ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

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

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

SIMONE SARMENTO
ROZANE REBECHI
MARINE LAÍSA MATTE
(ORG.)

Porto Alegre • 2024 • 1ª edição

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

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

Introduction

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

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

---

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

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

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

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

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

---

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

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

Combining corpus and genre-based approaches

Corpus Linguistics

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

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

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

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

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

1. Authentic Linguistic Data;

2. Readable Computer Segments;

3. Specially Organized Language Portions;

4. Texts Capable of Representing a Particular Language or Variety of Language.

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

Corpus-Based Pedagogy

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

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

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

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

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

Data Driven Learning (DDL)

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

Pedagogical Models Combining Corpus and Genre-Based Approaches

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

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

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

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

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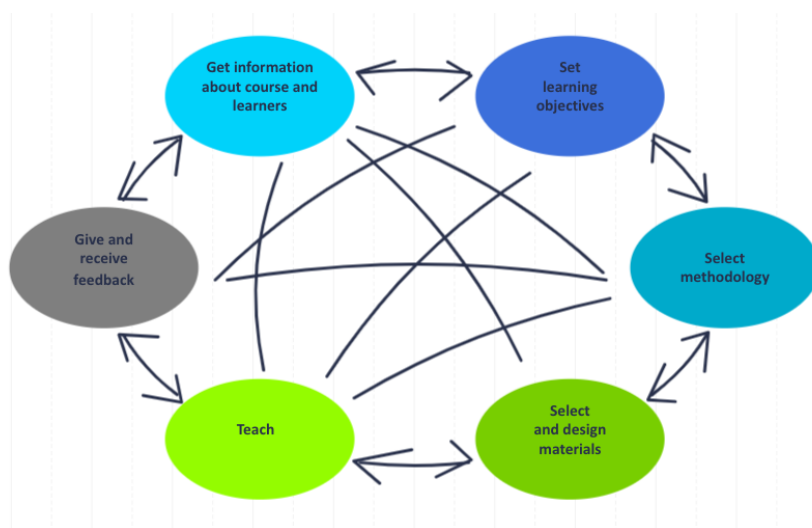

Figure 1. Stages involved in the process of designing and implementing an EAP course

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

The design of EAP materials

Framework

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

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

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

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

---

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

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

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

Table 1. Pedagogical unit structure for an EAP writing course

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

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

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

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

Step-by-step guide

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

Description of the EAP writing course

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

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

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

Table 2. Description of the EAP writing course

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

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

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

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

Table 3. Learning objectives for course and classes

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

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

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

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

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

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

ABSTRACT

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

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

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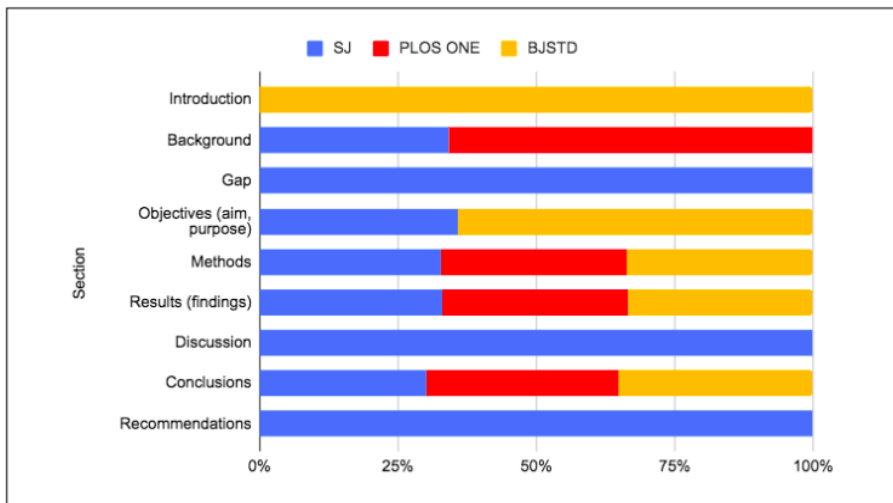

Figure 4. Rhetorical structure of Epidemiology structured abstracts. From Freitas and Bocorny (2021: 3)

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

PROCEDURE 1:

