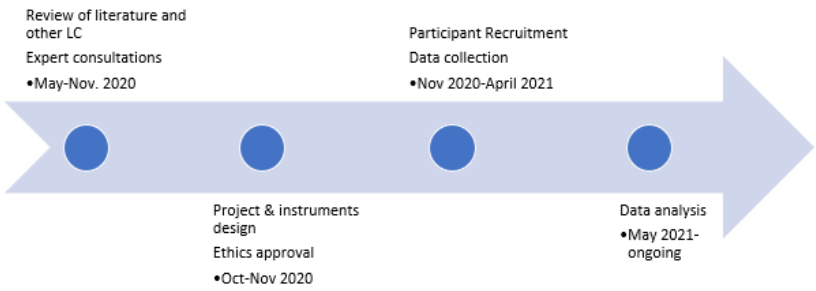


Figure 1. VanCor project Timeline

The first stage involved an extensive, updated review of the literature on learner corpora, with a focus on the creation and uses of LC for research and pedagogical applications. This literature review was complemented with consultations with experts in learner corpora in university contexts outside of Canada as well as with consultations with UBC librarians with expertise in data management.

The second stage involved defining the scope, objectives, procedures, timeline, and developing the data collection instruments. To collect the learner texts that our project participants were willing to share with us, we used a survey (hosted on the Qualtrics survey tool). This survey, reproduced in Appendix A, included a section for collecting demographic data about the participants, a second section about assignments information and for uploading assignments (up to 15 files), and a third section inviting participants for a debriefing interview. The interviews aimed to gather feedback from the students about their experience participating in the study (i.e., completing the survey), potentially offering deeper insights about the process of writing their assignments. To complete the survey, participants had to first provide their informed consent via a form included at the start of the survey. The survey also included the request for student

consent to the collection and use of their data. At this stage we also applied to the institutional ethics board for approval to conduct this pilot study.

The third stage involved participant recruitment and data collection. This stage spanned six months and took place virtually[3] in two courses taught by two instructors who are also members of this project team. In late November 2020 (end of our Fall term), we recruited participants in one section[4] of an academic writing course taught by Laura Baumvol in the Arts program and collected texts from this class until January, 2021. At the start of the Winter term, we recruited participants from two sections of an adjunct course taught by Alfredo Ferreira that links EAP instruction to courses in the Science program. The recruitment was carried out by the two graduate research assistants during a 15-minute class visit of a synchronous session when the instructors were not present. During this visit, the students were introduced to the project through a 5-minute video with an overview of the project goals and a description of what participating involved. This was followed by some Q&A time in case prospective participants had any queries. After the class visit, a link and QR code to the survey was posted as an announcement on the course learning management system sites.

In total, we collected nine assignments and two sets of instructions, and conducted two interviews were conducted; these took place once the final grades for their respective classes had already been awarded. Following data collection, the fifth stage involved data preparation and data analysis. To protect the identity of participants and systematize the process of data management, we assigned unique identifiers to each text and instructions, and removed all personal identifying information prior to starting with data analysis. Next, we used a metadata coding sheet to describe the relevant context and genre of each text. We developed our text metadata coding sheet partly based on a similar resource from Graves and Hyland (2017) with some adaptations for our context and project purposes. The

3 Since our project was carried out during the Covid-19 pandemic, all research activities – including recruitment and data collection – were carried out online.

4 Each course section has a student registration of 25, maximum.

coding sheet can be found in Appendix B[5]. For this classification, we are drawing on Systemic Functional Linguistic theory. Section 5 includes an illustration of the analysis of corpus data for use in research and instruction.

Key Reference Literature

As mentioned above, we consulted canonical texts on learner corpora (Granger, 2002, 2009, 2015; Gardner & Nesi, 2013; Römer & O'Donnell, 2011) to gain insights on types of data to collect, steps, and sequencing to follow, as well as tips to avoid common pitfalls and minimize challenges in data retrieval and analysis. Recent articles focusing on the process of designing and implementing a LC were helpful to learn from insights the authors gained through trial and error.

For instance, Granger et al's (2020) *International Corpus of Learner English* (ICLE)[6], which is composed of texts written by upper intermediate and advanced learners of 25 different language backgrounds offers an excellent model for gathering metadata on the texts that allow for an in-depth view of both the learners and the tasks.

Some projects have expanded their scope to provide additional types of resources to assist with writing research, support instructors' professional development, and train those intending to design and use an LC. Two such corpora we found impressive in this regard are the *Multilingual Academic Corpus of Assignments: Writing & Speech* (MACAWS) and the *Corpus & Repository of Writing* (CROW), both with Dr. Shelley Staples as a lead investigator. MACAWS (Staples et al., 2019) is an ongoing building corpus of assignments written by students enrolled in language programs at the University of Arizona. CROW (Staples & Dilger, 2018) contains texts that L1 and L2 first-year undergraduate students write in their composition classes in three universities in the US. Access to these resources is available by requesting registration to their customized websites. Once registered

5 This genre classification system will be revised as we collect more texts from different genres.

6 https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

with the MACAWS website, for example, we were able to access a repository of pedagogical materials associated with the assignments, such as syllabi, assignment sheets, lesson plans, and instructional materials; and language learning activities in Portuguese and Russian designed based on the language patterns that emerge from the corpus. In CROW, we also accessed demographic data and a repository of resources intended to help others – like us – with the design and use of LCs. The resources shared in these two projects guided our decisions about several aspects of our own project. For example, the demographic information helped us further refine the kind of metadata we would collect. The corpus helped us reflect on whether collecting drafts of our students' final writing was useful for our project. The pedagogical resources provided several suggestions for ways in which to involve other VC instructors in the next phases of our project.

Another corpus that informed our pilot is the *The Civil Engineering Writing Project* Conrad (2017)[7], led by Susan Conrad at Portland State University. The corpus includes student and practitioner writing in the field of engineering, as well as an impressive collection of open-source instructional materials for use by course instructors and students in self-study. The studies that emerged from the project (reviewed below) relate to genre and linguistic analysis, grammar and mechanics errors, as well as holistic evaluations of writing effectiveness. The genre classification in this corpus helped us reflect on the way we would approach genres in our own texts.

We also drew from "Writing assignments across five academic programs" by Graves (2017), a chapter in an edited book by Canadian researchers who created a corpus of undergraduate student assignments. From this resource we used the writing assignments coding guide, which served as the basis for our coding guide. This coding sheet is used to record standard information (e.g. genre, length of text, topic, grade, etc.) collected about each text, which is then entered into the web-based application that compiles them and creates reports. In addition to guiding our coding procedures, the process of adapting the coding sheet became an opportunity for our team to revisit and adjust as needed the goals and scope of the project.

7 http://www.cewriting.org/

Alongside canvasing websites and scholarly publications, we also reached out to a number of the scholars who led those works, primarily with questions related to preferred communication and collaboration practices within their research team, and questions related to data collection and management.[8] The guidance that was generously offered extended well beyond these questions. The scholars candidly shared their experiences of learner corpora development (e.g., MACAWS, CROW) and lessons learned along the way. They highlighted important yet often overlooked aspects of corpus development such as steps to ensure a sufficient number of texts are collected and advised starting small, staying focused on the scope, which may involve starting with a smaller project before scaling it up. Based on these insights, we adjusted our timeline for collecting the texts, and decided that in the scaled-up version of our project we will not provide monetary incentives for participation (as these can prove challenging for distribution as well as add significant cost to the project). These projects also provided access to a wealth of resources encompassing the lifecycle of a corpus-building project, from detailed information on developing the backend of the corpus, such as the database structure, automated tools, indexing, text-processing tools, and illustrations of how corpora can be used to create relevant pedagogical materials.

Drawing on the LC community of practice helped us reflect on our research questions and practices with experts in the field, make informed choices that strengthened our project, and enabled us to refine the project and move forward with heightened confidence. We also consulted experts in digital scholarship through workshops that provided crucial training on the choice of digital tools available for project management, data collection and storage, and the dissemination of project outcomes. These training sessions also introduced us to institutional norms and best practices pertaining to handling sensitive research data (e.g., institutional requirement to store data on Canadian servers).

8 We are extremely grateful for the generosity of our colleagues from the Corpus Linguistics field who have kindly shared their knowledge with our team.

Analysis of selected data from pilot project

Insights into student writing based on quantitative analysis of a large sample is a key goal of local learner corpora collected for immediate pedagogical purposes. While a corpus that is sufficiently large for quantitative analysis would have been a welcome outcome of our pilot project, as mentioned earlier, the first aim of the project was to test approaches to data collection in our context. Having described the steps we followed in designing, collecting, and storing the data for our project, we now focus on a small subset of four texts collected from Science students to illustrate how even such a small sample can inform teaching, research, and the VanCor project in valuable ways.

The contributions to instruction of a very limited sample of student texts from the same instructional setting can be likened to those of qualitative, case study analysis (Duff, 2008), the primary overlap being that the data emerge from a specific, well-defined context. As with case study data, such pilot corpus data suggest hypotheses about learner practices that can be subsequently explored in a wider study, targeted for collection in larger corpora, and, on the teaching side, can inform the development of instructional materials to test in classrooms for fit with student needs and interests.

Given the interest in forming hypotheses and developing instructional materials from the pilot data, two aspects of the data come directly into play. First, it is important to recognize that the texts are not representative of the Science program cohort or all students in the class: these are relatively successful texts voluntarily submitted by four high-performing students within the top 10% of the class. This quality of these data point to a weakness in the opt-in approach to the collection of student texts: generally, high-performing students submitted writing assignments that were also high-performing in terms of the grade received.

The second aspect of the data that inform their use for instruction is the nature of the writing undertaken in this assignment and our focus on pedagogical application. The students wrote comparative discussions, approximately 1,400 words in length, across three drafts with instructor and peer feedback. The student-writer selects two scientific theories, concepts,

or approaches in the history of science to compare in relation to a specified criterion. This critically-engaged discussion typically concludes with claims about the different motivations for these concepts in the history of science[9].

Correspondingly, instruction focused on expository genres, specifically comparative discussions in the history of science. Comparison is a semantic domain relevant in the discussion assignment as well as the first-year Physics, Chemistry, and Mathematics textbooks used by the students, as it is in science discourse more widely. The instructor in this case, Alfredo, observed that students were frequently challenged when using comparative language in their reports, such as from Chemistry labs, as well as in longer writing assignments. This particular discourse analytic research out of VanCor arose from an interest in developing materials that would help address this observed need for instructional materials in the history of science module of the VC science Content and Language Integrated Learning (CLIL) course. Students in the Vantage science stream received no other explicit instruction in the language and functions of comparison.

Qualitative analysis of these texts following SFL theory (Halliday & Matthiessen, 2014) led to a number of instruction and research-worthy insights into the functions of comparison in historical expositions and academic writing more generally. Table 1 outlines the functional range of comparative language identified in student writing across metafunctions (i.e., organization, interpersonal positioning, and representation) and some subfunctions. For background on the one more technical subfunction listed in the table, theme, understood in SFL as the informational point of departure for the clause, see Kang (2016). Within the function of representation, the genre-specific distinction between focal and non-focal compared things is explained below.

9 For a relevant outline of the development of disciplinary literacy practices in history, see Coffin (1997).

Meta-functions	Sub-function	Example of Comparative Language in Students' History of Science Writing
Organization	Title	A **Comparative** Exposition of Celestial Mechanics and Quantum Electrodynamics in relation to the Description of the State of Motion. (Text 09 - Science)
	Thesis statement; topic sentence	In the exposition that follows, phlogiston theory and oxygen theory **are compared** from macro and micro perspectives in studying science. (Text 10 - Science)
	Theme	Neo-Darwinism places greater emphasis on natural selection, whereas eugenics affirms that artificial selection is required to conserve the useful features of individuals (Paul, 2013). […] **These contrasts** will be further discussed within the section below. (Text 07 - Science)
Interpersonal Positioning	Hedge: claim	A **more detailed** exploration of the kinematic relation between two or more objects in macro and micro perspectives is provided to consider the difference between the types of acting force. (Text 09 - Science)
	Hedge: disciplinary category	"**better adapted individuals**" can be described as a group of organisms with higher reproductivity which enables their **"more useful"** genetic characteristics to pass onto their offspring and onto the future generations, whereas "**less adapted individuals**" are less likely to survive (Abbey & Abalaka, 2011). (Text 07 - Science)
Ideational	Comparison of focal things	**Comparing** the symmetrical aspect of nature has the possibilities to predict the existence of unknown materials or phenomena in the universe (Capra, 1975). (Text 08 - Science)
	Comparison of non-focal things	A **more detailed** exploration of the kinematic relation between two or more objects in macro and micro perspectives is provided to consider the difference between the types of acting force. (Text 09 - Science) **Both** observation and experiments are indispensable in studying science, making science more rigorous and accurate (Ainsworth et al., 1991). (Text 10 - Science)

Table 1. The functional scope of comparative language in high-performing expositions in History of Science by first-year science students.

These data highlight several important features of comparative language that can help develop hypotheses about this area of discourse for use in teaching and research. The main finding is that comparative language

realizes all three main metafunctions and various subfunctions. An interesting example of this is within the function of representation (technically in SFL, the experiential function), which indicates two levels of focus when analyzing genres that explicitly set out to compare things: the comparison of two or more things in focus in the comparative text, and the comparison of everything else for purposes such as organizing ideas for the readers, that is the comparison of non-focal things. The latter function arises, for example, in comparing relative degrees of information detail across the text, where the writer signposts "a more detailed exploration of… is provided". This finding indicates that comparison is both a defining feature of some genres and a more broadly functional resource in academic discourse.

These corpus data also help qualify the comparative exposition as a useful genre for understanding the development of student writing. This claim is based on the wide functional scope of comparative language, its Field (realized in the lexicogrammatical choices of for representing ideas, Tenor (interpersonal positioning), and Mode (textual organization) (on these register variables, see Halliday & Matthiessen, 2014). Such a map helps us to chart trajectories of development of language and academic writing within and across functions by focusing on comparative language. In this way, the data also lend validity to the assignment in relation to the course learning objectives which aspire for development across the three metafunctions.

In relation to language and writing development, it is worth noting the potential of extending this map. This opportunity has arisen in the transcript of a visiting lecture (which led the history of science module in the course) by an established historian of science[10]. The lecture includes instances of comparative language used for engagement (superlative/hyperbole used to bait readers into the counter-argument before arguing against it) and politeness through reverse polarity (reference to an unreliable academic source as "not the most impartial judge"). The extension of the semantic potential of comparative language suggested by the practices

10 The transcript referred to here comes from a lecture on Ancient Greek protoscience delivered by Dr. Sylvia Berryman, Philosophy Professor at UBC.

of a more mature scholar shows how comparative language can realize increasingly fine-grained functions in accordance with disciplinary and linguistic development, illustrating Halliday's (1993) conception of language development as increasing one's registerial repertoire or capacity to mean across situated contexts; for discussion in advanced language development, see Matthiessen (2006). These insights indicate potential directions for researching language and writing development in this context.

Moving to a lexicogrammatical view of comparison in History of Science arguments, an analysis of the comparative lexis from the history texts in the pilot corpus yielded the results shown in Table 2 below. The word lists on the right-side columns are classified by grammatical and semantic/functional units, subunits, and whether the words instantiate the semantic domains of similarity or difference. The ordering of grammatical units from nominal group (noun phrase in traditional grammar) at the top of the table down to verb, adverbial, and conjunction at the bottom is motivated by the degrees of information density afforded by these units (i.e., from most abstract and/or general to most concrete) per the concept of ideational grammatical metaphor (Ferreira, 2020; Halliday, 1998). The wordlist in each of the subunit categories are ordered from most to least frequently occurring with the number of tokens listed on the right-hand column.

Grammatical/ Functional Unit		Subunit Similarity	Instances in Pilot Corpus		#
			Difference		
1	Nominal Group / Participant	Head Noun / Thing	comparison/s		8
			similarities		8
				difference/s	7
				contrast/s	2
			alignment		1
				opposite	1
				superiority	1
		Adjective & premodifier/ post-pointer, describer		different	8
				opposite	7
				better	3
				greater	3
			both	better adapted	2
			comparative	broader	2
				higher	2
				more likely	2
			similar	deeper	1
			corresponding	less adapted	1
				less likely	1
				more appealing	1
			same	more like	1
				more predictable	1
				more regular	1
				more useful	1
				opposing	1
				proportional	1
				superior	1
2	Verb / Process	Relational process	overlap		1
			share		1
		Material & other process	compare/d		9
			comparing	contrasts	3
				distinguished from	2
			correlated with		2
3	Adverbial / Circum-stance		also	by contrast	1
			like	in contrast with	1
			similarly	more frequently	1
4	Conjunc-tion/ Relator		as	while	4
			as	whereas	2
				however	1
				rather than	1

Table 2. Comparative lexis by grammar and function in four high-performing History of Science expositions in 1st-year EAP

As can be seen, the tokens of comparative language cluster significantly in the nominal group (e.g. "These **contrasts**"; "A **more detailed** exploration of the kinematic relation between two or more objects in macro and micro perspectives"). This result can be understood to reflect the relatively high functional load of the nominal group in academic writing especially with regards to the specification of concepts and foci that is associated with disciplinary writing development in university (Duff et al., 2015). Unsurprisingly, abstract concepts involving comparison are central to texts and genres that set out to compare historical theories in science.

The finding of high frequency of comparative language in nominal groups relative to its use in more dynamic processes (verbs), circumstances (adverbs) and logical reasoning (conjunctions) points to a need for additional attention to this role of comparative language in construing abstract concepts in writing instruction. A cursory examination of two popular EAP writing textbooks, both fourth editions (Oshima & Hogue, 2005; Blanchard & Root, 2017), highlights a potential emphasis on the latter dynamic, syntactically more complex and material meanings, while the more frequent realizations of abstract concepts involving nominal groups receive little explicit attention. The "comparison signal words" recommended as useful for comparative writing in one of these textbooks, shown in Table 3, illustrates this tendency:

Comparison Signal Words
Transition Words and Phrases: similarly; likewise; also; too
Subordinators: as; just as
Coordinators: and; both… and; not only… but also; neither… nor
Others: like (+noun); just like (+noun); similar to (+noun); (be) like; (be) similar (to); (be) the same as; (be) the same (be) alike; (be) similar; to compare (to/with)

Table 3. Words and phrases used in comparisons recommended by popular EAP writing textbook (Oshima & Hogue, 2005: 116-117)

According to this edition of the textbook, students should focus their attention on realizations of comparison for these functions of logical ordering and transition with minimal attention given to elements of the nominal

group (noting that the "+noun" elements under "Others" do not themselves realize a comparative meaning). Such an emphasis does not align with the functional distribution of comparative language in the sub-corpus of high-performing texts in the History of Science.

These results suggest potentially useful insights for research and instruction. We have found that the semantic scope of comparison encompasses a wide functional range of language: ideational, interpersonal, and organizational meanings, and various sub-functions of these such as evaluation, affect and multiple scales of text organization including signposting through topic sentences and various cohesive devices.

Given the wide functional scope and grammatical realizations of comparative language in the comparative exposition genre, a relatively holistic perspective on language and writing development in EAP contexts can be operationalized by focusing on comparative language in this genre. The same results suggest various corpus-based approaches and tasks for instructional curricula involving comparative and related genres of academic writing. In these and other ways, the focus on a few student texts within a relatively specific written genre has yielded useful insights to apply to teaching, research, and the next phases of the VanCor project.

Current status and next steps

Through our collaboration on the VanCor project, the team has taken the first steps in designing, compiling, storing, and applying learner corpora: reviewing the literature, consulting with experts, piloting the various sub-tasks involved in data collection, and analyzing the results. These experiences in the pilot phase of the project will inform the next phase of the project.

Our efforts to disseminate our ideas and experiences range from the local to the global. We introduced our LC project and preliminary findings to our VC program colleagues with the aim of generating interest in collaborating on the larger scale of the project through realizing its potential for curriculum development, instruction, and research. Additionally, we have engaged in dissemination efforts, which include presentations at

professional organization annual conferences[11], with the intent of sharing the insights gained from our pilot and sharing our preliminary findings.

As for the next steps in VanCor itself, we plan to implement the project by inviting all VC instructors as collaborators and thus expand the nature of the student texts included in the LC. A higher number of instructor collaborators across all VC programs will allow us to collect texts from, ideally, all courses included in first year programs. This scope of text types will result in a diversity of genres across several disciplinary fields, expanding the potential contributions of the corpus to research.

Acknowledgments

The UBC graduate research assistant positions for this project, fulfilled by Greta Perris and Sara van Dan Acker, were supported by an institutionally administered Social Sciences and Humanities Research Council (SSHRC) grant (UBC Explore SSHRC grant), for which we are very grateful. Simone Sarmento would like to thank the support of CNPq Productivity Grant and CAPES Print. Our gratitude also goes to the students who participated in this pilot phase of our project, generously sharing samples of their writing. We would also like to thank Brian Wilson, Curriculum Manager at VC, for his feedback and advice on survey design; Dr. Shelley Staples, CROW and MACAWS project leader, and two of her team members, Dr. Bruna Sommer-Farias and Dr. Nina Conrad; for their time to meet with us and give us access to their project materials and sharing their lessons learned with us; and Dr. Susan Conrad, leader of the Civil Engineering corpus project, for making available for us a number of helpful teaching materials derived from that corpus.

11 Zappa-Hollman, S., Ferreira, A. A., Perris, G. & Matte, M. L. (March 2022). Designing a local learner corpus for pedagogical applications and research. Paper presentation at the Virtual *TESOL Annual Convention*.

References

Blanchard, K. & Root, C. (2017). *Ready to write 3: From paragraph to essay* (4th Edition). Pearson.

Coffin, C. (1997). Constructing and giving value to the past: An investigation into secondary school history. In F. Christie & J. R. Martin (Eds.), *Genre and institutions: Social processes in the workplace and school* (pp. 196-230). Cassell

Conrad, S. (2017). A comparison of practitioner and student writing in civil engineering. *Journal of Engineering Education, 106*, 191-217. doi:10.1002/jee.20161.

Duff, P. (2008). *Case study research in applied linguistics*. Lawrence Erlbaum/Taylor & Francis.

Duff, P. A., Ferreira, A. A. & Zappa-Hollman, S. (2015). Putting (functional) grammar to work in content-based English for academic purposes instruction. In M. A. Christison, D. Christian, P. A. Duff, & N. Spada (Eds.), *Teaching and learning English grammar: Research findings and future directions: A festschrift for Betty Azar* (pp.139–158). Routledge.

Ferreira, A. A. (2020). Sociocultural development in the spectrum of concrete and abstract ideation, *Mind, Culture, and Activity, 27*(1), 50-69, doi:10.1080/10749039.2019.1686027

Ferreira, A. & Zappa-Hollman, S. (2019). Disciplinary registers in a first-year program. A view from the context of curriculum. *Language, Context and Text, 1*(1), 148-193. https://doi.org/10.1075/langct.00007.fer

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, *34*(1), 25-52. https://doi.org/10.1093/applin/ams024

Granger, S. (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. John Benjamins Publishing Company.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching.* John Benjamins, 13–32. https://doi.org/10.1075/scl.33.04gra

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In Granger, S., Gilquin, G. & Meunier, F. (eds.) *The Cambridge handbook of learner corpus research.* Cambridge University Press, pp. 486-510.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020). *The International Corpus of Learner English*. Version 3. Presses universitaires de Louvain. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Graves, R. (2017). Writing assignments across five academic programs. In R. Graves & T. Hyland (Eds.). *Writing assignments across university disciplines*. Trafford Publishing.

Graves, R. & Hyland, T. (Eds.) (2017). *Writing assignments across university disciplines*. Trafford Publishing.

Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education, 5*, 93–116. doi:10.1016/0898-5898(93)90026-7

Halliday, M. A. K. (1998). Things and relations: Regrammaticising experience as technical knowledge. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives*. Routledge. pp. 185–235.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.

Kang, J. (2016). A functional approach to the status of theme and textual development. *Theory and practice in language studies, 6*(5), 1053-1059. http://dx.doi.org/10.17507/tpls.0605.20

Matthiessen, C. M. I. M. (2006). Educating for advanced foreign language capacities: Exploring the meaning-making resources of languages systemic-functionally. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 31–57). London, UK: Continuum.

Oshima, A. & Hogue, A. (2005) *Writing academic English* (4th ed.). Pearson-Longman.

Römer, U. & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 6*(2), 159-177. doi:10.3366/cor.2011.0011

Staples, S., Novikov, A., Picoral, A. & Sommer-Farias, B. (2019-). Multilingual Academic Corpus of Assignments – Writing & Speech (MACAWS). Available at https://macaws.corporaproject.or

Staples, S. & Dilger, B. (2018). *Corpus and repository of writing [learner corpus articulated with repository]*. Available at https://crow.corporaproject.org

Zappa-Hollman, S. (2018). Collaborations between EAP and disciplinary instructors: Factors and indicators of positive partnerships. *International Journal of Bilingual Education and Bilingualism, 21*(5), 591-606. doi:10.1080/13670050.2018.1491946

Zappa-Hollman, S. & Fox, J. (2021). Engaging in linguistically responsive instruction: Insights from a first-year university program for emergent multilingual learners. *TESOL Quarterly, 55*(4), 1081-1091. https://doi.org/10.1002/tesq.3075

Zappa-Hollman, S., Ferreira, A. A., Perris, G. & Matte, M. L. (March 2022). *Designing a local learner corpus for pedagogical applications and research*. Paper presentation at the Virtual TESOL Annual Convention.

Appendix A

Vantage Corpus of Student Texts Across Disciplines Project Survey

[Embedded institutional student consent form included here in original survey. The survey can be completed only after students provide informed consent]

Part 1 - Demographic information

Q1 What is your name? (as it appears in your UBC ID)

Q2 Please write down your preferred e-mail so that we can contact you:

Q3 Please confirm your email:

Q4 What Vantage Program are you in?
- Arts
- Science
- Engineering

Q5 How old are you?
- 17 to 19 years old
- 20 to 22 years old
- older than 22

Q6 What is your preferred gender?

