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 multi-dimensional 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.

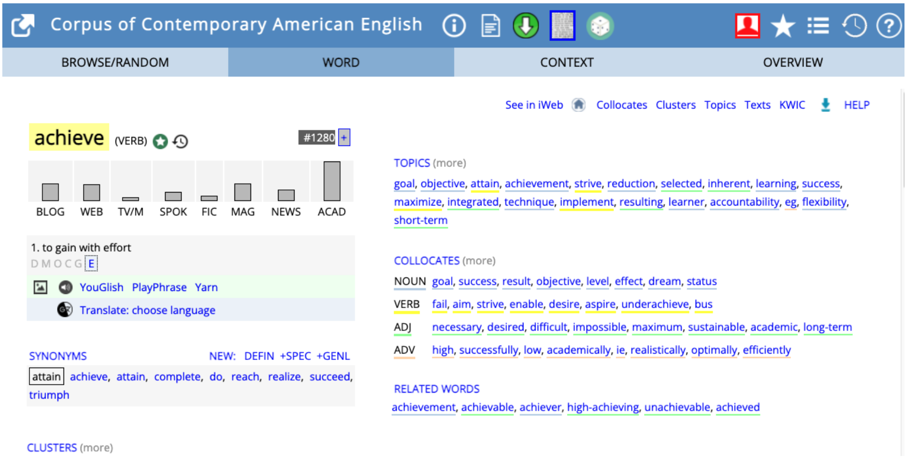
heart and periphery, nourishing all." Therefore, phraseology should be vital to language teaching in general and to EAP in particular.

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.



1. to gain with effort

attain achieve, attain, complete, do, reach, realize, succeed, triumph

TOPICS (more)

goal, objective, attain, achievement, strive, reduction, selected, inherent, learning, success, maximize, integrated, technique, implement, resulting, learner, accountability, eg, flexibility, short-term

COLLOCATES (more)

NOUN	goal, success, result, objective, level, effect, dream, status
VERB	fail, aim, strive, enable, desire, aspire, underachieve, bus
ADJ	necessary, desired, difficult, impossible, maximum, sustainable, academic, long-term
ADV	high, successfully, low, academically, ie, realistically, optimally, efficiently

RELATED WORDS

achievement, achievable, achiever, high-achieving, unachievable, achieved

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.

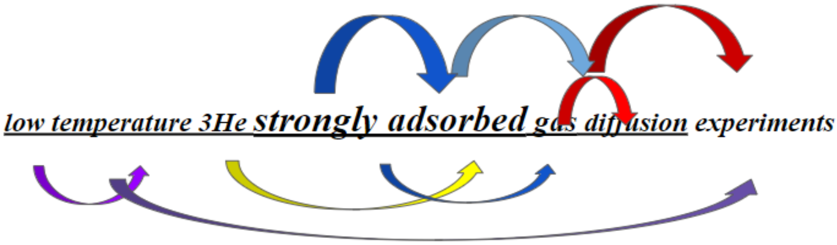

low temperature 3He strongly adsorbed gas diffusion experiments

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*.

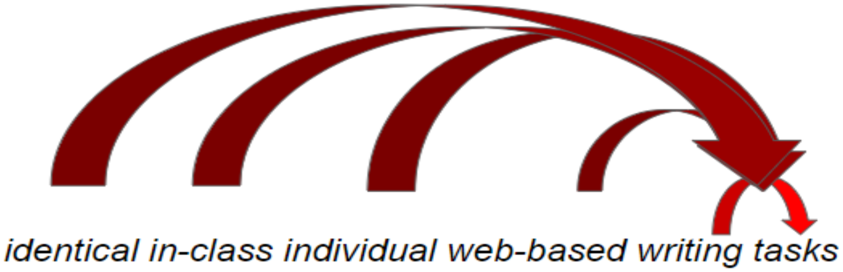


Figure 3. Sample of interrelations of modifiers from an applied linguistics corpus

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: 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., 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.