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

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

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

Figure 5. Section heading of the structured abstracts being studied

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

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

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

---

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

8   http://micusp.elicorpora.info/

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

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

---

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

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

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

Type	Rank	Freq	Range	S1_TT	S1_Ent	S2_TT	S2_Ent	S3_TT	S3_Ent
1 i + i i + i	1	1596	530	0.155	0.682				
2 p + p the + of	2	1385	1299	0.021	0.65				
3 the + of + study was	3	1379	1377	0.011	0.399			0.003	0.4
4 this study + to + the	4	1251	1250			0.01	0.518		
5 the + of this + was	5	1234	1231	0.015	0.394				
6 the + + this study was	6	1171	1169	0.014	0.403	0.008	0.03		
7 p results p + + of	7	1002	1002					0.11	0.403
8 aim of + study + to	8	982	982			0.005	0.371		
9 aim of + + was to	9	897	897			0.004	0.436	0.011	0.123
10 aim of this + + to	10	863	863					0.013	0.148
11 the aim of + + was	11	859	859					0.005	0.431
12 aim + + study was to	12	857	857	0.004	0.016	0.005	0.442		
13 aim + this study + to	13	823	823	0.005	0.066				
14 the aim + + study was	14	822	822			0.004	0.017	0.005	0.436
15 p + p a + of	15	790	774	0.023	0.37				
16 ...	16	771	754	0.021	0.575				

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

PROCEDURE 2:

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

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

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

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

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

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

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

	Lemma	Frequency Per Million ? ↓	DOCF ?			Lemma	Frequency Per Million ? ↓	DOCF ?	
1	patient	11,678.11	3,043	•••	251	curve	198.90	201	•••
2	risk	4,591.83	1,938	•••	252	prescription	198.16	123	•••
3	study	4,584.41	3,014	•••	253	classification	196.68	184	•••
4	p	4,560.66	1,789	•••	254	region	196.68	176	•••
5	cancer	4,359.53	1,395	•••	255	category	196.68	147	•••
6	ci	3,440.72	1,431	•••	256	program	195.19	157	•••
7	result	3,026.59	3,245	•••	257	endpoint	195.19	163	•••
8	group	2,885.57	1,296	•••	258	efficacy	193.71	199	•••
9	disease	2,683.70	1,529	•••	259	growth	193.71	141	•••
10	method	2,645.11	3,118	•••	260	community	192.97	156	•••
11	year	2,550.85	1,717	•••	261	experience	192.22	154	•••
12	conclusion	2,432.10	3,234	•••	262	hemorrhage	190.74	124	•••
13	analysis	2,277.73	1,927	•••	263	prediction	190.00	158	•••
14	factor	2,270.31	1,489	•••	264	adherence	189.25	90	•••
15	treatment	2,259.18	1,296	•••	265	safety	188.51	184	•••

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

Using Sketch Engine and searching for concordance lines with the lemma 'study' as a noun, it is possible to retrieve language data that could be easily integrated into exercises to be used in the course **Written Production of Structured Abstracts in the Area of Health Sciences**. Figure 9 shows the results:
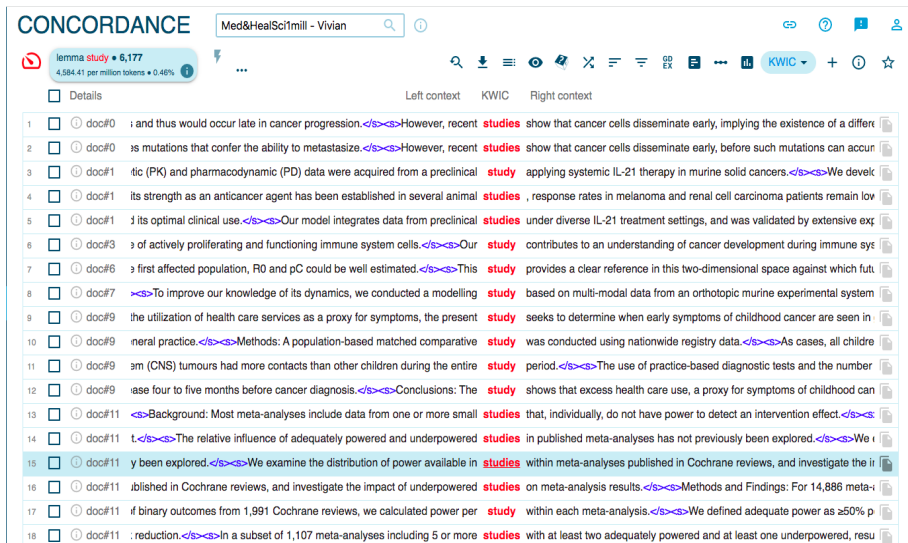
Figure 9. Concordance lines with the lemma 'study' as a noun. From Kilgarriff et al. (2004)