- Male
- Female
- Other _____

Q7 What is (are) your native language(s)? You can select one or more, as it applies to you, to a maximum of three.

☐ Arabic
☐ Cantonese
☐ Farsi
☐ French
☐ German
☐ Hindi
☐ Indonesian
☐ Japanese
☐ Korean
☐ Malay
☐ Mandarin
☐ Mongolian
☐ Portuguese
☐ Russian
☐ Spanish
☐ Other _____

Q8 How many years of high-school education did you complete in **English**?
None

- 1
- 2
- 3
- 4
- More than 4

Q9 In what country did you receive your high-school diploma? (If none of the countries apply to you, please select Other at the end of the list.)

- Brazil
- Cambodia
- Canada
- Chile
- China
- Taiwan
- Ecuador
- Egypt
- France
- Germany
- Hong Kong
- India
- Indonesia
- Iran
- Japan
- Korea
- Macao
- Malaysia
- Mexico
- Mongolia
- Panama
- Russia
- Other

Q9a Other: In what country did you receive your high-school diploma?

End of Part 1 - Demographic Information (participants complete this once)

Part 2 - Assignment Information and upload

Q10 Would you like to upload another assignment?
- Yes

- No

Q11 Vantage College Corpus of Texts Across Disciplines Assignment information and uploading. Please, answer the following questions and then upload your assignment.

You will be prompted to answer the same questions for every assignment you upload.

Q12 Assignment upload:
Is this a single document?
- Yes
- No

> Q12a You can upload only one document at a time. Please select another document and continue answering the questions.

Q13 Are you the only author of this assignment?
- Yes
- No

> Q13a You can only submit an assignment completed by you only. Please select another assignment that you completed by yourself. Assignments completed together with your peers or classmates as part of pair/group work cannot be accepted.

Q14 Have you received a grade for this assignment?
- Yes
- No

> Q14a You can only submit assignments that have been graded. Please select an assignment you have completed by yourself and for which you received a grade.

Q15 Course you completed this assignment for:
(NOTE: you can only upload assignments submitted only for the courses listed below)

Course
- VANT 140
- WRDS 150
- ASTU 204

Q16 What grade did you receive for this assignment?
- 90 - 100
- 70 - 89
- 60 - 69
- 50 - 59
- below 50
- Prefer not to answer

Q17 Upload your assignment:

Q18 If available, please upload the instructions you received to complete this assignment.

End of Part 2 - Assignment Information and upload
Part 3 - Interview Invitation

Q19 Thank you for uploading your assignment(s).
How easy or difficult was it to answer the questions and upload the assignment?
- Extremely easy
- Somewhat easy
- Neither easy nor difficult
- Somewhat difficult
- Extremely difficult

Q20 Would you be available to participate in a 30 minute interview to share your experience in this pilot project and to share with us information about the process of writing your assignment(s)?

For your participation in the interview you will receive a $20 UBC Bookstore web gift card.
Do you want to participate?

- Yes, please send me more information about the interview.
- No

Q21 Is this the email you would like to be contacted at: [email entered by participant]?

- Yes
- No

Q22 Please, provide your preferred e-mail so that we can send you more information about the interview.

End of Part 3 - Interview Questions

Appendix B

VanCor Metadata Annotated Coding Sheet

Date coded: [yyyy/mm/dd]

Coder: [Name of person who coded]

Project: [name of LC project]

Assignment UID: [unique ID assigned to text being coded]

- Date submitted to instructor: [yyyy/mm/dd]
- Date submitted to VanCor: [yyyy/mm/dd; this is the dat the assignment was uploaded by the participant to the Qualtrics survey]

Vantage Program: [select what applies]

Science

Engineering

Arts

Type of course: [Include here dropdown menu with list of courses from the corresponding program]

- EAP Writing course
  - LLED 200
  - LLED 201
- EAP disciplinary-linked course
  - VANT 140
- Other writing and communication course
  - ASTU 204
  - WRDS 150
- Disciplinary courses

Semester:

W1 [September-December]

W1-2 [September-April]

W2 [January-April]

S [May-July]

Course length in weeks: [include number of weeks]

Demographic Info:

- Age:
- Gender:
- Native language(s):

- Years of high-school education in English:
- Country received HS diploma:

Assignment

- Grade received:
- Percentage of final grade:
- Researcher's rating:
- Assignment **instructions** provided?
    - Yes
    - No
- Genre:
    - Instructor's label if provided: [this refers to the way the instructor called the genre of the assignment; e.g., annotated bibliography]
    - Student's label if provided: [this refers to how the student may have labeled the genre of the assignment; e.g., "in this discussion" - this is determined by looking at "clues" related to the overall structure of the text; e.g., "On the one hand…on the other hand…"]
    - Researcher's label: [use SFL-based classification]

- Is this assignment a component of a larger assignment? Yes/No

    No

    Yes: (link to genre of final assignment) (e.g. Results part of IMRD)

Length/# words:

Title:

Visuals included in the text? (e.g., figures, images, symbols, tables, graphs):

    No

    Yes

Completed In-class?

    Yes

    No

Completed out of class?

    Timed

    Not timed

The role of genre in academic language use: the case of Critiques and Case Studies in BAWE

Marine Laísa Matte (UFRGS)
Deise Amaral (UFRGS)
Larissa Goulart (Montclair State University)

Introduction

As users of any language, we know that different linguistic features are employed when we write or speak for different purposes. When we write a Facebook message, for instance, we use colloquial linguistic resources, like contractions and subject omission, that we do not include in a course paper (Biber, 2006). In other words, texts with different communicative purposes aimed at different interlocutors adopt distinct linguistic features in order to convey these purposes. However, it is only recently that academic discourse started to be treated not as a single unit; instead it has been shown that there is variation between different genres in academic writing (Biber, 2006; Biber & Gray, 2016; Hardy & Friginal, 2016; Staples et al., 2016, 2018; Staples & Reppen, 2016). Undergraduate argumentative essays and research articles, for instance, are both part of what we call academic discourse; nevertheless, these two genres have distinct characteristics (i.e., length, methodology description, the use of visual elements, etc.), which are reflected in the language used in their texts.

Although the texts required by teachers in university settings are referred to as assignments or course papers, in university writing variation becomes even more salient as the required texts can vary from laboratory reports to case studies or explanations. Gardner and Nesi (2013) suggest that some university assignments are written in preparation for professional practice (Case Studies, Designs, Proposals, among others) while others

are written as a form of showing independent reasoning and of developing critical thinking (Essays, Critiques, etc). The goal of this paper is to investigate how linguistic features vary in two academic genres of unpublished university writing: Case Studies and Critiques. The research questions to be answered are:

a) To what extent is there linguistic variation between Case Studies and Critiques?
b) How is this variation reflected in the way different academic language features are used to express the communicative purposes of Case Studies and Critiques?

Academic writing and genre/register[1] studies

Academic writing is usually considered more complex than writing in non-academic contexts. But what does complex mean? In this paper, we align with Biber and colleagues' definition of complexity (e.g., Biber et al., 2021), where grammatical complexity is defined as the addition of optional structural elements to simple phrases and clauses. Biber et al. (2011) used corpus-based analyses to contrast the grammatical complexity of academic research articles and conversation through 28 lexico-grammatical features associated with structural complexity in previous studies (Biber, 1988,

1 The terms "genre" and "register" have been used alternately depending on the time the research study has been produced and/or on the distinct conceptualizations they represent. Most researchers choose not to make any distinction between the terms and use one or the other without specifying the construct being followed. However, when they are theoretically distinct, genre studies tend to focus their analyses on "the conventional structures used to construct a complete text within the variety" (Biber & Conrad, 2009: 2) while register studies analyses search for "characteristic lexico-grammatical linguistic features" (Biber, 2006: 11). Both perspectives look for "linguistic varieties associated with particular situations of use and particular communicative purposes" (Biber, 2006: 11). According to Berber Sardinha, "register has been proposed as a central construct in corpus linguistic research (Biber, 2012) and as the driving force behind the analysis, rather than as an afterthought: 'the practice advocated […] is to begin a research study with the hypothesis that […] register differences exist, and to include analysis of those differences unless they are empirically shown to be unimportant' (Biber, 2012: 34)." (Berber Sardinha, 2014: 241).

1992, 2006; Biber et al., 1999). Their findings show that conversation is characterized by clausal elaboration (complement clauses, adverbials, etc.), while academic writing contains more phrasal compression (e.g., nominalization, non-finite clauses, etc.). This means that the phrasal constructions seen in "the use of different transformations would have significant effects on our perceptions of spatial patterns in kelp holdfast assemblages" (Biber et al., 2011: 27) are characteristic of academic language writing, such as prepositional phrases modifying the noun (*of different transformations*), attributive adjectives (*different, significant, spatial*), and nominal premodifiers (*kept holdfast*).

Several researchers have focused their attention on the analysis of grammatical complexity to account for language development in L1 (Biber et al., 2011; Ansarifar et al., 2018) or L2 writing (Bulté & Housen, 2018; Goulart, 2020; Kuiken & Vedder, 2019). Ansarifar et al. (2018) compared 99 MA-level abstracts and 64 PhD dissertation abstracts written by L1 Persian writers with 149 research article abstracts by expert writers in Applied Linguistics. The authors found that phrasal features would develop along university years of study, with the MA group differing significantly from the expert writers, while the PhD abstracts did not show such a difference in relation to the published articles, corroborating Biber et al.'s (2011) hypothesized stages of complexity development.

The use of phrasal constructions is also discussed in Staples et al. (2016). The authors conducted a study on academic writing development by analyzing texts retrieved from BAWE[2] organized in four levels of study (three years of undergraduate and first year of MA), four different genres (Essays, Critiques, Case Studies and Explanations), and in the four disciplinary groups present in BAWE (Arts and Humanities, Social Sciences, Life Sciences, and Physical Sciences), meaning that three separate analyses were done. By including 23 linguistic features that previous research has shown to account for grammar complexity (Biber et al., 2011; Biber & Gray, 2013; Biber et al., 2016), their study corroborates the assumption that there

2 The British Academic Written English corpus (BAWE - Nesi et al., 2008-2010) is a collection of proficient texts written by university students from 2004 to 2007 with representation of discipline areas and genre families.

is a high incidence of phrasal features in advanced academic writing - as students gain experience, their writing tends to become more compressed, that is, less explicit, with a preference for using more phrasal features which "are more economical and allow writers to package information more densely" (Staples et al., 2016: 179).

In their analyses of disciplines and genres, Staples et al. (2016) conclude that the writers in Arts and Humanities used more clausal features, such as finite clauses (e.g., *although the number of participants in the coup itself was indeed small*) than the ones in Life and Physical Sciences, who tend to use more phrasal features, such as attributive adjectives (e.g., *unique*, *efficient*). That implies that, in BAWE, Case Studies present more phrasal features than Critiques, as most of the Case Studies are from these two areas of the hard sciences. This coincides with the results of a study of academic sub-genres in Biber and Gray (2016) which finds that research articles in Humanities use more clausal modification while the ones in the Natural Sciences rely on phrasal structures modifying nouns (e.g., *patient report*).

Other studies have considered genre differences in analyses of lexico-grammatical variation, finding that certain language features are more recurrent in specific situationally-defined varieties than in others (Biber, 2006, 2012; Biber & Conrad, 2009). From the studies that compare more general registers, such as oral against written (Biber et al., 2011, 2016), to investigations of differences among academic genres (Biber, 2006; Biber & Gray, 2016; Hardy & Friginal, 2016; Staples et al., 2016, 2018; Staples & Reppen, 2016), there seems to be a growing understanding that genre mediates variation in language. In the present study, although looking only at genre variation in academic writing, we believe that as academic experience is gained, particular lexico-grammatical features are developed, like phrasal constructions (Biber et al., 2011; Ansarifar et al., 2018; Staples et al., 2016). Differently from many of the studies mentioned above that compare levels of academic writing, from undergraduate to PhD and expert published articles, we chose to work exclusively with texts written by MA students, the highest level in BAWE, which might indicate that they have already had a greater exposure to academic language when compared to

less experienced university students. Next, we present the corpus of study as well as the methodological procedures.

Methodology

The corpus

In order to answer our research questions, we have selected two university genres from the BAWE corpus, Critiques (CR) and Case Studies (CS). According to Gardner and Nesi (2013), CS are written in order to prepare students for professional practice and usually take large amounts of data into account. The authors describe the purpose of CS as "to demonstrate understanding of professional practice through the analysis of a single example" (Gardner & Nesi, 2013: 37). CR, on the other hand, require that students show informed and independent reasoning while also developing "understanding of the object of study and the ability to evaluate and/or assess its significance" (Nesi & Gardner, 2012: 94). As discussed above, we have selected CR and CS - two genres with different communicative purposes - written by first year MA students. As for native language, we included texts written in English as L1 and L2, without making the distinction, since our object of investigation is the development of academic writing irrespective of L1.

Our final corpus contains 95 CS and 83 CR from the 4 disciplinary groups in BAWE, Arts and Humanities (AH), Social Sciences (SS), Life Sciences (LS), and Physical Sciences (PS). Table 1 describes the number of texts, the number of words, and the average text length for each genre. As this table shows, CS are somewhat longer texts than CR. In addition, we can see a difference in length across disciplines, with SS CR being the shortest, and PS CS the longest.

Genre	Discipline group	Nr of texts	Nr of words	Mean text length
Case study	AH	-	-	-
	LS	66	161,080	2,440.6
	PS	10	38,139	3,813.9
	SS	19	64,692	3,404.8
	Total	***95***	***263,911***	***2,778***
Critique	AH	15	37,447	2,496.5
	LS	30	73,306	2,443.5
	PS	13	33,225	2,555.8
	SS	25	48,572	1,942.9
	Total	***83***	***192,550***	***2,319.9***

Table 1. Description of the corpus

These texts were tagged for a series of grammatical features associated with academic language using the Biber Tagger[3] (Biber, 1988). The features included in our analysis are described in the next section.

Lexico-grammatical features

The linguistic features used to contrast academic language in CR and CS were chosen based on a review of studies that examined academic writing in English as well as complexity features associated to academic language (Biber, 2006; Goulart, 2020; Gray, 2015; Parkinson & Musgrave, 2014; Staples et al., 2016; Staples & Reppen, 2016). We have included phrasal features (nouns - group, stance, abstract and cognitive nouns - plus stance nouns followed by prepositional phrases, attributive adjectives, premodifying nouns, and nominalizations), clausal features (verbs, subordinate clauses - causative, conditional and others -, *that*-complement clauses controlled by verbs, and clausal coordinating conjunctions) and also intermediate features encountered in these previous studies. Intermediate features are

3 Although originally designed to be used in Multidimensional Analysis studies, the Biber Tagger has been recently used by researchers that analyze grammatical complexity in L2 writing. (Biber et al., 2011; Biber et al., 2016).

clauses that add information to the noun phrase, such as relative clauses, or complement specific types of nouns. We have selected passive voice, relative clauses, *to*-complement clauses controlled by verbs of desire and stance nouns, and that complement clauses controlled by nouns - attitudinal, stance and nouns of likelihood. The 23[4] linguistic features selected can be seen in Table 2.

Linguistic feature	**Example**
Phrasal features	
Nouns	keyboard, engine, patient
Attributive adjectives	short term, previous research
Nominalizations	satisfaction, interference, invasion
Premodifying nouns	customer satisfaction, patient report
Group nouns	the committee took account of the severity
Stance nouns	reason, claim, assumption
Abstract nouns	a description of the progress
Cognitive nouns	analysis, decision, concern, idea
Stance nouns followed by prepositional phrases	advantages of having a large database
Clausal features	
Verbs	believe, make, propose
Subordinate clauses (causative)	this cannot happen because there would be arbitrage opportunities
Subordinate clauses (conditional)	they are only acceptable if one first accepts the existence of memes
Subordinate clauses (others)	increments of twice the error until the function value goes positive
That verb complement clauses	the findings show that cash flows map into returns with significantly higher coefficients
Clausal coordinating conjunctions	This can be applied for all multiple cylinders engine but is more commonly found in four- and six-cylinders engines
Intermediate features	
Passive voice	the high-level solution was selected
Non-finite to- verb complement clauses controlled by verbs of desire	Those speculators intend to save money to obtain benefits

4 The grammatical variable stance nouns is referred to as *stance nouns in other context*s and *new stance nouns* in Table 3. Thus, Table 2 and 3 have 22 and 23 features, respectively.

Wh-relative clauses	potential flights to Cape town, which will be stored for future access
That-relative clauses	an antidepressant that has a low toxicity
That-noun complement clause controlled by attitudinal noun	it is therefore no surprise that the main objective of the article is to aid conservation agencies in their management of present woodland
That-noun complement clause controlled by noun of likelihood	because of the assumption that more is better or that qualitative research is incomplete
Non-finite to- complement clauses controlled by stance nouns	it might take time to change laws but that is not a reason to inhibit new inventions

Table 2. Grammatical variables included in this study

Statistical analysis

The frequency of occurrence of each feature was normed to 10,000 to account for different text lengths (see Table 1). Since the data did not meet the assumptions of normality and linearity, as shown by Shapiro-Wilk normality tests, we ran Mann-Whitney U tests with the linguistic features as dependent variables and the two genres as independent variables using R version 4.0.5 (R Core Team, 2020), and calculated effect sizes[5] using the *R Companion* package (Mangiafico, 2015). The Mann-Whitney U tests indicated whether there are statistical differences between genres for each feature.

Results

The results of the analysis can be observed in Table 3, which contains the descriptive statistics for each feature in both genres (Critiques and Case Studies): means and medians of the features for each genre as well as the Mann-Whitney U test results, the statistical significance (*p*-value), and the effect size (r_{rb} - rank biserial correlation).

5 "Effect size is a standardized measure, that is a measure comparable across different studies that expresses the practical importance of the effect observed in the corpus or corpora" (Brezina, 2018: 14)

Feature	Mean (SD) CR	Mean (SD) CS	Median CR	Median CS	Mann-Whitney U	Alpha (*p*)	r_{rb}
nouns	328.79 (30.3)	343.49 (32.9)	327	342	2864	0.003*	-0.222
attributive adjectives	70.34 (17.2)	74.94 (12.7)	68.6	73.1	3028.5	0.014	-0.185
nominalizations	86.59 (23.1)	75.51 (18.3)	87.7	70.7	5085	<0.001**	0.275
premodifying nouns	44.93 (18.4)	50.19 (14.1)	42.7	53.3	3016.5	0.013	-0.187
group nouns	1.60 (2.37)	4.06 (3.04)	0.85	3.35	1611	<0.001**	-0.503
stance nouns in other contexts	3.87 (2.83)	2.28 (1.53)	3.45	2.1	5270.5	<0.001**	0.317
new stance nouns	5.014 (4.24)	3.87 (1.85)	4.15	3.65	4303.5	0.182	0.1
abstract nouns	28.86 (10.1)	24.20 (8.57)	27	22.4	5072.5	<0.001**	0.272
cognitive nouns	8.16 (5.4)	6.39 (2.56)	6.6	6.15	4284.5	0.202	0.0965
stance nouns followed by prepositional phrases	3.29 (1.87)	2,91 (1.35)	3.2	2.95	4278	0.208	0.095
verbs	112.57 (14.4)	116.13 (13.1)	114	115	3383	0.162	-0.106
subordinate clauses (causative)	1.09 (1.53)	0.54 (0.77)	0.65	0.3	4737.5	0.006*	0.204
subordinate clauses (conditional)	1.13 (1.36)	1.53 (1.35)	0.9	1.1	3022.5	0.013	-0.187
subordinate clauses (others)	4.07 (2.87)	3.66 (1.7)	3.4	3.4	3918.5	0.846	0.0144
that-verb complement clauses	4.21 (2.57)	2.77 (2.12)	3.8	2.2	5245	<0.001**	0.311
clausal coordinating conjunctions	5.75 (3.67)	4.83 (5.17)	5.1	2.7	4931.5	0.001**	0.241
passive voice	19.87 (7.22)	18.31 (5.08)	18.6	17.4	4196	0.311	0.0761

non-finite *to*-verb complement clauses controlled by verbs of desire	1.04 (1.1)	1.57 (1.12)	0.75	1.35	2608.5	<0.001**	-0.279
wh-relative clauses	4.02 (2.28)	4.55 (1.73)	3.6	4.6	2985.5	0.010*	-0.194
that-relative clauses	3.29 (2.55)	1.68 (1.21)	2.7	1.4	5606.5	<0.001**	0.392
that-noun complement clause controlled by attitudinal noun	0.82 (0.18)	0.03 (0.14)	0	0	4315	0.024	0.17
that-noun complement clause controlled by noun of likelihood	0.39 (0.702)	0.08 (0.16)	0	0	4655	0.004*	0.217
non-finite *to*-complement clauses controlled by stance nouns	1.07 (1.35)	0.60 (0.61)	0.7	0.4	4658.5	0.015	0.183

Table 3. Descriptive statistics for each feature across genres
Significance * <0.01. ** <0.001

Table 3 shows that out of the 23 linguistic features included in our analysis, 12 came up as statistically significant and are marked with asterisks, meaning that there is variation in their use between the two genres investigated. The effect sizes were all small, except for *that-verb complement clauses, stance nouns*, and *that-relative clauses*, which showed medium effects (r_{rb} = 0.31, 0.32, and 0.33 respectively) and *group nouns*, the only feature with a large effect size (r_{rb} = -0.5). These effect sizes are interpreted according to the range established in the literature: small effects are between 0.10 and < 0.30, medium from 0.30 to < 0.50 and large from ≥ 0.50 on (Cohen, 1988).

Within the phrasal features, the Mann-Whitney U tests indicate that six out of 10 features were more present in CR than in CS, five out of 10 showed a significant difference and, among these, only total *nouns* and *group nouns* were more frequent in CS (Mdn = 342 and 3.35) than in CR (Mdn = 327 and 0.85), with *group nouns* presenting the highest effect size

among all the comparisons made, r_{rb} = -0.5. Among phrasal features with a significant difference between genres, the trend observed in *nominalizations* (CR Mdn = 87.7, CS Mdn = 70.7, U = 5085, p<0.001, r_{rb} = 0.27), *abstract nouns* (CR Mdn = 27, CS Mdn = 22.4, U = 5072.5, p<0.001, r_{rb} = 0.27) and stance nouns (*stance nouns in other contexts,* CR Mdn = 3.45, CS Mdn = 2.1, U = 5270.5, p<0.001, r_{rb} = 0.32*)*, as well as the prevalence of *nouns* in CS coincide with the results previously found in the literature (Gardner et al., 2019; Staples et al., 2016, 2018).

When it comes to clausal features, it is possible to observe that half of the six features included in the analysis are statistically significant in the comparison between CR and CS, *subordinate causative clauses*, *that-verb complement clauses* and *clausal coordinating conjunctions*, and all of these are more frequent in CR. *That-verb complement clauses* (CR Mdn = 3.8, CS Mdn = 2.2, U = 5245, p<0.001, r_{rb} = 0.31), present a medium effect size, while *subordinate causative clauses* (CR Mdn = 0.65, CS Mdn = 0.3, U = 4737.5, p=0.006, r_{rb} = 0.2) and *clausal coordinating conjunctions* (CR Mdn = 5.1, CS Mdn = 2.7, U = 4931.5, p = 0.001, r_{rb} = 0.24), showed only small effects.

Four intermediate features are statistically significant, one with a medium effect size, r_{rb} = 0.33 for *that-relative clauses,* and small effects for *non-finite to-verb complement clauses controlled by verbs of desire*, *wh-relative clauses* and *that-noun complement clauses controlled by noun of likelihood.* When medians are observed, one is more frequent in CR, *that-relative clauses,* (CR Mdn = 2.7, CS Mdn = 1.4, U = 5606.5, p<0.001, r_{rb} = 0.39), and two in CS, *non-finite to-verb complement clauses controlled by verbs of desire*, (CR Mdn = 0.75, CS Mdn = 1.35, U = 2608.5, p<0.001, r_{rb} = -0.28), and *wh-relative clauses* (CR Mdn = 3.6, CS Mdn = 4.6, U = 2985.5, p = 0.01, r_{rb} = -0.19). *That-verb complement clauses controlled by nouns of likelihood, U = 4655, p = 0.004,* r_{rb} = 0.22, were also more used in CR, which can be seen from the means (CR M = 0.39, CS M = 0.08) as the frequency was too small for comparing medians.

In summary, of the 23 grammatical features of academic writing investigated, 8 were more recurrent in CS but only in 4 of these the difference was statistically significant. As for CR, on the other hand, 8 features

appeared significantly more frequently in this genre than in CS, showing small to medium effects, though. When we look at the types of features, CR had a bigger presence in the 3 groups - phrasal, clausal and intermediate features.

Discussion

Based on the quantitative results, the qualitative data analysis is presented in this section. Although these two academic genres have some characteristics in common, like the use of stance verbs and nouns followed by complement clauses to state the position of authors cited, the diverse objectives of preparing for professional practice (CS) and developing independent reasoning (CS) are expressed through the use of specific linguistic features. *Attributive adjectives* together with *nouns*, for instance, help in the description of technical terms in CS, mainly related to specifying diseases or products, whereas in CR they appear to evaluate previous studies. *Abstract nouns* and *nominalizations*, which can be preceded by *attributive adjectives*, are found in CR as a way of presenting and discussing theoretical concepts. Other types of nouns, such as *group nouns*, are used to indicate institutions, companies and/or universities and to describe situations or products developed by institutions, a typical feature of CS. Still among phrasal features, stance and hedging are also used in both genres to assess previous studies or objects of study in order to make sense of the case study being produced or the theory in discussion. When it comes to intermediate features, it is worth pointing out that the explanation or definition of objects of study is made through the use of *that-relative clauses,* while *to-clauses controlled by verbs of desire* are used when making recommendations in both genres.

From now on, we bring excerpts that exemplify how the features are used in both CR and CS. In the CR excerpts, the features are marked as follows: *nominalizations* are marked in italics (*argument*); *abstract nouns* in bold (**research)**, *stance nouns* between asterisk (*claim*), *attributive adjectives* between circumflexes (^nutritional^ treatment), *that- relative clauses* (elements that are similar), *that-verb complement clauses* (the authors state that) and *stance nouns + to/that clauses* (proposal that the researcher)

are underlined, and *coordinating conjunctions as clausal connectors* are between brackets (displayed by an individual, [and] to the mental structures).

Below, the excerpts are presented and followed by a discussion.

> On page 485 the authors state that they aim to argue against dryland farming. The first ***argument*** to support this is the pollen record that shows the **absence** of woodland *clearance*. Then, although the reader expects a second ***argument***, there is a ^long^ ^technical^ *narration* about alluvium **phases** at the end of which the point is made that Neolithic sites were located near a then ^active^ floodplain… (6006bCR)

In the excerpt above, the writer begins with a direct reference to a previous paper, a typical feature of CR, and reports the authors' ideas using the verbs "state", "aim" and "argue", followed by a *that-complement clause*, a *to-complement clause* and a *noun phrase*, respectively. The second sentence initiates the analysis of the argumentative structure of the text being evaluated, making use of a *relative clause* to explain the argument related to the "pollen record". The writer's stance can be recognized through the construction "although the reader expects" and the use of the *attributive adjective* in "long narration", which together show criticism of the text analyzed.

> Mayo and Jarvis referred to ***perception*** as "the ***process*** by which an *individual* selects, organizes, [and] interprets ***information*** to create a ^meaningful^ picture of the world." **Learning** is "**changes** in an *individual*'s **behavior** based on his **experiences**. ***Personality*** refers to 'the patterns of behavior displayed by an *individual*, [and] to the ^mental^ **structures** that relate **experience** and **behavior** in an orderly **way'**. **Motives** are described as 'the ^internal^ ^energizing^ **forces** that direct a person's **behavior** toward the ***achievement*** of ^personal^ **goals**. (3050bCR)

Even though in this second excerpt the same phenomenon of relying on a previous paper can be observed, the purpose here is to demonstrate comprehension of the concepts developed in previous work, another fundamental characteristic of Critiques (Gardner & Nesi, 2013). The concepts

definitions ("Personality refers to", "Motives are") are built with the use of *nominalizations* and *abstract nouns* ("perception", "personality", "experience", "achievement"), with *that-relative clauses* (defining "mental structure" and "energizing forces") and with coordination using "and" as conjunctions. These definitions corroborate that CR are a type of genre where students make sense of phenomena and claims in their disciplines (Nesi & Gardner, 2012: 37). Furthermore, this excerpt taken from a CR exemplifies what Ansarifar et al. (2018) and Staples et al. (2016) mention about academic writing being characterized by phrasal features rather than clausal ones. Considering that CR are expected to engage students in critical thinking more than CS, this argument is supported by the example above.

> But according to Scott […] It is important not to ignore ***experience*** [but] recognise its constructed **nature** and the **role** played by ***language*** and **discourse** (1992:25). In ***essence*** we need to be aware that ***experience*** is factual and socially constructed. ***Experience*** establishes the ***existence*** of *individuals* and operates within the ^ideological^ ***construction*** that makes the *individual* the ^starting^ point of **knowledge** (Scott, 1992:27). The ***argument*** about the **truth** in women's **account** only validates the *__claim__* by second wave *feminists* that women are ^different^. *Feminists* claim that using women's ***experience*** as a ^starting^ point is the only **option** left for feminist *researchers*. (0402cCR)

In this excerpt, the author shows his/her understanding of the theoretical context in which the research area - feminist research - is developed, which figures as one of the purposes of Critiques. We can easily notice that this discussion of theory demands the constant use of *abstract nouns* and *nominalizations*. The paraphrasing of one writer's ideas in the first part to validate the feminists' claim constructs the argument in support of the feminist point of view, emphasized by the sequence "only validates". The use of both the *likelihood noun* ("the claim […] that women are different") and the verb "claim" establishes a distance from the writer and a degree of uncertainty in relation to the propositions in the *relative* and in the *verb complement clause* ("claim that using women's experience […] is the only

option left") that follow. The first relative clause giving more information about "construction" is typical of this type of academic discourse.

> It is an ^enormous^ ***achievement*** that a project of such ***complexity*** was completed both on schedule and within **budget**… Another ^key^ ***success*** of the project was the company's ^final^ **expenditure** of only $470 million - approximately half of the originally allocated **budget**. This is indicative of ^effective^ **cost planning** on Eiffage's **behalf**. The ^low^ ^final^ **expenditure** also suggests the company may have had a rather ^large^ **Risk Budget** in place to deal with **uncertainty**. The ***fact*** that Eiffage were able to outbid the other three ^competing^ parties to build the bridge while maintaining such a ^sizeable^ **budget** demonstrates both ^effective^ **cost planning analysis** and **risk *identification*** and ***assessment***." (0177aCR)

The excerpt above evaluates the performance of a company in a construction project, one of the possible uses of the genre family Critique in BAWE. The purpose of the evaluation is clear in the use of the *attributive adjectives*, like in "enormous achievement" and "effective cost planning", as well as in the sequences with pre-qualifiers and adverbs, "rather large risk budget", "such a sizeable budget", "such complexity" and "only $470 million". The *relative clause* defining the "achievement" of the company, with the idioms "on schedule" and "within budget", suggest a more informal style, which is reinforced by the use of the emotive language also shown by the choice of adjectives used. On the other hand, the complement clause with the modal following the hedge verb "suggest" and the noun complement one initiated by "the fact that" both demonstrate a more distant style helping the writer to express his opinion in a less direct way.

For the CS excerpts, the features are marked as follows: *group nouns* are between asterisks (*laboratory*), *verb complement clauses controlled by verbs of desire* are underlined and the verb is bold (**prefer** to opt), *wh-relative clauses* are underlined (which indicate), *nominalizations* are in italics (*surgery*), *premodifying nouns* are between brackets ([work] hours), *nouns*

in bold (**patients**) and *attributive adjectives* between circumflexes (^nutritional^ treatment).

> Some **patients** **prefer** to opt for ***surgery*** at ***presentation*** rather than ^pharmacological^ or ^nutritional^ ***treatment*** of ^unknown^ ***duration***. Unfortunately, there is no ^controlled^ **data** to confirm the ^best^ **approach** for **patients**. There is a ^well-known^ ***association*** between ^ulcerative^ **colitis** and an increased **risk** of ^colorectal^ **cancer**, and **patients** with Crohn's **disease** are believed to be at ^increased^ **risk** of **cancer** of the ^small^ **intestine**. **Studies** have shown that the ^relative^ **risk** of ^colorectal^ **cancer** in **patients** with Crohn's **colitis** is approximately 5.6 and should raise the same **concerns** as in **patients** with ^ulcerative^ **colitis**. (0203iCS)

The excerpt above is from a Case Study and some of the highlighted features are different from the ones in the previous excerpts, which were taken from Critiques. As Nesi and Gardner (2012: 40) state, CS are used "to demonstrate/develop an understanding of professional practice through the analysis of a single exemplar" and are common in the Health area. In this part, the author discusses the possible treatment for a patient with Crohn's disease and the recommendation of surgery, a common stage of CS. The regular presence of *abstract nouns* and *nominalizations* is comparatively lower than what can be observed in the examples of CR, but *nouns* (in bold) are quite frequent. Moreover, in this example, there is no explicit judgment on the previous studies of the area; instead, the writer refrains from being conclusive about recommended treatment with the use of the *stance verbs* in "studies have shown", "patients with Crohn's disease are believed to be" and "the relative risk […] should raise the same concerns". Here it is worth pointing out that, although not quantitatively analyzed, *stance verbs* appear in both CR and CS, thus being a feature in common.

> The ^above^ **guidelines** therefore suggest that Mr's **ischaemia** was most likely due to an ^embolic^ **event**. The **fact** that capillary **refill** occurred, albeit delayed, in the ^right^ **foot** suggests that either the ***obstruction*** was incomplete or that some ***collaterals***

> were available in order to maintain ***perfusion***. Although it is impossible to determine the ^precise^ **cause** of the **emboli** without further ***investigation***, there was no ***evidence*** to suggest the ***presence*** of ^atrial^ ***fibrillation*** or an ^abdominal^ ^aortic^ **aneurysm**." (0047dCS)

In this excerpt, there is a clear reference to a particular case and the writer relies heavily on the hedge verb "suggest" followed by *that-verb complement clauses* or *noun phrases*. The clause "that either the obstruction was incomplete or that some collaterals were available" is used to present possible reasons for the occurrence of capillary refill. The technical phrases with *attributive adjectives* and *nouns*, such as in "embolic event", "right foot" and "atrial fibrillation" confirm Staples et al.'s (2016) claim that *attributive adjectives* are frequently used in CS to help the description of technical terms, as it is very common in Life and Physical Sciences" (p. 169), which is also the case in assignments written by graduate students in BAWE.

> In ^traditional^ ^hierarchical^ ***organization***, ***promotion*** is one of the most ^useful^ HR **policies**. In Oticon, **people** may be [project] ***leaders*** for several **times**, but they are seldom promoted. So the **organization** **need** to use other **methods** to compensate, like arranging tailored [training] **program** for ^excellent^ **staffs**, or giving them the ***freedom*** to choose ^suitable^ **tasks**, [**work**] **hours**, **place** of **work** and so on… (0166aCS)

This analysis of one company's HR policies shows the writer's knowledge about the professional practices in his area and his capacity to analyze and evaluate them. We can see here the use of the two features that were significantly more frequent in CS than in CR, *group* or *institution nouns* - "organization" - and *to-clauses* following verbs of desire – "need to use". In the first case, it is expected that institutions like companies or hospitals are mentioned in these texts as they describe events and practices typical of these contexts. As for the verbs of desire, they appear to be very useful for the recommendations part of CS.

> There are several ^salient^ **features** of the **history** which indicate ^Infective^ **Endocarditis** (IE) as the most ^likely^ **cause** for Mr's **symptoms**. […] In ***addition***, the **onset** of [**loin**] **pain** may be due to **splenomegaly** or ^immune-complex^ ***deposition***, which is commonly seen in IE. (0047fCS)

In this excerpt, there is a discussion on the medical history of a patient that suggests he might be suffering from a disease, namely Infective Endocarditis (IE); this understanding is conveyed through the use of the *wh-relative clause* starting with "which indicate". Besides, the following sentence introduced by "in addition" gives a further explanation of the symptoms and problems that are a result of IE. As can be observed, there are occurrences of *nouns* ("onset", "pain", "splenomegaly"), *premodifying nouns* ("loin") and *attributive adjectives* ("immune-complex") to explain and detail the disease under discussion.

Based on these excerpts, it is worth highlighting that the uses of the linguistic features match the purposes of both genres analyzed; CR evaluates and reviews an object of study while CS demonstrates understanding of professional practices (Gardner & Nesi, 2013). The first CR excerpt illustrates references and evaluations of an object of study; in the second one the author makes sense of a specific object of study and the evaluation is not direct; the same phenomenon happens in the third example in which instances of stance utterances are observed; the fourth CR excerpt presents an evaluation of a business issue.

When it comes to the CS excerpts, the purposes of this genre can be observed as well in the use of the analyzed features. In the first excerpt, there is an understanding of professional practice in the health field as well as the use of hedge in recommending a treatment; the second one brings even more distancing from explicit judgements at the same time as it discusses a specific case; the third excerpt also shows recommendations but in the business area.

Overall, it is possible to affirm that the features under investigation are used in order to convey particular meanings in each of the two genres examined. This means that the way the linguistic features selected for this

study are used showed, above all, that they are at the service of the communicative purposes of Critiques and Case Studies, as described by Nesi and Gardner (2012).

Conclusion

The objective of this paper was to analyze academic language features in two university student genres, under the assumption that academic writing cannot be regarded as a homogeneous group. Thus, the first research question asked to what extent there is linguistic variation in the comparison between Case Studies and Critiques. The results of Mann-Whitney U tests revealed that most of the linguistic features analyzed appeared with a higher incidence in CR than in CS, possibly due to the distinct purposes of each genre. In CS, students show knowledge of a professional practice, whereas in CR they must demonstrate informed and independent reasoning as well as understanding of the object of study while evaluating and/or assessing its importance.

Regarding the second research question, which involved the understanding of the grammar features observed in this variation, some patterns emerged, leading to the diverse usages of the same features according to the communicative objectives of each genre. For instance, *attributive adjectives* are used in both genres, but in CS it is possible to observe that this feature helps in the description of issues, diseases or products, whereas in CR they are commonly used to evaluate previous studies or to present and discuss theoretical concepts.

As reported in the quantitative results, most of the differences between the two genres were not statistically significant, with the same features being used in both; sometimes for the same purpose, as in the use of *stance features*; others to convey different purposes in each genre, as is the case of *attributive adjectives* previously mentioned. Below, we summarize what some usages highlighted in the excerpts extracted from the corpus might suggest:

- Attributive adjectives appear to be helping the descriptions in CS, such as patients' issues and diseases, or products in the business area, by specifying technical terms; in CR they support the evaluation of previous studies.
- Abstract nouns and nominalizations are very frequent in both genres and seem to be very useful to present and discuss theoretical concepts in CR.
- Group nouns referring to institutions, companies or universities are considerably more frequent in CS as this genre describes situations or products that happen in or are produced by these institutions.
- Stance features and hedging are present in both genres but are very recurrent in CR to assess previous studies or one specific object of study.
- *That*-relative clauses and noun complement clauses controlled by nouns of likelihood are more common in CR and are used to explain or define the object of study.
- *To*-clauses controlled by verbs of desire appear to help the recommendations of CS.

This study has the limitation of analyzing a restricted set of linguistic features and not a comprehensive amount of what has been proposed in the grammatical complexity literature. As suggestions for future studies, besides including more grammatical features in the investigation, it would be of great value to expand the variety of genres under analysis in order to explore other uses of the same features and others that were not included in this study and that might contribute to the description of how these features are employed in different genres. Also, it would be interesting to include different levels of university study, apart from undergraduate and graduate first years, in order to account for language development in academic writing.

Although with a small scope, the results of this investigation might contribute to the understanding of how the expression of diverse communicative objectives is built in academic writing through its various genres.

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

Berber Sardinha, T. (2014). 25 years later. In Berber Sardinha, T.; Veirano Pinto, M. (eds.) *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*. (pp. 81-108). John Benjamins.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511621024

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15, 133–163. doi:10.1080/01638539209544806.

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. John Benjamins.

Biber, D. (2012). Register as a predictor of linguistic variation, *Corpus Linguistics and Linguistic Theory*, *8*(1), 9-37. doi: https://doi.org/10.1515/cllt-2012-0002

Biber, D. & Conrad, S. (2009). *Register, Genre, and Styl*e. Cambridge University Press.

Biber, D. & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT Test: A lexico-grammatical analysis (*TOEFL iBT Research Report iBT-19*). Princeton, NJ: Educational Testing Service.

Biber, D. & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

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

Biber, D., Gray, B. & Staples, S. (2016). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels, *Applied Linguistics*, 37 (5), October, 639–668, https://doi.org/10.1093/applin/amu059

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London, England: Longman.

Biber, D., Gray, B., Staples, S. & Egbert, J. (2021). Theoretical and Descriptive Linguistic Foundation of the Register-Functional Approach to Grammatical Complexity. In: Biber, D., Gray, B., Staples, S. & Egbert, J. *The Register-Functional Approach to Grammatical Complexity* (pp. 6-22). Routledge.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Bulté, B. & Housen, A. (2018). Conceptualizing and measuring syntactic diversity. *International Journal of Applied Linguistics*, 28, 147–164. https://doi.org/10.1111/ijal.12196

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203771587

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics,* 34(1), 25-52.

Gardner, S., Nesi, H. & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, 40(4), 646-674.

Goulart, L. (2020). Analyzing the patterns of lexico-grammatical complexity across Graded Reader levels. *Reading in a Foreign Language*, 32(2), 83-103. http://hdl.handle.net/10125/67375

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. Amsterdam, Netherlands: John Benjamins.

Hardy, J. A. & Friginal, E. (2016). Genre variation in student writing: A multi-dimensional analysis. *Journal of English for Academic Purposes*, *22*, 119-131.

Kuiken, F. & Vedder, I. (2019). Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics*. Special Issue. https://doi.org/10.1111/ijal.12256

Mangiafico, S.S. (2015). *An R Companion for the Handbook of Biological Statistics*, version 1.3.2. rcompanion.org/rcompanion/.

Nesi, H. & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge, UK: Cambridge University Press. Available from http://bit.ly/slWnd5

Nesi, H., Gardner, S., Thompson, P. & Wickens, P. (2008-2010). *The British Academic Written English (BAWE) corpus.* Available from: https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Staples, S., Biber, D. & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal* vol 102/2, pp. 310-332. https://doi.org/10.1111/modl.12465

Staples, S. & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, 32, 17e35. https://doi.org/10.1016/j.jslw.2016.02.002.

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, 149–183. https://doi.org/10.1177/0741088316631527

Investigating Brazilian English Learners' Use of Academic Collocations: A Corpus-Based Study

Marine Laísa Matte (UFRGS/IFSul)
Simone Sarmento (UFRGS)

Introduction

Writing has a special role in academic contexts as it is one of the main skills students have to master in order to achieve academic success (Biber & Gray, 2016). In addition, at the higher education (HE) level, academic literacies are being learned and tested all the time. New ways of constructing knowledge are constantly being discovered (Lea & Street, 1998), and these practices are necessarily dependent upon academic writing. In spite of the importance writing plays in academic contexts, it is usually assumed that students should already know the rules and conventions of this practice. However, these rules are not transparent, forming what Lillis (2001) called "practices of mystery". For Lillis, students who are not familiar with academic writing conventions may have their participation in HE impaired. Thus, these conventions should be explicitly taught since we cannot depend on incidental learning or on a hidden curriculum, as students "must now gain fluency in the conventions of English language academic discourses to understand their disciplines and to successfully navigate their learning." (Hamp-Lyons, 2002: 1)

Academic language is a specific subset of general language, and differs considerably from the type of language used in daily life situations, not only in terms of formality but also in terms of language features (Simpson-Vlach & Ellis, 2010). The language features specific to academic contexts may range from, for instance, choice of verb to combination of words, i.e. collocations. Collocations are also important in general language, but

gain even more importance in academic registers. According to Sinclair (1991: 110), any language user will have a repertoire of "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." In other words, proficient language users resort to collocations to convey meaning. Therefore, mastering collocations is imperative for guaranteeing fluency in a text, as writing proper academic English goes beyond knowing isolated words. When it comes to judging text quality, one of the criteria a reader has in mind, however unconsciously, is how conventionalized language is. This conventionality is partly guaranteed by the appropriate use of collocations.

Bearing the importance of collocations for academic texts in mind, the main goal of this study is to analyze how Brazilian students produce collocations in academic texts written in English by comparing two corpora of unpublished texts: one with texts produced by Brazilians studying in British universities (BrAWE) whose grades are unknown, and the reference corpus with texts written by students from multiple nationalities studying in British universities but which were graded with merit or distinction (BAWE). The latter will be used as baseline data. The following research questions will be addressed:

a) Is there a statistically significant difference in the frequency of the noun nodes and their respective collocates in BrAWE and BAWE?
b) Are there differences in syntactic structures of collocations between the two corpora?

Collocations

"You shall know a word by the company it keeps" is probably a sentence that immediately comes to mind of anyone acquainted with collocational studies. This sentence formulated by J. Firth (1957: 11) has inspired a great deal of research in the field, as it summarizes the core meaning of collocations, i.e., the likelihood of two or more words occurring together ((Sinclair, 1991; Hill, 1999; Durrant, 2009). Sinclair (1991) proposed the idea that language operates according to the open-choice principle and the idiom principle. The former considers language as the result of complex

choices to complete each unit (word, phrase and clause) that composes a text, i.e., all slots of a text can be filled with any word as long as grammaticality is preserved. The latter assumes that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments." (Sinclair, 1991: 110).

Regarding language learners, evidence shows that they do make use of collocations but tend to have a more limited repertoire of conventional combinations (Granger, 1998; Lorenz, 1999; Nesselhauf, 2005). The comparison between native (NS) and non-native (NNS)[1] collocational performance is presented in Howarth (1998), who analyzes adult learners of English writing academically in Social Sciences postgraduate courses and focuses on the use of collocations composed of verb + noun. The study reveals that the NNS "produced, on average, a much lower density of conventional combinations (25%), suggesting either a generally lower level of knowledge of collocations, or a lack of awareness of how to deploy them appropriately, or both." (Howarth, 1998: 36).

Granger (1998) analyzes intensifying adverbs ending in –ly that function as amplifiers and modifiers as the nodes of the collocations. By comparing a corpus of native English writers to a similar corpus of advanced French-speaking learners of English, the data revealed a statistically significant overall underuse of amplifiers in the learner corpus. However, when looking at some amplifiers individually, *completely* and *totally* were overused by the learners, while *highly* was underused. Granger suggests that this overuse can possibly be explained by the fact that these adverbs have direct equivalents in French and, consequently, students choose to translate them from French into English. Additionally, some combinations with amplifiers such as *acutely aware*, *bitterly disillusioned*, *gravely disorganised*, and *steeply dipping* are used exclusively by native speakers.

1 It is important to point out that most studies related to proper use of collocations rely on a contrastive analysis between native speakers (NS) and non-native speakers (NNS). However, in this study the comparison was not based on a NS vs. NNS dichotomy.

Collocations composed of adjective + noun or noun + noun are analyzed by Durrant and Schmitt (2009). The authors analyze a total of 96 texts organized in two sets: one containing NNS texts and the other NS texts. By classifying collocations into low-frequency and high-frequency and establishing collocational strength with t-score and Mutual Information measures[2], they came to three main findings: Firstly, native writers use more low-frequency combinations than non-natives. […] Secondly, non-native writers make at least as much use of collocations with very high t-scores as do natives. […] Thirdly, non-native writers significantly underuse collocations with high mutual information (MI)[3] scores in comparison with native norms (Durrant & Schmitt, 2009). These findings suggest that learners have a tendency to repeat favored items, as they quickly pick up frequent collocations because the less frequent and strongly associated items take longer to acquire (Durrant & Schmitt, 2009). Ellis, Simpson-Vlach and Maynard (2008) reinforce this idea that NS use a wider range of collocations, whereas NNS tend to use collocations they encounter more frequently. The issue of overusing collocations is discussed by Ackerman and Chen (2013: 3), who argue that "by using a less appropriate collocate, a non-native speaker will sound unnatural or may even become unintelligible among speakers of the target language."

Laufer and Waldman (2011) investigate *verb + noun* collocations produced by L1-Hebrew learners of English. Besides comparing the learner corpus to a NS one, the authors also compared the data within L1 Hebrew learners of English represented in the corpus. Results indicated that the NS produced almost twice as many collocations as the learners. Learners underused verb + noun collocations when compared to NS and produced significantly more deviant collocations. Advanced and intermediate learners

2 The t-score is an association measure that "highlights frequent combinations of words. [H]owever while all collocations identified by the t-score are frequent, not all frequent word combinations have a high t-score. [On the other hand], MI-score is negatively linked to frequency, meaning that the value is larger the more exclusively the two words are associated and the rarer the combination is." (Gablasova et al., 2017: 8-9).

3 MI is a measure of association between words. The higher the MI score, the stronger the relation between the items (Church & Hanks, 1990)

were the ones who produced more deviant collocations, probably because they feel more confident in relation to the English language when compared to basic students.

Chinese learners of English and their use of collocations in academic written texts were investigated by Wu (2016). The author analyzed verb + adverb and adverb + verb collocations comparing three academic English corpora, two of NS and one of NNs. Wu (2016) also shows that there are significant differences in terms of collocations chosen by Chinese learners of English who use, for instance, *develop quickly*, *widely use* and *abolish completely* more frequently than NS do. This difference regarding lexical competence and knowledge of collocations might be related to the fact that the teaching of collocations is not common in China, and that Mandarin and English have only few similarities.

Ohlrogge (2009) analyzed 170 written compositions written for an EFL proficiency test and found correlations between level of proficiency and collocations. Hence, the students who received higher grades presented a higher incidence of collocations. This follows what Crossley et al (2015) state regarding the relation between proficiency and collocations. After having investigated lexical proficiency in both oral and written texts produced by learners of three different levels (beginning, intermediate and advanced), raters judged the lexical proficiency according to analytical and holistic features, one of them being collocations. Results indicate that higher proficiency writers tend to use a wider range of collocations than lower proficiency writers, corroborating what was found in our study.

When it comes to the analysis of collocations used by Brazilian learners of English in academic genres, more specifically in argumentative essays, Guedes (2017) explored *verb + adverbs* ending in *-ly* collocations. The author found that the most common verbs used by the learners are action verbs (*apply* and *provide*). Also, there is a high frequency of verbs such as *improve*, *develop*, and *adopt* among learners of English. On the other hand, verbs such as *increase*, *include*, *occur*, *reduce*, and *require* are more frequent in BAWE. Due to the low frequency of *verb + adverbs* ending in *-ly* their collocational strength could not be statistically measured.

Matte and Rebechi (2019) analyzed the differences in the use of collocations of the Academic Collocation list (ACL)[4] (Ackermann & Chen, 2013) in the same corpora used in the present study. Their results show that only a few collocations of ACL are used differently in the comparative analysis of BAWE and BrAWE. Furthermore, the most frequent collocations in both corpora are not exactly the same presented in the list, which suggest a possible mismatch between what is presented in ACL and authentic language produced by students both in BrAWE and BAWE.

There are ready-made lists containing relevant collocations and formulas to be mastered, as those presented in the ACL (Ackermann & Chen, 2013) and the Academic Formulas List (AFL) (Simpson-Vlach & Ellis, 2010). However, despite the "progression in research from studies that provide evidence of the importance of collocations for L2 learners" (Boers & Webb, 2018), it is necessary to create pedagogical materials that fit students' needs. Thus, more than memorizing vocabulary and collocation lists, it is imperative to master collocations in terms of knowing their appropriate use, that is, collocational competence must be acquired in context. This argument is sustained by Frankenberg-Garcia (2018: 101), who points out that "the lexical knowledge is not just about understanding words, but also about employing words in context.".

The corpora

The BAWE corpus (Alsop & Nesi, 2009) was compiled with the objective of gathering unpublished written assignments from students of multiple nationalities studying[5] in four different British universities: Warwick University, Reading University, Oxford Brookes University, and Coventry University. Unlike other academic corpora that are mostly composed of texts written by experts and edited by professionals, the BAWE is composed of discipline-specific learner texts. Despite containing students' writing, this corpus is different from those compiled with essays written under

4 https://www.eapfoundation.com/vocab/academic/acl/

5 BAWE contains texts of undergraduate and master's students.

examination conditions for analyzing non-native-speaker error and language acquisition, as it contains assignments written during undergraduate and master courses which were graded merit or distinction. The BAWE corpus was, thus, designed to enable the investigation of academic literacy and disciplinary knowledge development. BAWE has 6,968,089 words and it is balanced into four areas[6]: Life Sciences (LS), Social Sciences (SS), Physical Sciences (PS), and Arts and Humanities (AH). Each area encompasses a variety of disciplines. Moreover, the corpus is organized according to 13 different academic genre families proposed by Gardner and Nesi (2013). A total of 2,858 texts were compiled, being 1,953 written by L1 speakers of English and the remainder by highly proficient English as an additional language (EAL) students.

The Brazilian version of BAWE is BrAWE (the Brazilian Academic Written English corpus) compiled by Goulart (2017). The organization of the corpus is similar to the British one, as it covers the same areas of expertise and gathers assignments produced by undergraduate students. Therefore, BrAWE also follows Gardner and Nesi's (2013) classification of academic genre families, but only 12 categories were found. The final version of the corpus contains 380 assignments of students from 59 universities. The high number of universities involved is due to the fact that most of the students were participants of the Sciences without Borders (SwB) program, which partnered with over 80 universities in the United Kingdom alone. The SwB was a Brazilian scientific mobility program created in 2011 with the objective of strengthening and expanding the internationalization of Brazilian higher education by providing scholarships for both students and researchers.

Overall, engineering, natural sciences, and health sciences were the areas covered by the SwB. Areas such as arts and humanities were not contemplated by the program, but some texts from this area were included in the corpus as some students from other mobility programs were also contacted. Despite being comparable to BAWE, the corpus is unbalanced in terms of size of subcorpora. Considering that Life Sciences (LS), Social

6 Alsop and Nesi (2009) refer to these areas as disciplinary groups.

Sciences (SS) and Physical Sciences (PS) are the most representative areas in BrAWE, a subcorpus of BAWE was created in order to make it comparable to the BrAWE corpus. Thus, whenever BAWE is mentioned, we are referring to BAWE's subcorpora that contain only assignments in the fields of LS, SS, and PS.

	BAWE	**BrAWE**
Words	3,312,196	768,323[7]
Number of assignments	2,761	380
Quality of assignments	Merit and distinction	Passing (and higher)

Table 1. BAWE and BrAWE corpora

As stated above, the attested quality of assignments distinguishes BAWE and BrAWE. In BAWE, students were attributed merit and distinction, whereas in BrAWE students may have obtained a passing grade by the minimum requirement, which does not necessarily mean that no one wrote outstanding texts. Although grades were not given because of the quality of language, one can speculate that language may indeed play an important part in the quality of an assignment. According to Kumar and Rao (2018: p. 9), "poor academic writing skills and lack of command over the knowledge of English language" feature among the reasons why manuscripts are rejected. Therefore, due to the quality of texts, and to the high level of English language proficiency of participants, BAWE may be considered an adequate reference corpus to fulfill the purposes of a contrastive corpus analysis.

Methodological procedures

Collocations can be analyzed according to the frequency of the words or to the strength of association between the composing words using statistical measures, such as MI, t-score, Log Dice (Brezina, 2018). In this study,

7 The size of BrAWE in Sketch Engine is 768,323 rather than 670,314, as shown in Table 3, because this software counts punctuation marks as words.

we used Log Dice to calculate the strength of association between words since this is the default statistical measure of Sketch Engine, the software used to extract the collocations.

Three different types of collocations[8] were investigated: modifier + noun, noun (subject) + verb, and verb + noun (object). For example,

- Modifier: adjectives that come before the node
  Ex.: *difficult* + *task*, *advanced* + *technique*

- Verb (object of): used when the node is the object of the verb
  Ex.: *execute* + *task*, *apply* + *technique*

- Verb (subject of): used when the node is the subject of the verb
  Ex.: *task* + *require*, *technique* + *use*

These categories of collocates follow Frankenberg-Garcia et al.'s list (2018) composed of 187 collocational nodes which is a merge of three lists: the Academic Vocabulary List[9] (AVL, Gardner & Davies, 2014), the Academic Keyword List[10] (AKL, Paquot, 2010), and the Academic Collocations List (ACL, Ackermann & Chen, 2013). Of these 187 nodes 125 are nouns, 38 are verbs, and the remaining 24 are adjectives.

We focused on the identification of overused and underused academic noun-node collocations, through the comparison of two different corpora, the British Academic Written English corpus (BAWE) and the Brazilian Academic Written English corpus (BrAWE). The cut-off point to include a collocation in the study was a minimum frequency of four occurrences in BAWE in at least two out of the three remaining areas, i.e. Life Sciences, Health Sciences, and Social Sciences. Thus, collocations of one-single area were not included, as it is the case of *health need*, a collocation that only appears in LS assignments. The five methodological steps were:

8 The main word of a collocation is called node, and the ones associated to the node are the collocates. Thus, the basic structure of a collocation is node + collocate.

9 Derived from BAWE.

10 https://uclouvain.be/en/research-institutes/ilc/cecl/academic-keyword-list.html

1st: Listing in descending order the 125 nouns from the Frankenberg-Garcia et al.'s list (2018) from the most to the least frequent in BAWE by using the "search" tool in Sketch Engine[11]. The node was typed in the "lemma" box and the PoS noun was selected. All the words that derive from the base form of the node came up as a result, for example for *approach*, the plural form – *approaches* – was also selected. This procedure was repeated for every noun, i.e., for the 125 nodes.

2nd: Extracting the collocates of the 125 nodes in both corpora using the "Word Sketch" tool. The following syntactic structures mattered to this study: *(different + approach)*, *object of* (verb) (*use + approach)*, and *subject of* (verb) (*approach + involve)*. Again, the node was typed in the "lemma" box in "word sketch", and the PoS – noun was selected.

3rd: Calculating the Log Likelihood (LL) value (Rayson, 2002) for the different frequencies of each one of the 125 nodes in both corpora. If the outcome of the statistical test is 6.63 or higher, there is a 99% chance that the results are not random ($p<0.01$).

4th: Calculating the statistical significance of the collocates using LL to determine whether the comparison of frequencies of the collocates of each individual noun in both corpora was statistically significant ($p<0.01$). The frequencies of each collocate were verified in both corpora, and the LL value was calculated.

5th: Verifying the syntactic structure of the collocates that go together with each of the 125 nodes in order to check if different patterns emerge in the comparison between both corpora.

Results and discussion

The most frequent of the 125 nodes in both corpora is *system* (1.38 per thousand words in BAWE and 1.60 per thousand words in BrAWE) and the node with the lowest frequency is *exception* (0.03 in BAWE and 0.02 in

11 "The Sketch Engine is a corpus query system which allows the user to view word sketches, thesaurally similar words, and 'sketch differences'" (Kilgarriff et al., 2004). Word sketches, the products of the "Word Sketch" tool, are summaries of the grammatical and collocational behavior of a word.

BrAWE) in both corpora too. From these 125 nodes, 36 are used with a similar frequency in both corpora, whereas 89 are used in a statistically different fashion based on the LL ratio. From these, 48 were underused in BrAWE (marked with **), while the remaining 41 were overused (marked with *) when compared to the reference corpus, BAWE. The complete data can be found in Table 2.

Table 2. Raw frequency and normalized values of the 125 nodes in both corpora

According to the table presented in appendix 1, we can observe that there are 2,679 collocates for the 125 nodes in BAWE. One exception is the node *contrast*, that does not have any collocate according to our cut off point. In BrAWE, there are only 1,015 collocates for the same 125 nodes, and there are no collocates for six of the 125 nodes (*contrast*, *exception*, *reference*, *attempt*, *tendency*, and *alternative*). Thus, there is a difference of 1,664 between the total number of collocates in BAWE as compared to BrAWE, showing a low density of conventional combinations in the corpus of Brazilian students.

The 125 nouns portray 287 collocates which show a statistically significant different use when comparing both corpora, being 190 underused and 97 overused, as shown below:

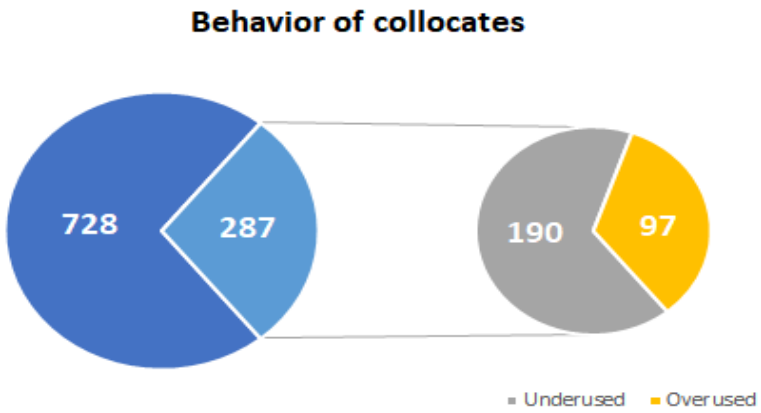


Figure 2. Behavior of collocates

Out of these 287 collocates, 202 are modifiers, 76 are verbs that collocate with nodes in the object position, and the remaining nine are nodes in the subject position. The types of collocates that go along with the 125 nodes in each corpus are displayed in Table 3.

	BAWE	**BrAWE**
Modifiers *whole system, final result*	1,359 (50.7%)	506 (49.8%)
Verb (object) *make process, conduct research*	1,049 (39.1%)	444 (46.7%)
Verb (subject) *result show, strategy include*	271 (10.1%)	65 (6.4%)
TOTAL	***2,679***	***1,015***

Table 3. Types of collocates used in each corpus

The results above account for both the variety and types of collocates of the 125 nodes (nouns) under analysis. Among the three categories, modifiers, i. e. words that occupy a position before the node, account for roughly half of the occurrences in both corpora (50.7% and 49.8% in BAWE and BrAWE respectively). Some examples are *whole system* and *final result*, in which *whole* and *final* are the modifiers and *system* and *result* are the

nodes. Nodes as objects are preferred in BrAWE (46,7%) as compared to BAWE (39.1%), as in *make process* and *conduct research*, with *make* and *conduct* being the verbs when the nodes *process* and *research* are the objects. Conversely, nodes as subjects are more frequent in BAWE (10.1%) than in BrAWE (6.4%) as in *result shows* and *strategy includes*, in which *result* and *strategy* are the subjects and are followed by the verbs *show* and *include* respectively.

When comparing collocations composed of the nodes with statistically significant differences (the overused nodes and the ones composed of the underused nodes), it is possible to observe a balance in terms of syntactic structures in BAWE and in BrAWE, as shown below:

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
Overused nodes	562 (50.4%)	423 (37.9%)	130 (11.6%)	219 (48.2%)	196 (43.1%)	39 (8.6%)
TOTAL	*1115*			*454*		
Underused nodes	468 (51.03%)	355 (38.7%)	94 (10.2%)	147 (49.3%)	134 (44.9%)	17 (5.7%)
TOTAL	*917*			*298*		

Table 4. Syntactic structures of collocations in both corpora

Modifiers that precede the nodes are the most productive ones, with 50.4% and 48.2% of occurrences in BAWE and BrAWE with the overused nodes, and 51.03% and 49.3% in BAWE and BrAWE with the underused nodes. Subsequently, verb + node (object) collocations have the second highest percentage of occurrences, with 37.9% in BAWE with overused nodes and 43.1% in BrAWE with the same nodes. When it comes to the underused nodes, the percentages are 38.7% and 44.9% in BAWE and in BrAWE respectively. Node (subject) + verb collocations account for the lowest percentages with both overused and underused nodes: 11.6% and 10.2% in BAWE, and 8.6% and 5.7% in BrAWE.

When analyzing the LL values of the nodes, there is a bigger difference in the range of LL values of the underused nodes than with the overused

ones. Table 5 illustrates the LL values of the nodes with the most significant differences in the comparison between both corpora. Considering that BAWE is the reference corpus, the terms overused and underused refer to the uses in BrAWE:

	Overused	**Underused**
Lowest LL	*Factor* (7.06)	*Difficulty* (-6.84)
Highest LL	*Example* (370.55)	*Data* (-615.78)

Table 5. Lowest and highest LL

Higher LL values indicate that the differences between the frequency scores are more significant (Rayson, 2002). Table 6 shows collocations with the node *data* (the underused node with the highest LL) in both corpora. Differences can be observed not only in the total number of collocations (55 in BAWE vs. 10 in BrAWE), but also in the syntactic patterns, since 90% of the words that collocate with *data* in BrAWE are verbs, as compared to 63,6% in BAWE.

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
data	20 (36.3%)	23 (41.8%)	12 (21.8%)	1 (10%)	7 (70%)	2 (20%)
TOTAL	55			10		

Table 6. Collocations with the node *data*

While 20 different modifiers[12] collocate with *data* in BAWE, in BrAWE the only modifier is "raw". A possible explanation is that the assignments which compose the BAWE corpus are mostly evidence-based studies, justifying the higher use of *data*. We can also speculate that Brazilian students prefer not to characterize the type of data under analysis by using the word individually rather than as part of a collocation. When it comes

12 *experimental, empirical, quantitative, historical, available, raw, recent, sample, past, primary, following, financial, other, survey, character, relevant, personal, important, actual, old.*

to the verbs that combine with *data*, regardless of whether the node is the object or the subject, the differences continue to be significant. Table 7 demonstrates the different behaviors:

	BAWE	**BrAWE**
Verb (object)	use, collect, obtain, show, analyse, contain, provide, give, record, gather, transmit, compare, present, produce, take, require, store, plot, interpret, send, receive, need, fit	collect, obtain, show, transmit, store, plot, need
Verb (subject)	show, suggest, use, collect, follow, gather, link, seem, demonstrate, support, indicate, exist	show, seem

Table 7. Verbs that collocate with *data*

Among the nodes with statistically significant differences, *difficulty* is the underused node with the lowest LL (-6.84). This means that overall *difficulty* is underused in BrAWE in comparison to BAWE. Table 8 portrays the syntactic structures of the collocations with this node.

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
difficulty	7 (46.6%)	8 (53.3%)	0	3 (37.5%)	5 (62.5%)	0
TOTAL	15			8		

Table 8. Collocations with the node *difficulty*

In total, there are 15 different collocations in BAWE and eight in BrAWE, with collocates in the modifier and verb (object) categories. While seven different modifiers collocate before the node in BAWE, only three are produced by Brazilians. As for the verbs that accompany the node when it is the object, eight go together with *difficulty* in BAWE whereas five are used in BrAWE, as shown in table 9:

	BAWE	BrAWE
Modifier	Great, technical, financial, main, economic, other	Great, main, other
Verb (object)	Face, cause, encounter, experience, pose, highlight, create	Face, cause, highlight, create

Table 9. Types of collocates with *difficulty*

Conclusion

This corpus-based study aimed to unveil the use of collocations by Brazilians studying in British universities. To that end, a comparative analysis of collocations of the Brazilian Academic Written English Corpus (BrAWE; Goulart, 2017) and the British Academic Written English (BAWE; Alsop & Nesi, 2009) was conducted.

Regarding the first research question *Is there a statistically significant difference in the frequency of the noun nodes and their respective collocates in BAWE and BrAWE?*, it is possible to state that from the 125 nodes analyzed, 36 have a similar frequency in both corpora, 48 were underused and 41 were overused in BrAWE. When it comes to the collocates, the 125 nodes produced 2,679 collocates in BAWE that met our inclusion criteria. In BrAWE, only 1,015 collocates occur with the same 125 nodes. Out of these collocates, 287 came up as having a statistically significant difference in use while analyzing the behavior of the 125 nouns, being 190 underused by Brazilians and 97 overused.

As for the second research question, *Are there differences in syntactic structures of collocations between the two corpora*?, the data revealed that from the 287 collocates which presented significant differences, 202 are modifiers, 76 are verbs in the object position, and nine are verbs in the subject position. In both corpora modifiers account for half of the occurrences (50.7% and 49.8% in BAWE and BrAWE respectively). Nodes as objects are more frequent in BrAWE (46,7%) as compared to BAWE (39.1%), whereas nodes as subjects are more preferred in BAWE (10.1%) than in BrAWE (6.4%). This discrepancy might be related to the type of study conducted by Brazilian students and to how proficient they are to employ different types

of verbs when nodes are used as subjects. For instance, studies conducted by students who wrote texts that compose BrAWE may be of different nature, thus the need to use a verb that best combines with the studies itself (make process, conduct research). On the other hand, when choosing verbs that are used after the node (subject of the sentence), their repertoire is narrower.

Based on the comparison of the two corpora used in this study – BAWE and BrAWE – we noted that academic collocations do not seem to be fully mastered by Brazilian students who write academic texts. For Sinclair (1991), learners operate more on the open choice principle than on the idiom principle, producing fewer collocations or collocations that do not sound natural. This lack of collocational competence was observed in the reduced number of collocations in BrAWE (1,015) when compared to BAWE (2,679) and in the number of outcomes that came up with statistically significant differences in the comparison between the data in the studied corpora. A node that illustrates this phenomenon is data, as displayed in Tables 6 and 7, in which it is possible to observe that the number of collocates used with data is significantly smaller in BrAWE than in BAWE.

The findings of this study suggest that Brazilian students have a limited variety of vocabulary as long as collocations are concerned. It is our belief that proper use of collocations is a major element in academic writing and should, thus, be treated as such in English teaching environments (AlHassan & Wood, 2015; Li & Schmitt, 2009; Martinez & Schmitt, 2012). For instance, the ones which are underused in BrAWE, such as *design* + *system*, *measured* + *value*, *good* + *value*, *decision-making* + *process*, *detailed* + *analysis*, *further* + *analysis*, *empirical* + *data*, and *quantitative* + *data* should be addressed with Brazilian students.

As pointed out by Hyland and Hamp-Lyons (2002: 10), "EAP offers the possibility of making even greater contributions to our understanding of the varied ways language is used in academic communities to provide even more strongly informed foundations for pedagogic materials." Some suggestions are given by Nesselhauf (2005: 253), for whom teaching collocations should begin with making students aware of this phenomenon. AlHassan and Wood (2005) also support the idea that a focus on formulaic

sequences in teaching reveals a development in L2 writing proficiency. Thus, a large repertoire of academic collocations improves students' writing, making it more formulaic and fluent, as formulaic sequences (such as collocations) provide fluency and conventionality to the language.

Considering that more information on the use of collocation by academic English learners would help us to establish a greater degree of accuracy on this matter, a natural progression of this work would be to thoroughly analyze and describe the collocates of all 125 nodes.

Acknowledgements

Marine Laísa Matte would like to thank CAPES for the financial support during her Masters. Simone Sarmento holds a CNPq research productivity scholarship level 1D.

References

Ackermann, K. & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235-247.

AlHassan, L. & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, 17, 51-62.

Alsop, S. & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1), 71-83.

Biber, D. & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Boers, F. & Webb, S. (2018). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, 51(1), 77-89.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), 22-29.

Crossley, S. A., Salsbury, T. & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36(5), 570-590.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, 47(2), 157-177.

Ellis, N. C., Simpson-Vlach, R. I. T. A. & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, 42(3), 375-396.

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (pp. 1-31). Special Volume of the Philological Society. Oxford: Blackwell. [Reprinted as Firth (1968)]Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, 35, 93-104.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P. & Sharma, N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23-39.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, 67(S1), 155-179.

Gardner, S. & Nesi, H. (2013) A classification of genre families in university student writing. *Applied Linguistics*, v. 34, n. 1, p. 25-52.

Gardner, D. & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327

Goulart, L. (2017). Compilation of a Brazilian academic written English corpus. Revista e-scrita: *Revista do Curso de Letras da UNIABEU*, 8(2), 32-47.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, analysis, and applications*, 145 - 160.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds.). (2009). I*nternational corpus of learner English (Vol. 2)*. Louvain-la-Neuve: Presses universitaires de Louvain.

Guedes, A. D. S. (2017). *Verbos do inglês acadêmico escrito e suas colocações: um estudo baseado em um corpus de aprendizes brasileiros de inglês*. PhD Thesis. Universidade Federal de Minas Gerais

Hill, J. (1999). Collocational competence. *Readings in Methodology*, 162.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied linguistics,* 19(1), 24-44

Hyland, K. & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for academic purposes*, 1(1), 1-12.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). *The sketch engine. Information Technology*, 105, 116

Kumar, V. P. & Rao, C. S. (2018). A review of reasons for rejection of manuscripts. *Journal for research scholars and professionals of english language teaching*, 8(2), 1-11.

Laufer, B. & Waldman, T. (2011). Verb- noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647-672.

Lea, M. R. & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in higher education*, 23(2), 157-172.

Li, J. & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, 18(2), 85-102.

Lillis, T. M. (2001). *Student Writing: Regulation, Access, Desire*. London: Routledge.

Lorenz, Gunter (1999). *Adjective Intensification – Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, 33(3), 299-320

Matte, M. L. & Rebechi, R. R. (2018). A quantitative analysis of collocations in Brazilian and British students' academic writing. *Entrepalavras*, 9(2), 195-213

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. *Formulaic language*, 2, 375-86.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis.* London: Bloomsbury Publishing.

Rayson, P. (2002). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD Theses, Lancaster University.

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.

Wu, J. (2016). *A Corpus-Based Contrastive Study of Adverb + Verb Collocations in Chinese Learner English and Native Speaker English*. Master degree project. Stockholm University.

Appendix 01

Node	BAWE	BrAWE	0 occurrences in BrAWE	Node	BAWE	BrAWE	0 occurrences in BrAWE
system	48	23	25	example	24	8	16
result	53	31	22	conclusion	8	6	2
value	50	23	27	conflict	7	2	5
figure	15	3	12	standard	25	8	17
process	52	20	32	reference	1	1	0
group	50	16	34	aspect	22	11	11
level	49	14	35	error	15	7	8
model	59	17	42	movement	3	1	2
development	45	12	33	task	20	13	7
data	55	10	45	measure	25	0	25
information	51	21	30	importance	25	12	13
research	41	15	26	support	18	5	13
analysis	34	15	19	feature	23	5	18
rate	55	18	37	discussion	4	1	3
effect	53	22	31	perspective	6	1	5
method	51	19	32	influence	13	6	7
change	55	20	35	requirement	21	8	13
strategy	43	13	30	extent	8	5	3

factor	68	25	43	characteristic	23	3	20
control	31	7	24	interaction	6	2	4
use	45	21	24	author	2	1	1
policy	30	8	22	degree	10	5	5
theory	20	3	17	capacity	12	5	7
approach	32	13	19	understand-ing	13	7	6
structure	26	11	15	concern	15	8	7
role	32	12	20	pattern	17	8	9
quality	29	16	13	reduction	10	5	5
difference	41	18	23	basis	9	4	5
function	28	12	16	definition	11	5	6
activity	37	11	26	procedure	9	5	4
organisation	16	5	11	trend	25	5	20
environ-ment	31	6	25	consideration	12	2	10
resource	26	11	15	observation	5	3	2
type	34	11	23	potential	11	3	8
society	5	2	3	improvement	11	6	5
condition	46	16	30	purpose	7	2	5
production	34	7	27	finding	13	8	5
form	20	4	16	assumption	9	3	6
section	16	5	11	outcome	10	5	5
interest	23	7	16	aim	5	2	3
relationship	35	12	23	presence	6	3	3
source	25	13	12	consequence	9	3	6
impact	30	16	14	explanation	6	4	2
practice	18	5	13	implication	7	0	7
need	46	20	26	variation	9	4	5
growth	23	8	15	category	10	2	8
material	26	11	15	difficulty	14	8	6
period	14	5	9	description	6	3	3
increase	28	11	17	link	8	3	5
review	6	3	3	attempt	1	1	0
term	16	6	10	shift	5	2	3
solution	24	17	7	significance	1	0	1
individual	6	0	6	limitation	2	1	1
concept	18	10	8	proportion	7	5	2
demand	25	9	16	phenomenon	7	5	2
population	26	10	16	recognition	2	1	1
element	24	12	12	contrast	0	0	0

knowledge	23	8	15	contribution	5	3	2
introduc-tion	3	0	3	alternative	4	4	0
benefit	35	15	20	insight	7	5	2
experience	17	6	11	tendency	1	1	0
technique	30	10	20	exception	1	1	0
range	21	9	12				
TOTAL	BAWE		BrAWE		0 occurrences in BrAWE		
	2679		1015		1664		

Appendix 02: Types of collocates for each node

NODE	Modifier	Object	Subject	Modifier	Object	Subject
	BAWE			BrAWE		
system	9	22	17	4	11	8
result	20	23	10	11	14	6
value	19	23	8	7	15	1
figure	9	5	1	0	2	1
process	14	24	14	5	9	6
analysis	12	14	8	4	10	1
group	18	20	12	7	5	4
level	21	25	3	6	8	0
model	14	24	21	4	11	2
development	25	16	4	10	3	0
data	20	23	12	1	7	2
information	24	25	2	10	11	0
research	22	9	10	9	3	3
rate	24	24	7	7	10	1
effect	25	25	3	11	10	1
method	17	23	11	11	6	2
change	24	24	7	10	8	2
strategy	18	19	6	6	6	1
factor	25	25	18	14	8	3
control	14	16	1	3	5	0
use	25	18	2	8	12	1
policy	9	16	5	1	5	2
theory	3	9	8	0	3	0
approach	10	15	7	3	9	
structure	10	15	1	3	8	0
role	18	14	0	8	4	0

quality	12	16	1	6	10	0
difference	24	15	2	9	9	0
function	11	16	1	5	7	0
activity	16	19	3	4	7	0
organisation	5	3	8	1	2	2
environment	23	7	1	4	2	0
resource	16	9	1	9	2	0
type	22	12	0	8	3	0
society	5	0	0	2	0	0
condition	25	17	4	11	4	1
production	22	9	3	4	3	0
form	16	4	0	3	1	0
section	10	1	5	3	1	1
interest	13	10	0	4	3	0
relationship	17	17	1	5	7	0
source	20	4	1	10	2	1
impact	21	8	1	12	4	0
practice	14	4	0	4	1	0
need	25	21	0	10	10	0
growth	13	10	0	5	3	0
material	14	9	3	7	4	0
period	11	2	1	5	0	0
increase	21	6	1	7	4	0
review	4	2	0	2	1	0
term	10	5	1	4	2	0
solution	9	13	3	5	9	3
individual	1	3	2	0	0	0
concept	9	9	0	2	8	0
demand	14	11	0	5	4	0
population	19	5	2	8	2	0
element	18	7	0	8	4	0
knowledge	13	10	0	3	5	0
introduction	3	0	0	0	0	0
benefit	2	19	2	8	6	1
experience	13	4	0	4	2	0
technique	18	10	3	4	5	1
range	14	7	0	6	3	0
example	14	8	2	5	3	0
conclusion	4	4	0	2	4	0
conflict	2	3	2	0	2	0
standard	12	12	1	5	2	1

reference	0	1	0	0	1	0
aspect	15	7	0	10	1	0
error	5	10	0	3	4	0
movement	2	1	0	1	0	0
task	11	8	1	8	4	1
measure	16	7	2	0	0	0
importance	11	14	0	6	6	0
support	12	6	0	1	4	0
feature	16	5	2	3	1	1
discussion	4	0	0	1	0	0
perspective	6	0	0	1	0	0
influence	11	2	0	6	0	0
requirement	15	6	0	5	3	0
extent	6	2	0	5	0	0
characteristic	17	5	1	2	1	0
interaction	4	1	1	2	0	0
author	2	0	0	1	0	0
degree	7	3	0	3	2	0
capacity	8	4	0	5	0	0
understanding	8	5	0	4	3	0
concern	12	3	0	6	2	0
pattern	12	5	0	5	3	0
reduction	4	6	0	2	3	0
basis	6	3	0	3	1	0
definition	8	3	0	2	3	0
procedure	5	3	1	1	3	1
trend	15	7	3	1	4	0
consideration	8	5	0	1	1	0
observation	2	3	0	1	2	0
potential	5	6	0	2	1	0
improvement	5	6	0	2	4	0
purpose	7	1	0	2	0	0
finding	5	4	4	2	3	3
assumption	4	4	1	1	2	0
outcome	5	6	0	4	1	0
aim	4	1	0	1	1	0
presence	2	4	1	0	2	1
consequence	9	0	0	3	0	0
explanation	3	3	0	2	2	0
implication	5	2	0	0	0	0
variation	6	3	0	2	2	0

category	8	1	1	2	0	0
difficulty	7	8	0	3	5	0
description	4	2	0	1	1	0
link	4	4	0	2	1	0
attempt	0	1	0	0	1	0
shift	1	4	0	0	2	0
significance	1	0	0	0	0	0
limitation	2	0	0	1	0	0
proportion	6	1	0	5	0	0
phenomenon	3	4	0	2	3	0
recognition	1	1	0	1	0	0
contrast*	0	0	0	0	0	0
contribution	4	4	0	2	1	0
alternative	2	2	0	2	2	0
insight	2	5	0	1	4	0
tendency	1	0	0	1	0	0
exception	1	0	0	1	0	0
TOTAL	1359 (50.7%)	1049 (39.1%)	271 (10.1%)	506 (49.8%)	444 (46.7%)	65 (6.4%)
	2679			1015		

**contrast* is an academic noun classified in Frankenberg-Garcia et al.'s (2018) study that does not have productivity in BAWE nor in BrAWE.

From corpus to classroom: evaluating Web-based tools to teach collocations

Larissa Goulart (Montclair State University)
Maria Kostromitina (Northern Arizona University)
Jennifer Klein (Coconino Community College)

Introduction

Throughout the years, researchers have defined collocations in different ways. Men (2018) for instance, defines collocations as sequences of words that are transparent in meaning (e.g. *make a decision*). Durrant and Mathews-Aydınlı (2011: 60) on the other hand, focus their definition on the linguistic aspect, stating that collocations are "successions of linguistic entities that are best learned as integral wholes or independent entities (…) (collocations) occur with sufficient frequency that their independent learning will facilitate fluency". These, sometimes conflicting, definitions of collocations emerge from the different approaches used in the study of collocations. Here we will focus on two of those: the phraseological approach and the frequency approach (Wolter & Gyllstad, 2013; Gablasova et al., 2017).

Researchers such as Paquot and Granger (2012) have focused on the phraseological approach to analyze collocations. Some examples of collocations in the phraseological approach are: *face a problem*, *take a step*, and *reach a conclusion* (Paquot & Granger, 2012). As we can see, the phraseological approach focuses on the semantic relationship between words in a collocation and their idiomatic nature. The frequency approach, in contrast, focuses on which words frequently occur together, such as *prepare meals*, *fixed an error*, and *conquered the city*. Research using the phraseological approach can also adopt measures of association, such as MI

(mutual information) scores or *t*-scores (see Brezina, 2018: 74 for a complete description of measures of association). It is worth noting, however, that some measures of association can be misleading. MI scores, for example, identify words that occur together frequently, but do not necessarily identify collocations that are frequent in the overall language. This distinction between the way collocations can be identified is reflected in collocation teaching materials. That is, when teaching collocations, some materials focus on the most frequent constructions, while others focus on idioms. This is one reason why it is important for teachers to be able to evaluate ready-made tools for learning collocations, as teachers seek to teach the most frequent collocations to their learners.

In addition to this divide between the phraseological and the frequency approach, there are other aspects that cause confusion when defining a collocation. Within the frequency approach, many studies will focus exclusively on collocations with lexical words, such as verb-noun combinations (Boers et al., 2014; Tsai, 2020) or adjective-noun combinations (Wolter & Gyllstad, 2013). Another point of disagreement is how to account for intervening words; some research allows for intervening words in a collocation (*bring* to *light*) (Tsai, 2020), while others do not (*give thanks*) (Yamashita & Jiang, 2010). Additionally, some researchers investigate n-grams as collocations, that is, they examine longer sequences of words in terms of collocational use (Gablasova et al., 2017).

Studies have also taken dispersion into account when defining collocations. Dispersion refers to the degree to which collocations are used frequently in different texts in a corpus (Gablasova et al., 2017). This is particularly important for language teachers, as teaching only the most frequent collocations, without accounting for dispersion, could mislead the learner to acquire a collocation that, in reality, only occurs in one particular text, or in one particular discipline.

Collocation research has also been connected to L2 learning. One relevant concept found in the domain of L2 writing research is the distinction between congruent and non-congruent collocations (Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010). Congruent collocations have similar lexical elements as collocations in the learner's first language, while

non-congruent collocations do not exist in the learner's native language. This distinction is especially relevant in research examining language transfer from a learner's L1 to an L2 and should be kept in mind when teaching collocations in the L2.

Even though the appropriate use of collocations is usually associated with native-like English proficiency (Bahns & Eldaw, 1993; Chen, 2011), research in second language acquisition has shown that collocations can be a challenge to language learners. Granger (1998), for example, shows that L2 learners of English tend to use more collocations that are congruent with collocations that exist in their native language. Nesselhauf (2011) also finds the same results when examining the production of German learners of English suggesting that non-congruent collocations are considered the most difficult ones for second language learners. Ellis (1996) argues that L2 learners' acquisition of formulaic sequences differs from that of native speakers in the sense that native speakers process formulas relying on semantic associations, while L2 learners rely on orthography and phonology, driving them to, possibly, making incorrect associations based on orthographic or phonological confusion. In a comparatively recent study, Ellis et al. (2008) confirm that native speakers process formulas based on different criteria than L2 learners; while the latter used formulas that are more frequent, the former used formulas that had a stronger association between words.

In sum, collocations present a challenge for learners due to four main aspects: First, language learners have difficulty identifying exactly which words collocate (Jiang, 2009), for example, there is no grammatical reason why *conduct research* is more common than *perform research*. Second, a node-word can have more than one collocate and each of these combinations can have different meanings (Nesselhauf, 2003). One such case is the word *face,* when it collocates with *to face* (as in *face to face*) it means *stand in front of*, and when it collocates with *away* it means to look to the other side. Third, collocations do not transfer from students' first language. Chan and Liou (2005), for example, point out that the difference between *take medicine* and *eat medicine* is not clear for learners with a Chinese background because this difference does not exist in mandarin. Finally, Cobb

(2018) suggests that even though collocations are pervasive in language, it is unlikely students will encounter them a meaningful number of times in their classroom readings and textbooks in order for these structures to be acquired in a classroom environment.

Considering the challenge that collocations can present to language learners, researchers have proposed a number of ways to help learners acquire these constructions. Cobb (2018) argues for the use of concordance lines in the teaching of collocations because, in contrast to textbooks, concordance lines give students standard associations that are possible in language, while textbooks expose students to non-standard collocations in exercises such as fill in the gap. Chan and Liou (2005) also highlight the fact that without the use of concordancing tools it is unlikely that students will encounter a collocation enough times to learn it inductively. Another argument for the use of corpus tools in the learning and teaching of collocations is that it allows learners to work independently, which is crucial for the acquisition of collocations (Woolard, 2000; Conzett, 2000). In spite of these arguments for the use of corpus tools in teaching collocations, Cobb (2018) notes that most Computer Assisted Language Learning (CALL) tools developed for English language learners focus on a single unit (i.e., only words), with concordancing tools that integrate collocations (and other formulaic language) remaining somewhat limited. Therefore, the goal of this chapter is to evaluate what collocation tools have to offer to language teachers and suggest tasks that integrate the use of these tools in the EFL classroom.

Web-Based Learning tools for collocations

Web-based learning tools (WBLT), or learning objects, are "interactive, online learning tools that support the learning of specific concepts by enhancing, amplifying, or guiding the cognitive processes of learners" (Kay, 2011: 1849). WBLTs allow students to manipulate different aspects of language in order to understand how language works. In the case of collocations, this manipulation can be in the form of the node-word, the

position of the collocates, the types of texts in which they occur, among other variables.

To date, most studies evaluating WBLTs have been conducted by the tools' developer, usually upon the launch of the tool (see Chen, 2011; L'Huillier, 1990; Nesbitt, 2012, for examples). There are two issues with this type of evaluation: first, this approach focuses on the evaluation of a single tool at a time, therefore, it does not present comparisons between existing tools. These comparisons are relevant to understand the tool that fits better in a specific context. Second, since each researcher is conducting an independent evaluation, the criteria set for evaluation varies widely. Chen (2011), for instance, includes part of speech tagging, frequency summary, retrieval speed, link to examples, search options and corpus size as their criteria, while Nesbitt (2012) focuses only on the design of the WBLT. These differences in evaluation criteria make it impossible to compare results across studies.

Evaluating several WBLTs using the same criteria allows teachers and researchers to determine which tool is more appropriate for specific learning contexts (i.e., teaching English to high school students, teaching Academic English to L2 graduate students, etc). Kay and Knaack (2009) also argue for the need of a structured and organized evaluation criteria that can later be used for teachers to evaluate new tools as they appear on the market. Therefore, we seek to propose an evaluation scheme that can be used by teachers and tool developers to assess the applicability of their tools for specific classroom contexts.

Previous studies such as Nurmukhamedov (2015) have investigated collocation tools from a learners' perspective; nevertheless, this author has explored online collocation dictionaries and a printed version of word and phrase. The current study seeks to evaluate WBLTs developed to teach collocations that are completely online and free to use. We believe that by evaluating these tools we can a) help inform the development of better tools in the future; b) inform teachers' decisions of the tools to use in each context; c) suggest tasks for integrating these tools in the classroom; and d) push developers to make more information available as to how they developed these tools.

Methods

Selecting the tools

The first step in addressing the research questions in the present study consisted of selecting the web-based collocation tools for evaluation and comparison. For this purpose, we developed specific inclusion criteria for the tools to be selected for the evaluation. Thus, to be included in the study, the web-based collocation tools had to meet the following criteria:

(1) be hosted on a specific website (i.e., online);

(2) be free to access;

(3) be corpus-based (i.e., grounded in corpus-based methodology)[1];

(4) allow word searches.

These criteria allowed us to exclude such collocation software as *Antconc* as it is not hosted online and generally requires installation on a computer. Additionally, *SketchEngine*, while widely used in research, was not included because it is a paid tool allowing only for a free 30-day trial with limited access to its tools. We also excluded pre-made collocation lists, such as a dictionary of collocations, and websites like *CollocAid* that do not have an option to search for collocations for a word of interest. In the end, five web-based collocation tools were identified for the evaluation: *FLAX, SKELL, Just the Word, Linggle,* and *Netspeak*.

Evaluation Rubric

After the web-based collocation tools were selected, they were assessed using an evaluation framework that was developed on the basis of: a) existing rubrics for the evaluation of education tools (e.g., Rosell-Aguilar, 2017), b) findings of previous research that focused on language learning apps overall. Broadly, research in learners' evaluation of English-learning computer or mobile apps has indicated that on top of the content quality,

1 The extent to which a tool was corpus-based or corpus-informed was determined by reading the information available on the tool's website. More specifically, we examined the source of the collocations presented to the user.

learners value the usability (also defined as operation and design) and customization of a tool, as well as its ability to give feedback (Chen, 2016; Smith & Ragan, 2004). In relation to usability, Rosell-Aguilar (2017) named navigation, accessibility, clear instructions, and the quality of sound and image among the factors that contribute to the success of a tool. Haughley and Muirhead (2005) also propose that learning tools need to be linked to the communicative experiences of the learners and thus encourage engagement. Other categories that are often used to evaluate computer-based language learning tools include the feasibility of a platform, such as its flexibility and reaction speed (Nesbitt, 2012).

While the criteria above needed to be taken into consideration in developing the evaluation rubric for the present study, we also accounted for additional characteristics specific to collocations. McEnery et al. (2006) provide a list of criteria that are required in collocation learning tools. They emphasized that a tool has to provide substantial information about the collocation and its use. For instance, a tool should allow its users to check the frequency of a collocation and its distribution across source texts in case the collocation is register-specific. A collocation tool should also report on the statistical measure(s) (*t* scores or MI scores) and positional information regarding the collocate to the node. Finally, learners should be able to adjust the distance between collocating items (or the collocation window) and between colligations (a type of collocation where lexical items are tied to grammatical ones, e.g., verbs of perception colligate with object and a non-finite verb complement clause) and collocations (McEnery et al., 2006). In addition, Chan and Liou (2005) and Yoon and Hiverla (2004) comment on the presentation of collocations, reporting that learners experienced difficulty with collocation tools as they presented cut-off sentences in the concordancer and learners were unable to locate the appropriate collocates. Along with these essential features, Chen (2011) highlights the importance of the corpora underlying a collocation tool, saying that the corpus used in a tool needs to be large in size (more than 100 million words) and pre-tagged for parts of speech as well as text registers.

Synthesizing and adapting the characteristics of language learning tools and collocation apps that are pervasive across the existing frameworks,

we developed an evaluation rubric that consisted of four major categories: content quality, interface, presentation of search outcomes, and feedback. The first category addressed the concerns about the quality of the presented collocations in terms of the underlying corpus research as well as the ability of the tool to be adjusted based on a learner's needs. Thus, the content quality was operationalized as the amount and quality of linguistic research conducted and corpora used to create the tool. Additionally, the criterion included the tool's ability to account for register variation, to filter the presented collocations based on specific criteria, and to account for adjacent words between the nodes in the search. The interface criterion encompassed the usability of a tool, clearly defined menu options, navigation in the tool, and user-friendliness (i.e., how many clicks does it take a learner to get to the information they are looking for?). The third criterion, presentation of search outcomes, involves the tool options related to the identification of the collocations. The criterion accounts for the ability of a tool to provide learners with the information about the part of speech of the collocates and the frequency of the collocation. It also evaluates the flexibility in the way learners can search for collocations including misspellings, the side of the collocates in relation to the node, etc. Finally, the last criterion in the rubric addressed the issue of feedback and assessed whether a tool provided learners with an evaluation of the collocation they produced. The complete evaluation rubric can be found in Appendix A.

Scoring

We evaluated each collocation tool in the study on a five-point Likert scale. That is, each criterion in the rubric was assigned five points with one being the lowest and five being the highest score a tool could receive. The total score each tool could receive was 15 points. While the length of Likert scales is often defined arbitrarily, five-point scales have been commonly used in various domains of Second Language Acquisition (SLA) research, such as speech comprehensibility and accentedness (Trofimovich & Isaacs, 2013), learners' individual differences (MacIntyre & Vincze, 2017), and L2 writing (Becker, 2018) among others. Each tool was first evaluated by the

first and second author separately, after which the raters then met to discuss the discrepancies and the agreement reached 100%.

Results

Description and evaluation of tools

This section focuses on describing the five web-based tools and their evaluation. The tools examined in this study were *SKELL*, *FLAX*, *Linggle*, *Just the Word*, and *Netspeak*. First, each tool is described according to its three aspects: a general description of the tool, searches using that tool, and results of searches using the tool. Next, the reference corpus/corpora for each tool are described, followed by examples of searches and results. Then, a summary of the three main evaluation criteria (content quality, interface, and presentation of search outcomes) is given for each tool.

*S*KELL
https://skell.sketchengine.eu/

SKELL, or Sketch Engine for Language Learning (Baisa & Suchomel, 2014), is a search engine that allows users to search for words or phrases and see concordance lines and possible collocations of the word by part of speech. Information regarding the reference corpus/corpora for this tool was not available.

Searches in SKELL. First, users will click on the "Try SKELL" button on the homepage of the website. Then, users will type in the word or phrase that they wish to find information about. There are three different tabs that learners are presented with when they enter their word in SKELL: 1) Examples (which are concordance lines), 2) Word Sketch (which provides collocational information), and 3) Similar Words (providing synonyms). So, learners will need to be instructed by their teacher to click on the "Word Sketch" tab to access collocations.

Results of searches in SKELL and examples. "Word Sketch" provides learners with information related to the word that they searched. First, Word Sketch allows users to select the part of speech of the word

they searched. Possible collocations of a word are then organized by part of speech and function. Finally, users can see examples of how these collocations are used (portions of concordance lines). For example, if one searches for the word *purpose* (we must use the base form of a word when performing the search), they can choose if they want to search the word as a noun or verb (the most frequent part of speech will be automatically selected, a dropdown menu provides options for other parts of speech). For this example, the word *purpose* was searched as a 'noun'. The following categories are presented when this search is performed: 1) verbs with *purpose* as subject (*a purpose built*), 2) verbs with *purpose* as object (*serve the purpose*), 3) adjectives with *purpose* (*purpose is twofold)*, 4) modifiers of *purpose* (*for the sole purpose of*), 5) nouns modified by *purpose* (*all-purpose flour*), 6) *words and* (this shows phrases like *purpose and meaning*, *purpose and direction*), and 7) *or purpose* (showing phrases like *motive or purpose*).

Content Quality of SKELL. Of all the tools examined, SKELL is the most complete in terms of content quality and usability. The collocations presented to learners come from large corpora that are available on Sketch Engine, as a result, learners can find more information about the collocations of interest if they log on to Sketch Engine. The only issue in terms of content quality is that learners do not have the option to select collocations that occur in specific genres or disciplines.

SKELL's interface. The interface clearly indicates that SKELL was developed for English learners. In terms of the language used and the menu design, the website is clearly targeted for learners, using only a single word to describe the menus and providing a limited number of options. The fact that SKELL provides only three types of search, described above, also makes it easy for learners to locate the right option for their needs.

Presentation of search outcomes. In comparison to the other WBLTs evaluated, SKELL seems to be the most appropriate to use with students without extensive classroom training. It only requires that learners know how to type the word that they are searching. The results are then presented divided by part of speech, as detailed above. One of the great advantages of SKELL is that learners can click on a collocation and find example sentences of this collocation being used in context.

FLAX

http://flax.nzdl.org/

FLAX, or Flexible Language Acquisition (Fitzgerald et al., 2015), provides a tool for searching for collocations based on the British National Corpus (BNC), British Academic Written English (BAWE) corpus, and the Wikipedia corpus.

Searches in FLAX. The collocation tool in FLAX is under the menu Learning Collocations. This tool searches the reference corpora (BNC, BAWE, or Wikipedia corpus) for collocations, and presents them to the user. FLAX allows for collocation searches in six different registers: 1) Contemporary English (Wikipedia corpus), 2) Standard English (BNC), 3) Academic English in Physical Sciences, 4) Academic English in Social Sciences, 5) Academic English in Life Sciences, and 6) Academic English in Arts and Humanities. Users need to choose which register to search in from a dropdown menu. Users simply enter a word, choose the register, and click "go".

Results of Searches in FLAX and examples. The results of a search display collocates by part of speech, and include the frequency of the collocation in the reference corpus. The ten most frequent collocations are automatically displayed for each part of speech and users can click "more" to see less frequent collocations.

Continuing with the previous example of *purpose* from above, the most frequent collocations for each of the six registers are displayed in Table 1.

Register	Collocations
Contemporary English	*main purpose*, *primary purpose*, *sole purpose*
Standard English	*main purpose*, *sense of purpose*, *primary purpose*
Academic English in Physical Sciences	*used for this purpose*, *suited for this purpose*, *developed for this purpose*
Academic English in Social Sciences	*non-commercial purpose*, *main purpose*, *commercial purpose*
Academic English in Life Sciences	*used for this purpose*, *purpose in life*, *purpose of this study*
Academic English in Arts and Humanities	*purpose of the study*, *main purpose*, *different purpose*

Table 1. Most frequent collocations of *purpose* by register

Content Quality of FLAX. As described above, FLAX relies on a combination of different corpora to extract collocations; nevertheless, differently from SKELL, learners using FLAX can choose the specific text types that they are interested in. In terms of content quality, FLAX also has good documentation of the criteria used to extract collocations, making it one of the best tools in this criterion.

FLAX's Interface. FLAX seems geared towards advanced learners and teachers. Unlike SKELL, FLAX's interface can be confusing due to the extensive number of menu options. While these options can be beneficial for learners who are interested in learning collocations in a specific discipline (e.g. law, life sciences, etc.) or register (university writing, abstracts, etc.), the menus are not clearly labelled, which can be distracting.

Presentation of search outcomes. This tool provides plenty of materials for learners and teachers interested in academic English, from lesson plans to lists of collocations. The advantage of this tool is that results are organized in terms of part of speech, which can help learners visualize language patterns.

Linggle
https://linggle.com/

Linggle (Boisson et al., 2013) is a search engine that allows users to search for collocations, specifying the number of words and parts of speech of the collocations. The search engine draws from several reference corpora including Google Web 1T 5-gram, the BNC, and the New York Times Annotated Corpus.

Searches in Linggle. Users need to know wildcards[2] to use this search engine. Linggle also allows users to search with part of speech tags. This makes searching for collocations more challenging for users than some of the other tools described which allow users to search for words or phrases with the use of buttons rather than wild cards and parts of speech.

Results of searches in Linggle and examples. Results of a search in Linggle are collocations displayed by frequency. The results provide a frequency of the collocation and percentage. It is not stated whether these are raw frequencies or normed frequencies. In addition, there is no explanation for the percentages. So, the percentages might represent how often the collocate appears in collocation with that node or the frequency of that word in percentage to the number of words in the reference corpus. If a user clicks on a collocation, concordance lines are displayed under the collocation.

We will again use the word *purpose* for the example of a search in Linggle. For this example, the search was "det. purpose _ n.", which searches for collocations of *purpose* that have a determiner, followed by *purpose*, followed by any single word, followed by a noun. The collocations returned, in order of frequency, are *the purpose of making*, *the purpose of carrying*, *the purpose of conducting*, and others.

Content Quality of Linggle. As stated above, Linggle extracts collocations from a combination of corpora that are available online. One pitfall of Linggle is that it combines these texts, without accounting for the possibility of variation in the collocations used across text types. In addition, it

2 A wildcard is used in a search to substitute one or more characters in a string or word.

is unclear the criteria used for the extraction of collocations and as a result the meaning behind the percentages presented on the output is ambiguous.

Linggle's Interface. Linggle requires extensive training to be accessible for learners. As the search has to be conducted with wild cards, learners need to learn the wild cards and their meanings in order to search for collocations. Nevertheless, once learners become familiar with these wild cards they could search for collocations with or without intervening words and look for collocates on both sides of a node. Extensive training aside, the tool has clear menu options.

Presentation of search outcomes. The results of a Linggle search are displayed in terms of frequency, but it is unclear what exactly this frequency represents in the overall corpus, especially the percentage that is presented on the right. Another issue with the output from Linggle is that the example sentences are not centralized, leaving it to the learner to find the context of the collocation.

Just the Word
http://www.just-the-word.com/

Just the Word is a web-based tool that allows users to see collocations of a word they have searched. The reference corpus for Just the Word is the BNC.

Searches in Just the Word. To search for collocations, users enter a word into the search bar and click the "combinations" button. Users can also use the "alternatives" button to search for suggested alternatives to the collocation they have entered in the search bar.

Results of searches in Just the Word and examples. When users search for a word using the "combinations" feature, the results are organized by part of speech and function. If a user wants to see examples of a collocation, they can click on the collocation and will be taken to a page with all concordance lines that include that collocation. The results also display a raw frequency for the number of times the collocation appears in the reference corpus. Each collocation has a green or red bar next to it to indicate whether the collocation is a "good" or "bad" combination. The length of the green or red bar indicates frequency as well.

Returning to our example with the word *purpose*, if we search this word using the "combinations" feature, the results indicate that *purpose* is a noun and provide different categories of collocations by part of speech. The collocations returned for our example are *achieve purpose*, *acquire for purpose*, *assume for purpose*, and others.

Content Quality of Just the Word. First, Just the Word examines only collocations encountered in the BNC, which limits the search in terms of text types. Interestingly, the tool does not provide an option to select the collocations of specific text types, even though this should be possible considering the reference corpus. Similar to Linggle, there is no documentation on the criteria used to determine which collocations were extracted from the corpus. In addition, the "good" and "bad" evaluations for the collocations presented seem to be based solely on frequency.

Just the Word's Interface. The results of Just the Word are very similar to the ones provided by SKELL, with two main differences: the reference corpus and the interface. The interface is similar to SKELL's in terms of giving the option to search by part of speech and the way the results are presented. The disadvantage is that the results for all different parts of speech are presented in the same screen, which might be confusing for an inattentive learner.

Presentation of search outcomes. As a result of the way the search results are presented, all in the same screen, the output can be confusing for learners. One main advantage is that learners can click on a collocation and see example sentences with the collocation centralized, so that learners can notice language patterns around the collocation.

Netspeak
https://netspeak.org/

Netspeak (Potthast et al., 2010) is another search engine that allows users to search for collocations using wild cards. Netspeak provides collocations for both English and German. The reference corpus for the English version of Netspeak is Google Books.

Searches in Netspeak. To conduct searches for collocations in Netspeak, users must use wild cards. The user enters their search terms into the search bar and collocations are automatically displayed as they search.

Results of searches in Netspeak and examples. Results of a search in Netspeak display the most frequent collocations first. Netspeak also provides raw frequencies of collocations and a percentage in the results. If users click on the collocation in results, they are provided with excerpts from Google Books that include examples of the collocation. These excerpts provide more of the text than just the concordance line in which the collocation is used.

For this example, "the purpose ?" was used to search for collocations, searching for collocations of *purpose* that are preceded by *the* and followed by at least one character (this includes punctuation marks). The results of this search returns collocations such as *the purpose of*, *the purpose for*, *the purpose and*, and others.

Content Quality of Netspeak. Netspeak provides almost no documentation; the only information available for how the collocations are extracted is that the reference corpus is Google Books. Therefore, it is difficult to use Netspeak without knowing the criteria used to extract collocations. Similar to Linggle, the results window present percentages on the right that are not defined, leaving it to the user to guess whether this is a percentage of the total number of words or a percentage of the collocates with that node.

Netspeak's Interface. Similar to Linggle, Netspeak requires that students learn wild cards to conduct the search, but differently from Linggle the results page is very limited. Overall, the menus are clear, but the fact that wildcards have to be used in the search limits its usability in the classroom.

Presentation of search outcomes. The output shows only the collocates organized by frequency. It does not provide information in terms of part of speech, dispersion, or text types. In addition, in order to obtain examples, the learner has to continue clicking on the website, which might make it difficult to return to the results page.

The following table details the evaluation of each tool according to all the criteria discussed in section 4.2.

		SKELL	FLAX	Linggle	Just the word	Netspeak
Content quality		3.5	3	3.5	2	1.5
Definition of Collocation	What is the definition of collocation used?	The website provides the following definition: "Word Sketch is a list of words which occur frequently together with the searched word and collocation is a typical combination of two words."	Not available	Not available	Not available	Not available
Linguistic research	Reference Corpus	SKELL relies on a combination of corpora including Wikipedia, English Web 2013, BNC, WebBootcat, among others.	FLAX uses a combination of corpora, including BNC, Google, Wikipedia, EThOS, MOOC, BAWE. Users have access to different tools to select which reference corpus to use.	Linggle relies on the Google 1T interface as the source for collocations.	BNC	Google Books
	Types of texts	Wikipedia articles and web registers. The developers define it as "standard, everyday, formal, professional English".	Wikipedia articles, university writing, PhD abstracts, law course materials, etc.	Not available	Not available	Not available
	Frequency threshold	Not available	Not available	Not available	Not available	Not available
	Measure of association	The developers use LogDice to determine the measure of association of the collocations.	Not mentioned	Not available	Not available	Not available

Language Variation	Does it account for register variation?	No, all corpora are combined in the search.	Users can select to search in different reference corpora accounting for different registers.	No, all texts are combined in the same search.	No, all texts are combined in the same search.	No, all texts are combined in the same search.
	Does it account for disciplinary variation?	No	In the university writing corpus, users can select to search based on different disciplinary groups.	No, all texts are combined in the same search.	No, all texts are combined in the same search.	No, all texts are combined in the same search.
Filter	Is there a teaching filter in place?	No	Not mentioned	No	No	No
Interface		4	4	2.5	2	2
Intuitivity	Does the tool have clear menus and icons to facilitate navigation?	The tool is clear to navigate, but the collocation tool is under the label "word sketch" which might make it confusing for students.	Yes	Yes	The search is well-indicated, but the options are confusing.	Yes
	Can learners use it without in depth training?	Yes	Yes	No, learners need to learn some special characters in order to conduct the search.	Yes	No, learners need to learn some special characters in order to conduct the search.
	How many clicks are necessary to finalize the search?	1 to 2	2 to 3	1 to 2	1 to 2	1 to 2

Design features	Is the interface appealing to students?	Yes, the fonts are large, and the menu labels are clearly defined.	The tool seems to be designed with teachers in mind. As a result of the many search options, learners might get a bit confused.	Yes, the interface is pretty clear, and the search is visible.	No, the interface is a bit polluted showing previous searches from users all over the world.	No, the interface does not specify how learners can conduct the search.
Presentation of search outcomes		4	4	4.5	4	2
Search	How is the search structured?	Users can search for only one word at a time.	Users can search for one word or a two-word sequences, but preference is given to the first word in the case of sequences.	Users can search for one word or a two-word sequences, but they will need to use special characters for the two-word searches.	Users can search for only one word.	Users can search for only one word.
	Does it provide collocates to both sides of the word?	The tool presents collocates that occur in both sides of the search word.	The tool presents collocates that occur in both sides of the search word.	Yes, but the user must specify which side of the node they want to see collocations.	Yes, the tool provides collocates on both sides of the node and also accounts for intervening words.	Yes, but the user must specify which side of the node they want to see collocations.
	Does it account for intervening words?	Collocations with intervening words are included in the results.	Collocations with intervening words are included in the results.	It is possible to account for intervening words if the user is familiar with regex.	Yes	It is possible to account for intervening words if the user is familiar with regex.

	Does the tool correct misspelled words in the search?	The tool does not correct and if the misspelled word occurs in the corpus, it will provide results to that search.	The tool does not conduct the search if a word is misspelled.	The tool does not correct and if the misspelled word occurs in the corpus, it will provide results for that search.	The tool does not correct the misspelled word, but also does not provide any matches.	The tool does not correct and if the misspelled word occurs in the corpus, it will provide results for that search.
	Can the search be limited by part of speech?	The tool allows users to limit the part of speech being searched.	The tool does not offer this option.	Yes, as long as user knows the tag set being used.	No	No
Presentation	Which criteria is used to determine the order of presentation?	Results are organized by part of speech and by frequency.	It is not clear, but it seems that frequency is used to decide which collocations appear first.	Frequency	Results are organized by part of speech and by frequency.	Frequency
	Does the tool provide frequency and dispersion for the collocations?	SKELL provides frequency only.	FLAX provides frequency only.	Linggle provides frequency only.	Just the word provides frequency only.	Netspeak provides frequency only.
	Does the tool show examples of the collocation being used in context?	If the learners click on collocation, the tool opens a new window with several example sentences of the collocation being used.	If the learners click on collocation, the tool opens a new window with several example sentences of the collocation being used.	If the learners click on collocation, the tool opens a new window with several example sentences of the collocation being used.	If the learners click on collocation, the tool opens a new window with several example sentences of the collocation being used.	If the learners click on collocation, the tool opens a new window with several example sentences of the collocation being used.

	Does the tool provide part of speech information in the output?	SKELL organizes the results around part of speech categories.	FLAX organizes the results around part of speech categories.	No	Just the word organizes the results around part of speech categories.	No
	Can the learners save the searches and examples?	No, only if the learners copy the results to a word file.	The tool has an option to save examples of the collocation in use.	No, only if the learners copy the results to a word file.	No, only if the learners copy the results to a word file.	No, only if the learners copy the results to a word file.
Feedback	Does the tool provide any type of feedback?	The tool does not provide any feedback to learners.	The tool does not provide any feedback to learners.	There is another tool in the Linggle suite that provides feedback to learners, but the collocations tool does not.	The tool does not provide any feedback to learners.	The tool does not provide any feedback to learners.
Applicability	Does the tool provide ready-made resources for the classroom?	No	FLAX has several different options of teaching applications, including an Android app and an option to pick collocations and develop didactics materials with them.	Yes, there are pre-prepared tasks available on the website.	No	No
Total		11.5	11	10.5	8	5.5

Table 2. Rating of the Collocation tools

Overall, the comparison between the tools showed that SKELL is the highest rated tool in terms of content quality and usability. FLAX is also a high rated tool, but it seems to be more useful for advanced learners. Linggle and Netspeak are difficult to implement in the classroom because of the search using wild cards, but Linggle still gained some points over Netspeak because it provides ready-made teaching materials. Finally, Just the Word is interesting, but it has a very similar output to the one presented in SKELL and SKELL has a cleaner interface. Considering this evaluation, in the next section we present four teaching ideas created to help the teacher include SKELL and FLAX in the ESL classroom. The goal of these classroom activities is twofold: familiarize learners with SKELL and FLAX, and aid learners in accurately producing collocations.

Pedagogical applications

Our final section provides suggestions for activities using the highest rated tools for searching and learning about collocations, FLAX and SKELL. Descriptions of each activity are below, followed by the activities themselves.

The first activity (5.1) targets academic collocations using FLAX. It is aimed toward upper intermediate and advanced learners and takes approximately 90 minutes. This activity begins with an introductory discussion, followed by a vocabulary activity in which FLAX is used to find frequent collocations of the vocabulary words. Learners then need to fill in gaps in a text using the collocations they have learned. Finally, learners complete a productive activity that requires them to use the new collocations in writing.

The second activity (5.2) also uses FLAX to target academic collocations but is targeted toward intermediate learners. The activity will take approximately 60 minutes. This activity begins with a short discussion about collocations. Learners then discuss the topic of the activity, password managers. Next, learners read a text with underlined collocations, highlighting those that they do not think are real. Learners then use FLAX to check the collocations and determine if their judgments match the results in FLAX.

Finally, learners are instructed to write comments and questions regarding password managers, incorporating the targeted collocations.

The third activity (5.3) aims to increase learners' familiarity with collocations and using the tool, SKELL. This activity is targeted toward pre-intermediate and intermediate learners and takes approximately 75 minutes. The activity includes a pre-listening discussion about multi-tasking, a TED Talk on monotasking, and an activity in which learners fill in collocations and provide examples from the listening or from SKELL, as well as the frequency of the collocation in SKELL. Finally, learners prepare a short talk arguing for or against monotasking, incorporating the targeted collocations.

The final activity (5.4) aims to help learners become familiar with looking up words by part of speech in SKELL by learners looking up collocations for nouns related to selfies. The activity is aimed at intermediate learners and takes approximately 60 minutes. Learners will discuss selfies, think of nouns related to selfies, use SKELL to find collocations for the nouns, and play a spoken game with the collocations.

Academic Collocations and Using FLAX

Goal: To teach academic collocations and how to use FLAX
Level: Upper intermediate/Advanced
Time: 90 minutes

Discussion

1. What are black holes?
2. Can you think of movies or shows that mention black holes? Why are they important in the storyline of these movies?
3. How do scientists study black holes?

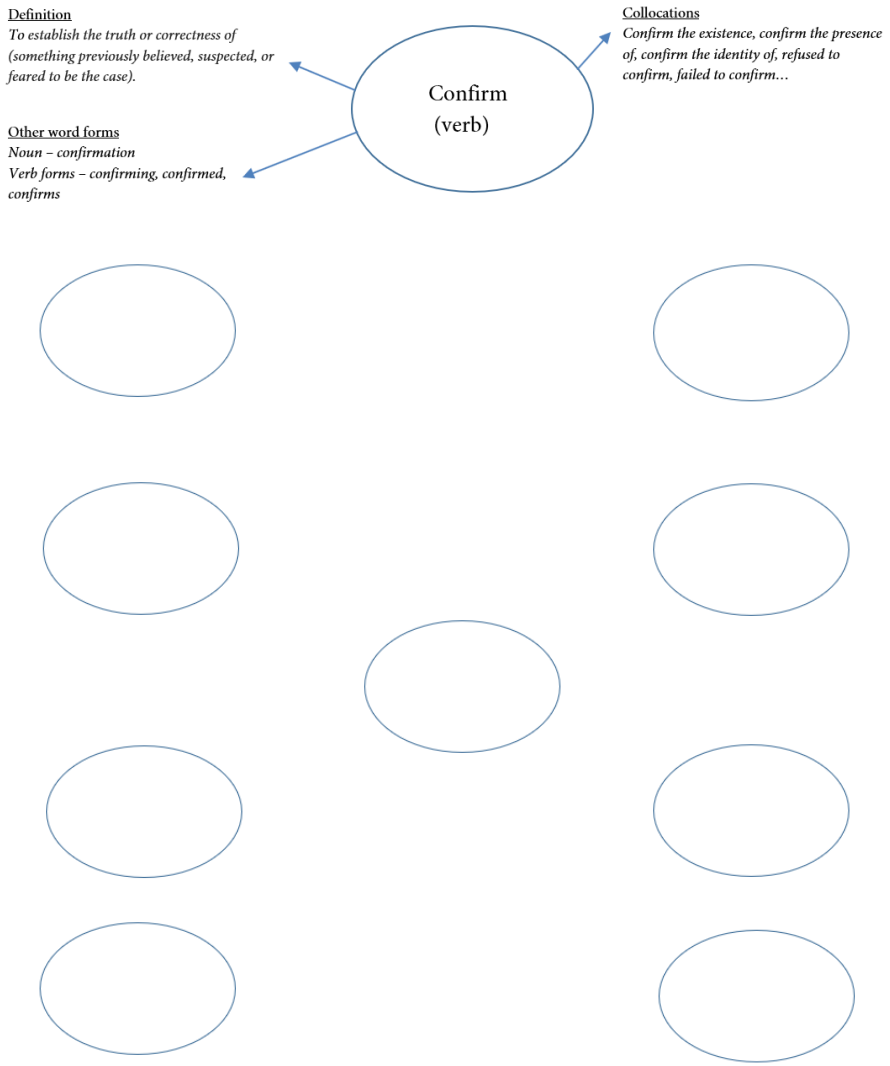
Vocabulary

1. The following list of words appears in the text we are going to read:

a) using an online dictionary find the definition of these words
b) using FLAX note the most frequent collocations of these words
c) complete the bubbles with what you know about each word

Circumstances (n)	Immediately (adv)	Energy (n)	Collapse (v)
Gravitational (adj)	Algorithm (n)	Breed (n)	Detect (v)
Finding (n)	Ultimately (adv)	Evidence (n)	Involve (v)

One example has been done for you.



2. Complete the collocation gaps in the text using FLAX and the list of collocations you created in exercise 1.

These black holes collide so hard they made space-time jiggle

By Paola Rosa-Aquino, September 3, 2020

[1] Some 7 billion light-years away, two black holes swirled closer and closer together over eons until they crashed together with a furious bang, creating a new black hole in the process. This disturbance in the cosmos caused space-time to stretch, collapse, and even jiggle, producing ripples known as gravitational _____ which reached our Earthly abode on May 21st of 2019.

[2] Using LIGO (Laser Interferometry Gravitational-wave Observatory), a pair of identical, two-and-a-half-mile-long interferometers in the United States, and Virgo, a roughly two-mile-long detector in Italy, an international team of scientists announced Wednesday that they had detected this cosmic collision, and it's racking up superlatives: it's the biggest, the farthest, and the most energetic black hole merger observed to date. This is also the first definite sighting of an intermediate-sized black hole, clocking in at about 142 times more massive than the Sun, forged from a black hole merger. The findings were _____ on Wednesday in a paper detailing the discovery in Physical Review Journals and another detailing the implications of the event in the Astrophysical Journal Letters.

[3] The merger signal, called GW190521, lasted only a tenth of a second—but scientists immediately _____ it was extraordinary in comparison to the low chirp of two colliding black holes LIGO detected in 2015, which confirmed Einstein's ineffable notions on space-time. "It's the biggest bang since the Big Bang that humanity has ever observed," says Alan Weinstein, an astronomer at the California Institute of Technology who was part of the study. It could offer clues as to why the Universe looks the way it does.

[4] _____ algorithms analyzed the signal, ultimately _____ scientists to pinpoint the masses of the merger and just how much energy was _____. The two progenitor black holes weighing in at about 66 and 85 solar masses merged into a black hole of 142 Suns. The remaining eight solar masses would have been converted into gravitational wave energy.

[5] Up until now, scientists have been able to detect and indirectly observe black holes in two different size ranges: stellar-mass black holes, which measure from a few solar masses up to tens of solar masses, and supermassive black holes that range from hundreds of thousands to several billions of times the mass of our sun. However, astronomers that detected GW190521 witnessed the birth of a _____breed of black hole: an "intermediate-mass" black hole. A few potential intermediate black holes have been spotted, but this is the first _____evidence of their existence.
[6] This strange signal was produced by the merger of two equally weird black holes: The heavier of the two merging black holes, at 85 solar masses, is the first black hole so far detected smack-dab in what is known as the "pair-instability mass gap." A star that collapses shouldn't be able to produce a black hole between the range of 65 to 120 solar masses because the most massive stars are obliterated by the supernova that comes hand in hand with their collapse. According to Weinstein, a possible explanation might be what astronomers call hierarchical mergers—when lighter stellar-mass black holes merge into heavier ones, which then merge into heavier ones still," consolidating until they become gargantuan black holes.
[7] Astrophysicist K.E. Saavik Ford says this finding is particularly exciting: "It's a bridge between the black holes that are formed directly when stars collapse and supermassive black holes that we find in the centers of galaxies." As Saavik Ford points out, it's actually very hard to make hierarchical mergers since black holes have to find each other, and then merge together. "That takes many, many, many lifetimes of the universe under anything like _____circumstances," Saavik Ford says, "so it had to have happened in a very dense stellar environment" like an active galactic nucleus or AGN.

Source: https://www.popsci.com/story/science/massive-black-hole-merger-gravitational-waves

3. Compare the collocations you used to a classmate's:
a) Did you use the same collocates?
b) If so, how did you come to the same choice of word as your classmate?

c) If not, how did the meaning of the text change based on the word you used?

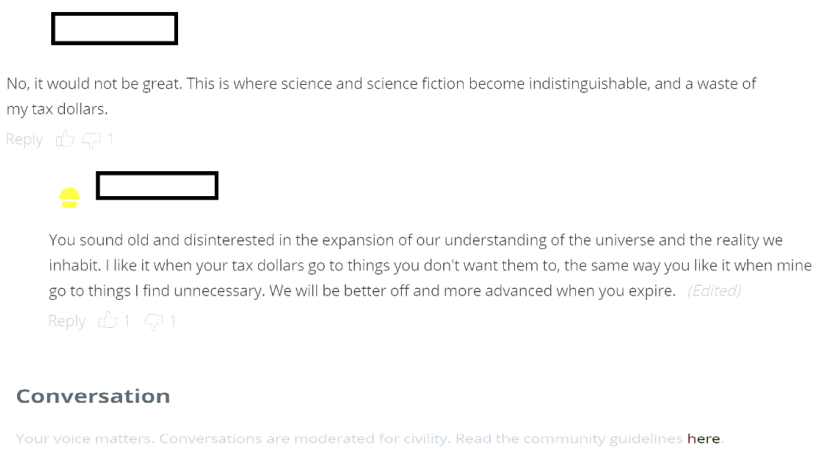
Writing

1. Look at the *conversation* section of the website:

a) What is the view of the first comment?

b) Does the second comment agree or disagree with the first one?

c) How would you respond to the first commenter? Write your comment in the conversation box and remember to use the collocations you found on FLAX.



Password Managers Activity (FLAX)

Goal: Notice the use of collocations in academic writing
Level: Intermediate
Time: 60 minutes
DIY: Password Managers

1. Collocations

In this class, we are going to learn about collocations. Before we get into it, discuss the following questions with your classmate:

a) Why do we say *heavy drinker*, but not *strong drinker*? Or *I am interested in*, but not *I am keen in*?
b) Can you think of other word combinations that always occur together? Which ones?
c) Why do you think it is important to learn these word combinations?

2. Password Manager

The text we are going to read talks about password managers, answer the following questions:

a) What is a password manager?
b) Why do people need it?
c) What are some of the dangers in using a password manager?

3. Noticing collocations

This text contains several collocations, all of them have been underlined for you.

a) Just by reading the text can you tell if these collocations are appropriate or not? Highlight the ones you don't think are real collocations.
b) Using FLAX (http://flax.nzdl.org/) check to see if these collocations are appropriate or not. Is there any mismatch between the ones you highlighted and what you found on FLAX?

How to get started using a password manager

By David Nield, September 8, 2020

Using a password manager is one of the best and easiest manners to keep your online accounts safe. If you're worried about making the jump, don't be—they're simple to set up and very much worth your while.

There might be slight differences between them, but all password managers work similarly. In our opinion, 1Password is one of the best available, so we'll go through that setup process so you know what to expect.

Signing up for a password manager

You can try 1Password for 30 days for free, but because it doesn't have a free tier, you will need to enter payment details to do so. After the trial period is up, it'll cost you at least $3 a month, billed annually. If you don't sense like expanding your list of paid services, LastPass and BitWarden have free tiers—the difference lies in the amount of features you'll be able to access, not the level of security protecting your information.

Registering for an online account is quite much the same no matter the platform, and we're going to assume you're fairly familiar with that process. What you really need to keep in mind when signing up for a password manager platform, though, is that you'll have to pick a master password.

Importing your passwords

Most password managers give you the option to import credentials from somewhere else, such as your browser. In the main 1Password portal on the web you can click your name (top right) then hit Import to get started.

This is certainly a good time-saver, but if you want to open again from scratch, that's fine too. Doing this will allow you to filter out those old and redundant logins that you may not want to carry over to your new password manager.

Editing settings and credentials

As you would expect, your password manager will come with a bunch of settings to explore. We'd recommend checking them out once you've got to grips to the basics of the software.

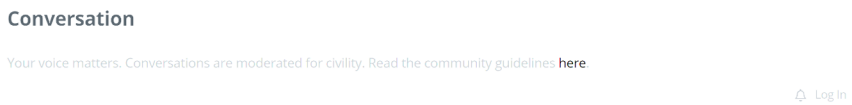
You can set a period of inactivity after which the desktop and mobile apps automatically lock. It's a good idea to set this as low as possible, just in case you briefly step away from your laptop or your phone.

Source: https://www.popsci.com/story/diy/password-manager-guide/

4. Writing

After reading the text do you think you would use a password manager, or do you still have questions to the author?

Use the conversation box to write your comments and questions about password managers. Try to incorporate collocations you saw in the reading.



Multitasking versus Monotasking (SKELL)

Goal: Practice looking up collocations in SKELL and use them to prepare a short talk presenting arguments for or against monotasking.

Level: Pre-intermediate and intermediate

Time: 75 minutes

Multitasking versus Monotasking

1. Pre-speaking activity.

Discuss the following questions:

Do you know what multitasking is?
What are some of the benefits and drawbacks of multitasking?
Do you think you are good at multitasking?

2. Watching the TED talk

You will watch a short TED talk[3] about monotasking as an alternative to multitasking.

3 https://www.youtube.com/watch?v=0YNeyBANrTI&ab_channel=TED

1st time watching:

Do you agree with the speaker's ideas about multitasking?
Do you see monotasking as a good alternative to multitasking?

2nd time watching:

The following target words will appear in the video:

Target word (s)	**Word following the target**	**Frequency**	**Example in context**
fly (v.)	[through]		
design (n.)	[process]		
story (n.)	[about]		
multitasking (adj.)	[environment]		
multitasking (adj.)	[life]		
sense (n.) (of)	[adventure]		
consider (v.) (the)	[option]		

Try to note the words that are used right after the words on the list and write them down in the right column. Using SKELL, find out the frequency of the collocation per million words and write an example of the collocations used in context (either from the video or from SKELL). What do you notice about the collocation frequencies in the list?

3. Prepare a 1-2-minute talk (similar to the one you just watched) arguing for or against monotasking. Make sure to include the collocations you've recorded in your talk and be ready to present.

The Selfie Culture (SKELL)

Goal: Practice looking up collocations for certain noun words and using them in speech
Level: Intermediate - Upper-intermediate
Time: 60 minutes

The Selfie Culture

1. Pre-speaking activity

Answer these questions in small groups:
- Do you take selfies?
- What is the best part about taking a selfie?

2. Brainstorming & working with SKELL

- Brainstorm nouns on the topic of selfies. Try to come up with 5-6 nouns.
- Using the **SKELL Word Sketch** function, look up at least 2 collocations for each noun you've brainstormed. Make sure that one of these collocations has a verb and the other has an adjective. Write down the nouns and their collocations in the handout provided (available on the next page).

3. Applying the collocations game

Go back to your small groups. Discuss the following question in your groups:

What are some of the issues you see in today's selfie culture?

Your task is to participate in the discussion using the collocations in your handout that you've found using SKELL. Cross a collocation off the list once you use it in the discussion.

The first person to cross all the collocations off the list wins the game.

Collocations recording sheet

Noun	Collocations	Examples
1.	**with a verb:** **with an adjective:**	
2.	**with a verb:** **with an adjective:**	
3.	**with a verb:** **with an adjective:**	
4.	**with a verb:** **with an adjective:**	
5.	**with a verb:** **with an adjective:**	
6.	**with a verb:** **with an adjective:**	

References

Bahns, J. & Eldaw, M. (1993). Should we teach EFL students collocations?. *System*, *21*(1), 101-114.

Baisa, V. S. (2014). SkELL – Web interface for English Language Learning. *In Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, pp. 63-70. ISSN 2336-4289.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing*, 37, 1-12.

Boisson, J., Kao, T., Wu, J., Yen, T. & Chang, J. (2013). Linggle: a Web-scale Linguistic Search Engine for Words in Context. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 139–144, Sofia, Bulgaria, August 4-9 2013. c 2013 Association for Computational Linguistics.

Boers, F., Demecheleer, M., Coxhead, A. & Webb, S. (2014). Gauging the effects of exercises on verb–noun collocations. *Language Teaching Research : LTR*, *18*(1), 54–74. https://doi.org/10.1177/1362168813505389

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Chan, T. P. & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. *Computer assisted language learning*, *18*(3), 231-251.

Chang, Y. C., Chang, J. S., Chen, H. J. & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. Computer Assisted Language Learning, *21*(3), 283-299.

Chen, H. J. H. (2011). Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, *24*(1), 59-76.

Chen, X. (2016). Evaluating language-learning mobile apps for second-language learners. *Journal of Educational Technology Development and Exchange (JETDE)*, *9*(2), 3.

Cobb, T. (2018). From corpus to CALL: The use of technology in teaching and learning formulaic language. In Anna Siyanova-Chanturia, Ana Pellicer-Sánchez (Edited) *Understanding formulaic language: A Second Language Acquisition Perspective* (pp. 192-210). Routledge.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 70–86). London: Language Teaching.

Durrant, P. & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58-72.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in second language acquisition*, *18*(1), 91-126.

Ellis, N. C., Simpson-Vlach, R. I. T. A. & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL quarterly*, *42*(3), 375-396.

Fitzgerald, A., Wu, S. & Marín, M. J. (2015). FLAX: Flexible and open corpus-based language collections development. In K. Borthwick, E. Corradini, & A. Dickens (Eds), *10 years of the LLAS elearning symposium: Case studies in good practice* (pp. 215-227). Dublin: Research-publishing.net. doi:10.14705/rpnet.2015.000281

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, *67*(S1), 155-179.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. *Phraseology: Theory, analysis and applications*, 145-160.

Haughey, M. & Muirhead, B. (2005). Evaluating learning objects for schools. *E-Journal of Instructional Science and Technology*, 8(1), n1, 1-23.

Jiang, J. (2009). Designing pedagogic materials to improve awareness and productive use of L2 collocations. In *Researching collocations in another language* (pp. 99-113). Palgrave Macmillan, London.

Kay, R. (2011). Evaluating learning, design, and engagement in web-based learning tools (WBLTs): The WBLT Evaluation Scale. *Computers in Human Behavior*, *27*(5), 1849-1856.

Kay, R. H. & Knaack, L. (2009). Assessing learning, quality and engagement in learning objects: the Learning Object Evaluation Scale for Students (LOES-S). *Educational technology research and development*, *57*(2), 147-168.

L'Huillier, M. (1990). Evaluation of CALL Programs for Grammar. *Computer Assisted Language Learning*, 1(1), 79-86.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Men, H. (2018). *Vocabulary Increase and Collocation Learning*. Springer.

Nesbitt, D. (2012). Student evaluation of CALL tools during the design process. *Computer Assisted Language Learning*, *26*(4), 371-387.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, *24*(2), 223-242.

Nesselhauf, N. (2012). Exploring the phraseology of ESL and EFL varieties. In T. Herbst, S. Faulhaber & P. Uhrig (Ed.), *The Phraseological View of Language: A Tribute to John Sinclair* (pp. 159-178). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110257014.159

Nurmukhamedov, U. (2015). *An evaluation of collocation tools for second language writers* (Doctoral dissertation, Northern Arizona University).

Paquot, M. & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32(Mar), 130–149. https://doi.org/10.1017/S0267190512000098

Potthast M., Trenkmann M. & Stein B. (2010) Netspeak: Assisting writers in choosing words. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. van Rüger & J. Rijsbergen (eds.) *Advances in Information Retrieval: Proceedings of 32nd European Conference on Information Retrieval* (ECIR 10), 672. Springer. Available at http://www.uniweimar.de/medien/webis/publications/papers/stein_2010e.pdf.

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO journal*, *34*(2), 243-258.

Smith, P. L. & Ragan, T. J. (2004). *Instructional Design*. Wiley.

Tsai, M.-H. (2020). The effects of explicit instruction on L2 learners' acquisition of verb–noun collocations. *Language Teaching Research : LTR*, 24(2), 138–162. https://doi.org/10.1177/1362168818795188

Trofimovich, P. & Isaacs, T. (2013). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15(4), 905-916.

Wolter, B. & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, *35*(3), 451-482.

Woolard, G. (2000). Collocation-encouraging learner independence. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 28-46). Hove, England: Language Learning Publication.

Yamashita, J. & Jiang, N. A. N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *Tesol Quarterly*, *44*(4), 647-668.

Yoon, H. & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of second language writing*, *13*(4), 257-283.

Appendix A: Rating Criteria

Criteria	Definition
Content quality	
Definition of Collocation	**What is the definition of collocation used?**
Linguistic research	**Reference Corpus**
	Types of texts
	Frequency threshold
	Measure of association
Language Variation	**Does it account for register variation?**
	Does it account for disciplinary variation?
Filter	**Is there a teaching filter in place?**
Interface	
Intuitivity	**Does the tool have clear menus and icons to facilitate navigation?**
	Can learners use it without in depth training?
	How many clicks are necessary to finalize the search?
Design features	**Is the interface appealing to students?**
Presentation of search outcomes	
Search	**How is the search structured?**
	Does it provide collocates to both sides of the word?
	Does it account for intervening words?
	Does the tool current misspelled words in the search?
	Can the search be limited by part of speech?

Presentation	**Which criteria is used to determine the order of presentation?**
	Does the tool provide frequency and dispersion for the collocations?
	Does the tool show examples of the collocation being used in context?
	Does the tool provide part of speech information in the output?
	Can the learners save the searches and examples?
Feedback	**Does the tool provide any type of feedback?**
Applicability	**Does the tool provide ready-made resources for the classroom?**
Total	

Driving forces to adopt EMI: scholars' perceived benefits of English medium of instruction in Brazilian higher education

Laura Baumvol (UBC-CA/UFRGS)
Lucas Marengo (UFRGS)
Simone Sarmento (UFRGS)

Introduction

Over the last decades, internationalization of higher education (HE) has become a high priority for policymakers and HE institutions (HEIs) (Knight, 2008). In countries situated in the geolinguistic global periphery, like Brazil, however, internationalization of HE must go beyond the system of prioritizing only academic mobility and shift to one which benefits a wider audience. The process of Internationalization at Home (IaH) has been seen as a counteract to the increased emphasis on academic mobility and an alternative for a more inclusive internationalization process (Baumvol & Sarmento, 2019; Beelen & Jones, 2015; de Wit et al., 2015; Teekens, 2007). IaH emphasizes the intercultural and international dimensions in the teaching and learning processes and research, the extracurricular of international students and teachers into local academic life, as well as the enhancement of education and research as a whole (Knight, 2008; de Wit et al., 2015). In fact, IaH is a paradigm for the development of strategic institutional internationalization policies, as it encourages respect for diversity while developing people "with a cosmopolitan mindset, with communication skills between and across cultures, at home" Teekens (2007: 6).

Within IaH processes, additional languages, especially English, play a key role in giving access to students and teachers to international practices while in their own countries and institutions. Teaching undergraduate

and graduate courses in English is one of the main strategies to internationalize HE in non-English dominant contexts.

EMI is a crucial part of IaH processes and can be defined as the use of the English language to teach academic content in countries or places in which English is not the language spoken by the majority of the population, i.e., non-English dominant contexts. Internationalization and globalization of education are usually the driving forces of EMI (Dearden, 2014; Gimenez et al., 2018; Macaro, 2018; Pecorari & Malmström, 2018). Considering the growing demand for more internationalized academic environments, this investigation aims to identify (1) whether EMI is present in Brazilian HE and (2) the perceived benefits of teaching in English. Data were collected through an electronic questionnaire sent out to Brazilian HE teachers[1]. The analysis compares the perceptions of teachers across eight fields of knowledge according to the classification of Brazilian funding agencies (Agricultural Sciences, Applied Social Sciences, Biological Sciences, Engineering, Exact and Earth Sciences, Health Sciences, Human Sciences, Linguistics, Literature, and Arts*).*

Prior research has focused on teachers' perceptions of EMI in different global contexts (Briggs & Dearden, 2018; Chapple, 2015; He & Chiang, 2016; Orduna-Nocito & Sánchez-García, 2022; Tatzl, 2011; Tran et al., 2021; Tsuchiya & Pérez Murillo, 2019; Wächter & Maiworm, 2014; Werther et al., 2014; Yeh, 2014). However, to our knowledge, this is the first large-scale study focusing on the Brazilian context. First, the importance and advantages of EMI will be highlighted. Next, the methodological procedures used for data collection and analysis will be introduced. Finally, the results will be presented and discussed along with concluding remarks.

1 The term "teachers" used throughout this paper includes both professors and researchers working in HE institutions.

EMI in the Context of Higher Education

Over the last decades English has achieved the status of global scientific and academic lingua franca (Ammon, 2010; Baumvol et al., 2021; Crystal, 2003; De Swaan, 2001; Jenkins, 2013; Lillis & Curry; 2010; Montgomery, 2013; Solovova et al., 2018). According to Hyland (2015), English is used in 95% of all the publications in the *Science Citation Index* (SCI). In a similar fashion, HE programs and courses all over the world have increasingly been adopting EMI in varied academic practices. Therefore, to better participate in these practices which happen largely in English, academics from all continents should have some mastery of the English language.

Muñoz (2012) suggests that the greater use of English contributes to establishing an environment that, indirectly, leads to language proficiency development. Individuals construct their dialogical relations in socially co-constructed practices using language (Clark, 1996) and, thus, English learning is grounded in interaction. The adoption of EMI could bring considerable linguistic benefits because instructors and students can take part in authentic language practices that require the use of English. This may lead to improvement in their proficiency for various practical purposes, such as participating in academic events, in Massive Open Online Courses (MOOCs), and in exchanges with international research partners. To join EMI classes, however, students are expected to already have a working knowledge of the English language. Important to point out that, although EMI can help improve students and teachers' English language proficiency levels, in EMI settings language learning is usually considered a by-product of the extensive use of English in the classroom and not its main goal (Airey, 2016).

The driving forces behind the implementation of EMI can be manifold. Internationalization is usually a primary motivation, so much so that in some cases EMI is believed to be an indicator of whether a HEI is internationalized (Jordão & Martinez, 2021). Apart from contributing to teachers and students' English language proficiencies (Briggs & Dearden, 2018), other perceived benefits of adopting EMI include increasing recruitment of international students, providing access to intercultural and international

learning materials (Liu & Fang, 2017), as well as creating opportunities for the students to enter a global academic and entrepreneurial community (Dearden, 2014). Furthermore, Hu and Lei (2014) state that the expansion of EMI in Asia has been considered advantageous because it allows, concurrently, for the learning of content itself and for the development of the English language for both students and teachers.

When examining the European context, Wächter and Maiworm (2014) indicate that the motivation for EMI comes from the need of engaging students from other countries and preparing local students for international mobility and for the international labor market, as well as from the target of elevating the profiles and the positions of the universities in rankings. In Asia, the governments of Indonesia, China, and Japan have implemented language policy and planning reforms over the last years to widely implement EMI to encourage students' English fluency (Indonesia) and to stimulate the internationalization of top universities (Japan) (Walkinshaw et al., 2017).

Regarding the Brazilian context, Gimenez et al. (2018) have shown that only a few isolated initiatives of EMI are being offered in Brazil, especially at the postgraduate level. It is important to note, however, that in Brazil, English proficiency is intrinsically related to social class. Disadvantaged students usually only have access to English classes in regular schools, which, in many scenarios would be good enough, but not in Brazil and the causes are manifold (Baumvol & Sarmento, 2019). First of all, there is a belief that additional languages are not to be learned in the official regular schools, making teachers demotivated from the start. Second, classes are large and there is usually only one hour of English class a week, making it impossible to acquire fluency. Also, public school teachers are underpaid in the country and, to counterbalance their low salaries, have to take more than one job and work very long hours, leaving no room for professional development. Therefore, the teaching of English has been relegated to the private sector, with over 6,000 private language courses in the country, with an annual increase of 15% (Windle & Nogueira, 2015). There are different types of private language courses, covering a variety of price ranges, hence, catering for different social classes, but not all of them. Considering

this, it is important that HEIs take into consideration the different levels of English proficiency of post-secondary students and even teachers, offering English classes to improve their proficiency before or while adopting EMI.

At the same time, the adoption of EMI has also faced a number of criticisms. For instance, Airey (2011) highlights that there is not enough support to ensure an increase in quality when English becomes the language of instruction at the post-secondary level. In addition, weakening the use of local languages in education could lead to problems in the expansion of the disciplinary use of local languages, domain loss and diglossia, and parallel language use (Jenkins, 2013; Josephson, 2005). Despite these criticisms, EMI is, as Macaro (2015) puts it, an "unstoppable train" and has been a growing trend in many parts of the world (Airey et al., 2017; Coleman, 2006; Martinez, 2016; Richard & Pun, 2022) and should, therefore, be further investigated.

Methodology

Data for this study were collected through an online questionnaire composed of 66 questions which was sent to HE teachers working in different types of Brazilian HEIs (i.e., public, private, research universities, technical institutions, and colleges) between May and October 2017. The design of the instrument was based on an extensive literature review about the role of languages in the internationalization of HE globally and in the Brazilian context, as well as on informal interviews with teachers from varied fields of knowledge working in Brazilian HEIs (Baumvol, 2018).

The identification of potential participants to whom the questionnaire was sent to was based on the Lattes Platform, an initiative of the National Council for Scientific and Technological Development (CNPq) which aims to integrate academic curricula databases of academics into a single platform. Participants were recruited so as to respect proportions related to the field of knowledge and location of the HEIs, i.e., in which state the Brazilian HEI is located. Thus, 29,747 online questionnaires were sent by email (10% of the cohort), out of a total of 297,515 Lattes CVs of teachers with a PhD and affiliated with a Brazilian HEI. By the end of the

process, 5,119 valid responses had been collected, representing a return rate of 17.2%. Regarding the fields of knowledge, the Lattes Platform categorizes researchers in the following major fields: Agricultural Sciences, Applied Social Sciences, Biological Sciences, Engineering, Exact and Earth Sciences, Health Sciences, Human Sciences, Linguistics, Literature, and Arts, Others, and Technologies. As there were no CVs registered under the fields of knowledge "Other" and "Technologies", only the other eight major fields were considered. The present study examines two questions of the questionnaire:

RQ1. Have you ever taught classes in English?

RQ2. In your opinion, what are the main benefits of classes taught in English at Brazilian higher education institutions?

The two questions were closed-ended questions. The first one allowed only (a) yes or (b) no answers, a multiple-choice type of question on Google Forms. In the second question, participants could select more than one of the following nine options, a check-boxes type of question: (A) students improve their level of English proficiency; (B) teachers improve their level of English proficiency; (C) classes take place in the language in which scientific and academic knowledge is disseminated; (D) students have an experience of internationalization, even though they are in Brazil; (E) teachers have an experience of internationalization, even though they are in Brazil; (F) students will be better prepared for their professional future and for the job market; (G) better quality of teaching in Brazilian HEIs; (H) foreign students can participate in classes and (I) there are no benefits. The answers to the two questions were compared across the eight fields of knowledge to allow for the understanding of each field's characteristics.

Results

This section presents the results of the responses to the two previously mentioned questions. The answers to the first question focused on finding out whether teachers from different fields of knowledge have or have not previously taught classes in English, i.e., adopted EMI, can be seen in Table 1.

Have you ever taught classes in English?			
Field of Knowledge	**Total of respondents**	**Yes**	**No**
Agricultural Sciences	446	63 (14.1%)	383 (86%)
Applied Social Sciences	656	89 (13.6%)	567 (86.4%)
Biological Sciences	520	82 (15.8%)	438 (84%)
Engineering	457	80 (17.5%)	377 (82.5%)
Exact and Earth Sciences	735	90 (12%)	645 (87.8%)
Health Sciences	814	110 (13.5%)	704 (86.5%)
Human Sciences	822	49 (6.0%)	773 (94.0%)
Linguistics, Literature, and Arts	257	65 (25.4%)	192 (74.7%)
TOTAL 4706		**13.5%**	**86.5%**

Table 1. Status of teachers regarding the use of EMI in class across the eight fields of knowledge.

All fields of knowledge have a much higher number of academics who have not yet taught in English. Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences, Social and Applied Sciences, and Engineering show a much closer pattern; in this case, between 13.5% and 15.8% of the teachers have taught classes in English. The field of *Human Sciences* has the lowest number of teachers who have taught classes in English, only 6%. On the other hand, *Linguistics, Literature, and Arts* has the highest number of academics who have adopted EMI (25.4%). Such behavior of the latter was expected, since many of the courses in this field such as Literature and English language teaching courses, are part of a TESOL major and, thus, taught in English throughout the undergraduate programs.

In relation to the low percentage of teachers who have taught classes in English, Dearden (2014) suggests that many of them may not even be aware of any EMI policy in their universities. In the case of Brazil, however, there is apparently also a lack of language policies around EMI (Gimenez et al., 2018), a fact that may corroborate teachers' lack of knowledge about it.

The second question examined in this study focused on the main benefits of teaching in English. As mentioned before, nine options were offered to respondents and they could choose all that applied. In Figure 1 you can see the percentages for each answer.

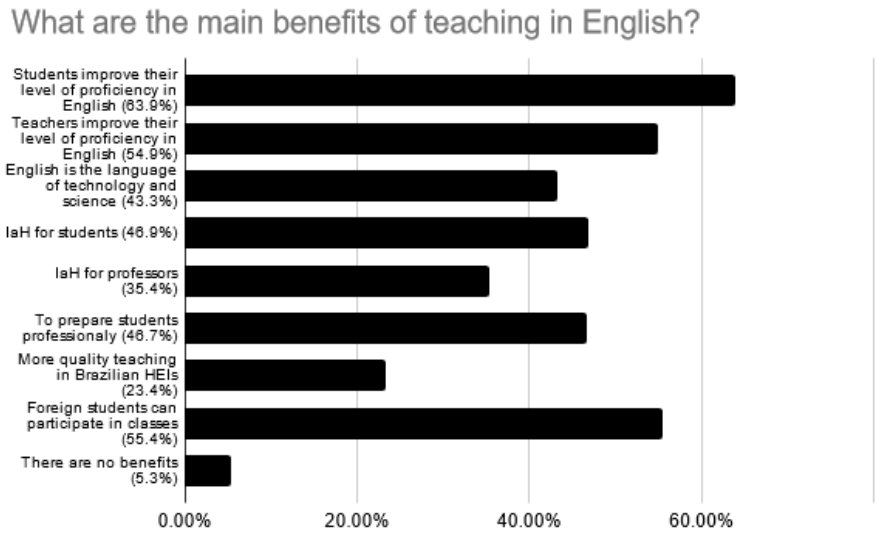


Figure 1. Perceptions of main benefits of EMI by respondents

In the respondents' opinion, the main benefit of offering classes in English was that Brazilian students could improve their level of proficiency in English (63.9% of the valid responses). The second most recurrent response was that foreign students could participate in EMI classes, with 55.4% of the responses. Following close was the one that mentioned the improvement of the level of fluency in English by the teachers themselves, with 54.9%. The other five options all had an incidence below 50%: IaH for students with 46.9% of responses; preparing students for their professional future with 46.7%; English as a language of science and technology with 43.4% of responses; IaH for teachers with 35.4% of responses, and, finally, improving the quality of classes, with 23.4%. The answer with the lowest number of informants, on the other hand, was that there were no benefits in teaching in English, with 5.3%.

According to the results, teachers perceive that classes taught in English might lead to improvements in the students' language skills. In this respect, Martinez (2016) acknowledges that students and teachers' proficiency is a recurring issue in the implementation of EMI. A study conducted with lecturers of an English-medium university in Turkey showed that they acknowledge the linguistic benefits of EMI (Collins, 2010), while an investigation in an Austrian HEI indicated that lecturers understand that students are encouraged to practice the language in EMI courses and then feel more confident in speaking skills (Tazl, 2011). In Japan, a mixed-method study using questionnaires and interviews with teachers revealed that EMI courses are mainly implemented to improve the English proficiency of HE students (Chapple, 2015). Another mix-methods research examined the perceived impact of EMI approaches on students' English language proficiency in Vietnamese HE (Tran et al., 2021). The authors found that most lecturers noticed an improvement in students' English language ability. According to Tran et al. (2021: 20), "students' language proficiency was improved because they used English as an everyday habit in class and during lesson preparations and having lectures in EMI classes". Finally, in a survey conducted by Martínez and Chichón (2020) in a Spanish medium-sized state university, 82% of the lecturers reported that students' English improves when they attend courses taught in English. Studies on students' perceptions of English proficiency in EMI settings point in the same direction, showing a perceived enhancement in their English proficiency levels (Tatzl, 2011; Wächter & Maiworm, 2014; Yeh, 2014).

The second most common benefit of EMI chosen by teachers in our study was that classes in English attract foreign students. Wächter and Maiworm (2014), when examining EMI programs across Europe, showed that one of the main motivations for the implementation of English-taught programs was to attract students from other countries. An investigation into the challenges faced by post-secondary international students in China also pointed to the importance of EMI to allow these foreign students to pursue their studies (He & Chiang, 2016). Thus, EMI is viewed as a way to increase the mobility of international students, aiming for the internationalization of these HE academic settings.

In regards to teachers' English language proficiency, 95% of the Spanish lecturers who responded to Martínez and Chichón's (2020) survey reported that teaching in English helps these lecturers improve their own language proficiency. In an investigation carried out in the Northern European context (Henriksen et al., 2018), some of the interviewed lecturers viewed the implementation of EMI in their HEIs as a good opportunity to improve their English proficiency levels. These results align with the third most recurrent response in our study (54.9% of the responses), according to which EMI could help improve teachers' English proficiency.

"IaH for students" and "To prepare students professionally" had very similar outcomes (with 46.9% and 46.7%, respectively). In fact, when integrating international and intercultural dimensions into the curriculum "at home", IaH can "enhance the quality of education and research for all students and make a meaningful contribution to society" (De Wit et al., 2015: 29). Therefore, respondents seem to be stating that classes taught in English benefit the IaH process since they aid students to better interact both globally and within the local community. The results reported by Botha (2014) when investigating students' perceptions of the Chinese EMI context point in the same direction. Almost 80% of the students who responded to a survey conducted at Sun Yat-sen University (SYSU) strongly agreed that English "internationalizes" their university (Botha, 2014). Concerning the preparation of students for their future careers, Briggs and Dearden's (2018) results indicate that preparing students for their professional lives was generally highly ranked among the teachers who completed the survey.

The option regarding the use of English as the global language of science and technology was chosen by 43.3% of the respondents. This aligns with several other investigations which have acknowledged the status of the global scientific language achieved by English (Ammon, 2010; Crystal, 2003; De Swaan, 2001; Jenkins, 2013; Lillis & Curry, 2010; Montgomery, 2013; Solovova et al., 2018). The lecturers of 10 HEIs across Europe who participated in a study conducted by Orduna-Nocito and Sánchez-García (2022) recognized the role of English as a Lingua Franca in reading and writing research papers, as well as in conferences and in research in general.

The option which stated that one of the benefits of teaching in English is IaH for teachers received 35.4% of responses. In this way, EMI also means

qualification opportunities for academics within the reality of globalization currently experienced in the country. Next, respondents chose "more quality teaching in Brazilian HEIs" (23.4% of responses). These results align with Hu and Lei's (2013) ideas regarding the Chinese context, in which EMI has been promulgated by the Ministry of Education as a "key police initiative improving the quality of undergraduate education in Chinese higher education since the turn of the 21st century" (2013: 557). In addition, the results of the study by Briggs and Dearden (2018) showed that, for teachers, the primary goal of teaching in English was providing home country students with a high level of education.

Finally, the answer with the lowest incidence was the one stating there are no benefits in teaching in English, with only 5.3% of the responses. In this respect, Brigss and Dearden (2018) found that 21.8% of the 167 respondents to their survey (EMI teachers) believe that EMI is not beneficial, a figure substantially higher than in our study. This shows that resistance to EMI in Brazil does exist, but there may be stronger obstacles to its implementation in Brazilian HE settings. A possible explanation for this resistance might be that EMI has been controversial because of political and pedagogical reasons, "including the desire to protect national languages and cultures, a concern that policies had not been clearly thought through, and that EMI was potentially divisive and could lead to social inequalities" (Dearden, 2014: 4).

In conclusion, for the vast majority of teachers in our study, the main benefits of classes taught in English are the enhancement of English proficiency for students and the possibility for foreign students to participate in classes. These results align with Tatzl (2011), Wächter and Maiworm (2014), Yeh (2014), and He and Chiang (2016), which also show a general belief that students can improve their proficiency and that international students can join classes taught in English.

We will now present the same data of Figure 1, i.e., the check-boxes responses to the question "What are the benefits of teaching in English?", but this time focusing on the comparison between the eight fields of knowledge. Again, the options were: (A) Students improve their level of proficiency in English; (B) Teachers improve their level of proficiency in English;

(C) Classes take place in the language in which scientific and academic knowledge circulates most; (D) Students have an experience of internationalization even though they are in Brazil; (E) Teachers have an experience of internationalization even being in Brazil; (F) Students will be better prepared for their professional future and for the labor market; (G) More quality teaching in Brazilian HE institutions (HEIs); (H) Foreign students can participate in classes and, finally, (I) There are no benefits. Based on the options provided, the results are shown in Table 2 below.

Benefits of EMI									
Field of Knowledge	**Option A**	**Option B**	**Option C**	**Option D**	**Option E**	**Option F**	**Option G**	**Option H**	**Option I**
Agricultural Sciences	71%	61%	50%	51%	33%	51%	30%	53%	4,2%
Applied and Social Sciences	56%	48%	39%	47%	33%	40%	20%	56%	4%
Biological Sciences	73%	62%	53%	52%	39%	58%	28%	64%	5%
Engineering	67%	57%	53%	47%	37%	52%	23%	70%	3%
Exact and Earth Sciences	70%	57%	50%	46%	33%	51%	19%	61%	3%
Health Sciences	67%	64%	46%	49%	41%	48%	30%	56%	4%
Human Sciences	51%	43%	26%	38%	27%	31%	18%	41%	10%
Linguistics, Languages, and Arts	58%	46%	37%	47%	34%	40%	18%	50%	8%

Table 2. Perceptions of main benefits of EMI by respondents across eight fields of knowledge.

Options (D), Students have an experience of internationalization even though they are in Brazil and (E), Teachers have an experience of internationalization even being in Brazil dealt with the idea that students and

teachers have the opportunity of IaH, that is, using the English language without leaving the country (Beelen & Jones, 2015; Baumvol & Sarmento, 2016, 2019). Teachers in the field of *Human Sciences* were the ones who gave the least importance to IaH as a benefit of EMI, with 38% and 27% for each of the options. In contrast, the fields of Biological Sciences, Agricultural Sciences, and Health Sciences showed higher numbers for IaH, with 52%, 51%, and 49%, correspondingly. When it comes to option (G) Quality of teaching due to the adoption of EMI, teachers from the *Human Sciences and Linguistics, Literature and Arts* were the ones who gave the least importance to this benefit (18%); while *Agricultural Sciences* and *Health Sciences*, for instance, had 30% of responses in this respect.

With regards to option I (no benefits in adopting EMI), the results showed that the fields of *Human Sciences* and *Linguistics, Literature,* and *Arts* were the ones with the highest percentages, with 10% and 8% of responses correspondingly. Even though these numbers are also low, when compared to the percentages of the other six fields of knowledge, results are twice as high as the other areas, since all the other fields had 5% (*Biological Sciences*) or less (all other fields). The two fields with the lowest figures in this option were *Exact and Earth Sciences* and *Engineering*, both with only 3% of participating teachers. These numbers may point to a difference in terms of resistance towards EMI, with the "softer" sciences, here represented by *Human Sciences* and *Linguistics, Literature,* and *Arts* presenting the higher resistance.

Overall, the "softer sciences" are those that least perceived EMI as a practice that brings benefits. When analyzing the number of responses from the *Human Sciences* in relation to other questions, such as "foreign students can participate in classes", this field had 41% of responses compared to 70% of teachers in the field of *Engineering* and 64% of teachers from *Biological Sciences*. The results suggest a pattern concerning the (non-)benefits of EMI. While the "harder" sciences (*Biological Sciences, Agricultural Sciences, Health Sciences, Engineering,* and *Exact and Earth Sciences*) had a higher acceptance of EMI, the fields in the "softer" sciences had lower figures regarding the possible benefits of EMI in HE classrooms.

Conclusion

The analysis of the first research question, which asked whether participants had taught classes in English, showed that most respondents (86.5%) had never taught classes in English, while only 13.5% answered that they had already done so. In fact, the British Council/FAUBAI Guide (Gimenez et al., 2018) shows the practice of EMI is still incipient in Brazil, as between 2017 and 2018, there were only 1,011 courses taught (undergraduate and graduate) in English across the country. Considering Brazil has 2,457 HEIs which offered 41,953 full Programs and countless number of courses (something like hundreds of thousands of courses) in 2020, roughly 1,000 courses offered in English point to a reality that EMI in Brazil happens due to isolated initiatives and is not part of an organized language education policy. Thus, while in some countries EMI is being considered an "unstoppable train" (Macaro, 2015), in Brazil EMI is a train still to be caught. We, scholars from applied linguistics in Brazil, do talk (for or against) extensively about the phenomenon, however, the phenomenon seems to hardly exist.

Comparing the different fields of knowledge, *Linguistics, Literature, and Arts* is the one with the most respondents who have taught in English (25.3%). All other fields had a percentage lower than 20%. As mentioned earlier, we believe that the main reason why the field of *Linguistics, Literature, and Arts* is the one that has the most taught classes in English is the fact that an additional language is the major, such as in TESOL programs. Thus, several courses in the curriculum are taught in English, like English literature, for instance. Conversely, *Linguistics, Literature, and Arts* is the second area which believes there are no benefits in EMI. *Engineering, Biological Sciences,* and *Agricultural Sciences* are next, with 17.5%, 15.8%, and 14.1% respectively. The areas of Applied Sciences, Health Sciences, and Exact and Earth Sciences comprise 13.6%, 13.5%, and 12.3%. Finally, the lowest incidence of classes taught in English was in the *Human Sciences*, with only 6% of the teachers answering they had already had this experience.

The responses to the second question, which asked for opinions about the main benefits of teaching classes in English, demonstrated that

the option with the highest percentage was to increase students' level of English proficiency. Before the analysis, and according to the literature in the field, we expected that the main perceived benefit would be attracting international students to Brazil (Macaro, 2015; Martinez, 2016; Wächter & Maiworm, 2014), but this came in second. Engineering was the only area that matched our expectation, as 70% of participants from this area considered that classes taught in English could enable international students to participate in the courses, compared to 67% of incidence for students improving their proficiency in English. Even though improving language proficiency is usually only a by-product of EMI classes, in a context such as Brazil, it might be a good idea for content teachers to have some knowledge of language issues so that learning can be facilitated. It is here that EAP teachers can act together with content teachers, this type of partnership is a pre-requisite for a successful implementation of EMI.

The least chosen option for this same question was the one that stated that there are no benefits to the EMI approach (only 5.3% of the responses). Different fields of knowledge have higher percentages than others in their perception of EMI. For instance, *Human Sciences* and *Linguistics, Literature, and Arts* (fields of the "softer" sciences) had respectively 10% and 8% of responses pointing to non-benefits in adopting EMI, while fields such as *Exact and Earth Sciences* and *Engineering* ("harder" sciences) had only 3%.

Whereas in some countries we can notice local languages being threatened by the widespread use of English in HE, it is our belief that in Brazil we still face a different problem: the one of inclusion. As the majority of the academic practices are only held in Portuguese, proficiency in academic English is a privilege of only a few students whose families can afford paying for English classes in the private sector or even abroad. Hence, the need for investments in English language education by institutions or by the government is paramount. In an under-resourced context like Brazil, English language teaching should prepare teachers and students for language competence at the post-secondary level. Teachers should be aware of the roles of professional development, especially in preparing their language competence for delivering content-area knowledge in English,

particularly improving their communicative skills. If Brazilian HEIs aim at a greater internationalization environment, they must understand that a broader adoption of the English language is the first step to be taken as it allows for the inclusion of different stakeholders in the international educational and scientific contexts.

Acknowledgements

Simone Sarmento holds a CNPq research productivity scholarship level 1D.

References

Airey, J. (2011). Talking about teaching in English: Swedish university lecturers' experiences of changing teaching language. *Ibérica, 22,* 35-54.

Airey, J. (2016). EAP, EMI or CLIL? In: K. Hyland, Ken & P. Shaw (Eds.), *Routledge handbook of English for academic purposes*. (pp 71-83). Routledge.

Airey, J. et al. (2017). The expansion of English-medium instruction in the Nordic countries: Can top-down university language policies encourage bottom-up disciplinary literacy goals? *Higher Education, [s.l.], 73*(4), 561–576.

Ammon, U. (2010). The hegemony of English. In: UNESCO. *World Social Science Report: Knowledge Divides*, 154-156.

Baumvol, L. K. & Sarmento, S. (2016). Internationalization at Home and the use of English as a Medium of Instruction. In Beck, S. et al. (Eds.), *ECHOES: Further reflections on language and literature* (pp. 65-82). EdUFSC: Florianopolis.

Baumvol, L. K. (2018). *Language practices for knowledge production and dissemination: The case of Brazil.* (Doctoral dissertation, Universidade Federal do Rio Grande do Sul, Brazil). Retrieved from https://lume.ufrgs.br/handle/10183/189174

Baumvol, L. K. & Sarmento, S. (2019). Can the use of English as a Medium of Instruction promote a more inclusive and equitable higher education in Brazil?. *Simon Fraser University Educational Review. Burnaby, BC, Canada. Vol. 12, n. 2 (Summer 2019), p.[87]-105*.

Baumvol, L., Sarmento, S. & Arêas da Luz Fontes, A. B. (2021). Scholarly publication of Brazilian researchers across disciplinary communities. *Journal of English for Research Publication Purposes*, *2*(1), 5-29.

Beelen, J. & Jones, E. (2015). Redefining internationalisation at home. In A. Curaj, L. Matei, R. Pricopie, J. Salmi & P. Scott (Eds.). *The European higher education area: Between critical reflections and future policies* (pp. 59-72). Springer.

Botha, W. (2014). English in China's universities today. *English Today 117, 30* (1), 3-10. DOI: doi:10.1017/S0266078413000497

Briggs, J. G. & Dearden, J. (2018). English medium instruction: Comparing teacher beliefs in secondary and tertiary education. *Studies in Second Language Learning and Teaching, 8*(3), 673-696. DOI: http://dx.doi.org/10.14746/ssllt.2018.8.3.7

Chapple, J. (2015). Teaching in English Is Not Necessarily the Teaching of English. *International Education Studies, Ontario, 8*(3). DOI: 10.5539/ies.v8n3p1.

Clark, H. H. (1996). The use of language. In H. H. Clark, *Using language* (pp. 3-25). Cambridge: Cambridge University Press.

Coleman, J. A. (2006). English-medium teaching in European higher education. *Language Teaching, 39*(1). DOI: 10.1017/S026144480600320X.

Collins, A. B. (2010). English-medium higher education: Dilemma and problems. *Eurasian Journal of Educational Research, 39*, 97–110. DOI: http://hdl.handle.net/11693/48964

Crystal, D. (2003). *English as a Global Language* (2nd edition). Cambridge: Cambridge University Press.

Dearden, J. (2014). English as a medium of instruction – a growing global phenomenon: Phase 1. *British Council*.

De Swaan, A. (2001). *Words of the World*. Global Language System. Cambridge: Polity Press and Blackwell.

de Wit, H., Hunter, F., Egron-Polak, E. & Howard, L. (Eds (2015). *Internationalisation of higher education: A study for the European parliament*. http://www.europarl.europa.eu/RegData/etudes/STUD/2015/540370/IPOL_STU(2015)540370_EN.pdf

Gimenez, T., Sarmento, S., Archanjo, R., Zicman, R. & Finardi, K. (2018). Guide to English as a Medium of Instruction in Brazilian Higher Education Institutions 2018-2019. FAUBAI. DOI: 10.13140/RG.2.2.31454.89921.

He, J. & Chiang, S. (2016). Challenges to English-medium instruction (EMI) for international students in China: A learners' perspective. *English Today 128, 32*(4). DOI: 10.1017/S0266078416000390

Henriksen, B., Holmen, A. & Kling, J. (2018). *English Medium Instruction in Multilingual and Multicultural Universities: Academics' Voices from the Northern European Context* (1st ed.). Routledge. https://doi.org/10.4324/9780429456077

Hu, G. & Lei, J. (2014). English-medium instruction in Chinese higher education: a case study. *Higher Education*, *67*(5), 551–567. DOI: 10.1007/s10734-013-9661-5.

Hyland, K. (2015). *Academic Publishing: Issues and challenges in the construction of knowledge*. Oxford University Press.

Jenkins, J. (2013). *English as a lingua franca in the international university: The politics of academic English language policy*. Routledge.

Jordão, C. M. & Martinez, J. Z. (2021). Wines, Bottles, Crises: A Decolonial Perspective on Brazilian Higher Education. *Revista Brasileira de Linguística Aplicada*, 21, 577-604.

Josephson, O. (2005). Parallelspråkighet [Parallel language use.] *Språkvård*, 3.

Knight, J. (2008). *Higher Education in Turmoil: The changing world of internationalization*. Rotterdam: Sense Publishers.

Lillis, T. M. & Curry, M. J. (2010). *Academic Writing in a Global Context: The politics and practices of publishing in English*. Routledge.

Liu, J. Fang, F. G. (2017). Perceptions, awareness and perceived effects of home culture on intercultural communication: Perspectives of university students in China. *System*, *67*, 25-37.

Macaro, E. (2015). English medium instruction: Time to start asking some difficult questions. *Modern English Teacher*, Shoreham by Sea, *24*(2), 4-7.

Macaro, E. (2018). *English medium instruction*. Oxford University Press.

Martinez, R. (2016). English as a Medium of Instruction (EMI) in Brazilian higher education: Challenges and opportunities. In K. Finardi (Ed.) *English in Brazil: Views, policies and programs*. Eduel.

Martínez, F. Z. & Chichón, J. L. E. (2020). EMI at Tertiary Level in Spain: Perspectives From Lecturers at a Medium-Sized State University. In M. M. Sánchez-Pérez (Ed.), *Teacher Training for English-Medium Instruction in Higher Education* (pp. 232-256). IGI Global. DOI: 10.4018/978-1-7998-2318-6

Montgomery, S. L. (2013). *Does science need a global language? English and the future of research.* University of Chicago Press.

Muñoz, C. (2012). *Intensive exposure experiences in second language learning*. UK: Multilingual Matters.

Orduna-Nocito, E. & Sánchez-Garcia, D. (2022). Aligning higher education language policies with lecturers' views on EMI practices: A comparative study of ten European Universities. *System, 104*, 2-14.

Pecorari, D. & Malmström, H. (2018). At the crossroads of TESOL and English medium instruction. *TESOL Quarterly*, 52(3), 497–515. https://doi.org/10.1002/tesq.470

Richards, J. C. & Pun, J. (2022). Teacher strategies in implementing English medium instruction. *ELT Journal*, 76(2), 227-237.

Solovova, O., Santos, J. V. & Verissimo, J. (2018). Publish in English or Perish in Portuguese: Struggles and Constraints on the Semiperiphery. *Publications, 6* (25), 1-14.

Tatzl, D. (2011). English-medium masters' programmes at an Austrian university of applied sciences: Attitudes, experiences and challenges. *Journal of English for Academic Purposes, 10*, 252–270. https://doi.org/10.1016/j.jeap.2011.08.003

Teekens, H. (2007). Internationalisation at home: An introduction. In H. Teekens (Org.). *Occasional paper 20: Internationalisation at home: Ideas and ideals (pp. 3-12)*, Amsterdam: Drukkerij Raddraaier.

Tran., T. H. T., Burke, R. & O'Toole, J. M. (2021). Perceived Impact of EMI on Students Language Proficiency in Vietnamese Tertiary EFL Context. *Journal of Education: Language LEarning in Education, 9*(3), 7-24.

Tsuchiya, K. & Pérez Murillo, M. D. (2019). Prospective Teachers' Perceptions of CLIL in Spain and Japan: Translingual Social Formation through EMI-CLIL Lectures. In: Tsuchiya, K., Pérez Murillo, M. (eds) *Content and Language Integrated Learning in Spanish and Japanese Contexts.* Palgrave Macmillan, Cham. DOI: https://doi-org.proxy.lib.sfu.ca/10.1007/978-3-030-27443-6_15

Wächter, B. & Maiworm, F. (2014). *English-taught programmes in European higher education: The state of play in 2014.* Lemmens.

Walkinshaw, I., Fenton-Smith, B. & Humphreys, P. (2017). EMI issues and challenges in Asia-Pacific higher education: An introduction. In B. Fenton-Smith, P. Humphreys, & I. Walkinshaw (Eds.), *English medium instruction in higher education in Asia-Pacific: from policy to pedagogy* (pp. 1-18). (Multilingual education; Vol. 21). Springer, Springer Nature. https://doi.org/10.1007/978-3-319-51976-0_1

Werther, C., Denver, L., Jensen, C. & Mees, I. M. (2014). Using English as a medium of instruction at university level in Denmark: the lecturer's perspective. *Journal of Multilingual and Multicultural Development*, *35*(5), 443-462.

Windle, J. & Nogueira, M. A. (2015). The role of internationalisation in the schooling of Brazilian elites: distinctions between two class fractions. *British Journal of Sociology of Education*, *36*(1), 174-192.

Yeh, C.-C. (2014). Taiwanese students' experiences and attitudes towards English-medium courses in tertiary education. *RELC Journal, 45*(3), 305-319. https://doi.org/10.1177/0033688214555358

About the authors

Alfredo Afonso Ferreira
Dr. Alfredo Ferreira is a Lecturer in the science stream of the Academic English Program at Vantage College in the University of British Columbia, Canada, where he teaches first-year courses in academic writing and science language and literacy. Alfredo's teaching and research are informed by systemic functional linguistics and Vygotskian sociocultural theory. His doctorate from UBC examines the development of grammatical metaphor in the writing of multilingual graduate students. Alfredo's projects include the development of a free, online, language and literacy-focused textbook for first-year physics and professional development programming in research writing with the NGO Academics Without Borders.
E-mail contact: alfredo.ferreira@ubc.ca

Ana Eliza Pereira Bocorny
Ana Eliza Pereira Bocorny is a Professor at the Department of Modern Languages of the Federal University of Rio Grande do Sul. She holds a Ph.D. in Language Studies, (UFRGS), and an MA in Education (PUCRS). Her research interests include Applied Linguistics, English for Academic Purposes, and Corpus Linguistics.
E-mail contact: ana.bocorny@ufrgs.br

Ana Luiza Pires de Freitas
Ana Luiza Pires de Freitas is an English language educator at the Department of Education and Humanities at the Federal University of Health Sciences of Porto Alegre/Brazil (UFCSPA). She holds a PhD in Language Studies and her main research interests are in academic writing in health sciences and English as a medium of education. She's also active in international teacher education programs.
E-mail contact: analuizaf@ufcspa.edu.br

Deise Amaral
Deise Amaral is a researcher on language assessment, with a focus on large-scale proficiency exams. Deise has worked as a language teacher and examiner for over 30 years. She holds an MA and is currently pursuing a PhD in Applied Linguistics at UFRGS. Her research interests are in language testing, writing development, vocabulary and phraseology studies, and the use of corpus methods to inform the description of proficiency levels.
E-mail contact:: deise.amaral@ufrgs.br

Deise Prina Dutra
Deise P. Dutra is a Full Professor at the Federal University of Minas Gerais (UFMG) in Brazil where she teaches English and Linguistics at the undergraduate and graduate levels. Her research interests are corpus linguistics, mainly with a focus on learner and specialized Corpora, as well as on data-driven learning. She was the coordinator of Language without Borders and of English for academic purposes subjects and the Deputy Dean for International Affairs at UFMG.
E-mail contact: deisepdutra@gmail.com

Diva Cardoso de Camargo
Diva Cardoso de Camargo is a Professor of Translation Studies at the São Paulo State University, Brazil. Her Pos-Doctoral studies was on translation at The University of Manchester. She lectures on translation theory and literary translation, and supervises a number of research students in the area of translation studies and literature. Her main area of research interest is the use of corpora as a resource for studying various features of translation, including the distinctive nature of translated text and the distinctive styles of individual translators.
E-mail contact: divaccamargo@gmail.com

Greta Perris
Greta Perris is a PHD student at UBC, Faculty of Education. As a Graduate Research Assistant, she has been engaged in research projects examining teaching multilingual students in EAP programs and across content

courses in post-secondary institutions. She has an MA in Second Language Education from OISE, University of Toronto and has been researching, managing, developing curriculum, and teaching in English as an additional language programs in Canada and abroad for over 15 years.
E-mail contact: greta.perris@ubc.ca

Jennifer Klein
Jennifer Klein is an instructor of English to adult learners at Coconino Community College. She is interested in technical writing and data analytics. Her background is in linguistics, education, and writing. She holds an MA in TESL Language Teaching/Applied Linguistics from Northern Arizona University.
E-mail contact: jlg642@nau.edu

Larissa Goulart
Larissa Goulart is an Assistant Professor of Linguistics at Montclair State university. She received her PhD in Applied Linguistics from Northern Arizona University in 2022. Her research focuses on register variation, corpus linguistics for language teaching, and English for University Purposes. Her research has been published in major peer-reviewed journals including Register Studies, Journal of English for Academic Purposes, and Corpora.
E-mail contact: goulartl@montclair.edu

Laura Baumvol
Dr. Laura Baumvol is a lecturer at the University of British Columbia School of Journalism, Writing, and Media. She holds a PhD in Applied Linguistics and was a recipient of the Emerging Leaders of America Scholarship (Global Affairs Canada). She has published papers and edited journals internationally and is an investigator in multiple research projects. Her main interests are the use of languages for knowledge production and dissemination, writing across the disciplines, research-informed practices in teaching and learning, the relationship between scholarly and public discourse, and internationalization of higher education.
E-mail contact: laura.baumvol@ubc.ca

Lucas Marengo
Lucas Marengo is a Ph.D. student and researcher at the Federal University of Rio Grande do Sul. He holds a Master's degree in Language Studies at the same university. His main topics of research are focused on the internationalization processes of Brazilian higher education, English as a Medium of Instruction (EMI), and the offer of additional languages in academic settings.
E-mail contact: lucashenriquefog@gmail.com

Luciano Franco
Luciano Franco is a professor at the Federal Institute of Education, Science and Technology of Paraná (IFPR). He holds a Master´s degree in Linguistic Studies at the São Paulo State University "Júlio de Mesquita Filho" (Unesp), and is currently a doctoral student at the same institution. He is a member of the Research Group En-Corpora: Corpus-Based and Corpus-Driven Teaching Research Group) at Unesp. His research areas are: English Language Teaching, Corpus Linguistics and English for Specific and Academic Purposes.
E-mail contact: luciano.francco@gmail.com

Maria Kostromitina
Maria Kostromitina is a Ph.D. candidate in Applied Linguistics at Northern Arizona University. Her research interests lie at the intersection of second language prosody and pragmatics. She has also been involved in research projects in the domains of speech perception, language assessment, and corpus linguistics. Maria's research has been published in major peer–reviewed journals in applied linguistics including Studies in Second Language Acquisition.
E-mail contact: masha@nau.edu

Marine Laísa Matte
Marine Laísa Matte is a researcher on English for Academic Purposes. She is currently a PhD candidate at the Federal University of Rio Grande do Sul, where she also received her Master's degree in Applied Linguistics in

2019. Marine is a Portuguese and English teacher at Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSul). Her research interests are in language teaching, corpus linguistics and academic writing. E-mail contact: marinematte@ifsul.edu.br

Paula Tavares Pinto
Paula Tavares Pinto is a Lecturer at São Paulo State University (Unesp, Brazil). She developed part of her Ph.D research at the University of Manchester and was a visiting scholar at the University of Surrey. She is the leader of the Research Group En-Corpora: Corpus-Based and Corpus-Driven Teaching and has supervised a number of research students. Her publications lie in the areas of Translation Studies, Terminology, Corpus Linguistics, Data-Driven Learning and English for Specific and Academic Purposes. E-mail contact: paula.pinto@unesp.br

Rozane Rebechi
Rozane Rebechi is a professor and researcher at the Federal University of Rio Grande do Sul (UFRGS). She holds a Master and a Ph.D. degree in English Language and Literature from the University of São Paulo (Brazil). Her main areas of research are Translation, Terminology, and Discourse, to which she applies Corpus Linguistics as methodology. She was chair of the Brazilian Association of Researchers in Translation (ABRAPT) through 2020-2022 and is Associated Partner and member of the management board to the European Masters in Technology for Translation and Interpreting (EM TTI).
E-mail contact: rozane.rebechi@ufrgs.br

Sandra Zappa-Hollman
Sandra Zappa-Hollman is an Associate Professor in the Department of Language and Literacy Education at the University of British Columbia and Director of Academic English Programming at Vantage College. Her research has mainly examined processes and outcomes concerning academic discourse socialization of multilingual learners in English-speaking institutions; the application of genre-based pedagogies for language and con-

tent integration in EAP curricula; interdisciplinary collaborations between EAP and other subject-area instructors; language ideologies underlying university instructor's beliefs about working with multilingual learners; and the use of paid academic support services by post-secondary students. For more details, please visit https://lled.educ.ubc.ca/sanda-zappa-hollman/
E-mail contact: Sandra.zappa@ubc.ca

Simone Sarmento
Simone Sarmento is a professor at the Federal University of Rio Grande do Sul (UFRGS). She holds a doctorate in Terminology and Lexicography from UFRGS (2008), a Masters in Language Studies from Lancaster University (2005), and a Masters in Applied Linguistics from UFRGS (2001). She was a visiting scholar at the Faculty of Education at the University of British Columbia (CAPES) and at the University of Lisbon (Santander Universities). Her main research interests are in the fields of Corpus Linguistics, English as a Medium of Instruction, English for Academic and Special Purposes, Language Education Policies, and Teacher Education. She holds a CNPq level 1D research productivity grant.
E-mail contact: simone.sarmento@ufrgs.br

Talita Serpa
Talita Serpa holds a PhD in Linguistic Studies from the Postgraduate Program in Linguistic Studies at São Paulo State University "Júlio de Mesquita Filho" (IBILCE/Unesp) with research on Corpus-Based Translation Pedagogy and Translation Studies (2017). She also was an Academic Visitor at The University of Manchester (UK) (2015-2016) financed by FAPESP. Nowadays, she is Postdoctoral Researcher and Collaborating professor at the same Postgraduate Program in Linguistic Studies at Unesp, with CAPES/PNPD support (2019-2023). Most of her works since 2017 has been on Translation and Corpora, Corpora and Terminology and DDL.
E-mail contact: talita.serpa@unesp.br

Tony Berber Sardinha

Tony Berber Sardinha is Professor of Applied Linguistics at the Pontifical Catholic University of Sao Paulo (PUCSP), Brazil. He has published numerous books and research articles. In addition to being on the board of major international journals and book series, he is the editor-in-chief for the linguistics journal DELTA (https://www.scielo.br/j/delta), and the current corpus linguistics area editor for the multi-volume Encyclopedia of Applied Linguistics (2nd ed., ed. Carol Chapelle). His interests include corpus linguistics, multidimensional analysis, multimodality, social media, discourse analysis, and metaphor.

E-mail contact: tonycorpuslg@gmail.com