PROCEDURE 3:

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

Figure 10 illustrates the search for 'study':
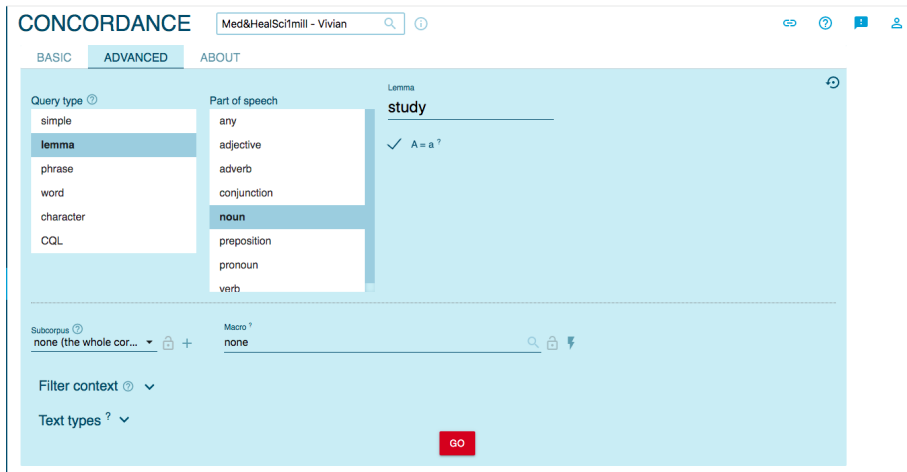
Figure 10. SE interface for PROCEDURE 3. From Kilgarriff et al. (2004)

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

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

PROCEDURE 4 presents the steps for filtering data:

1) Use the results obtained with PROCEDURE 3 (search for the lemma 'study', as a noun)
2) Click on the Filter icon, as shown in Figure 12:
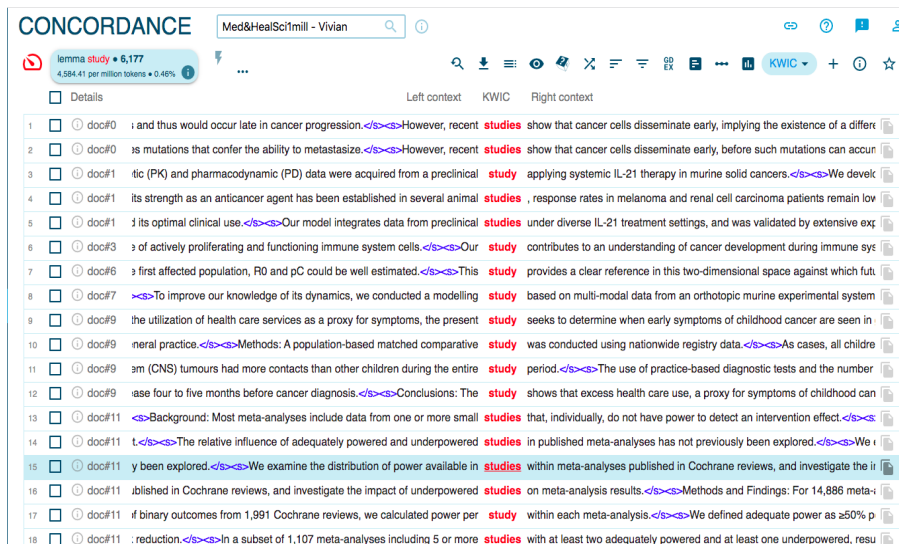
Figure 12. Filtering data in SE. From Kilgarriff et al. (2004)

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

Figure 13 illustrates the search:
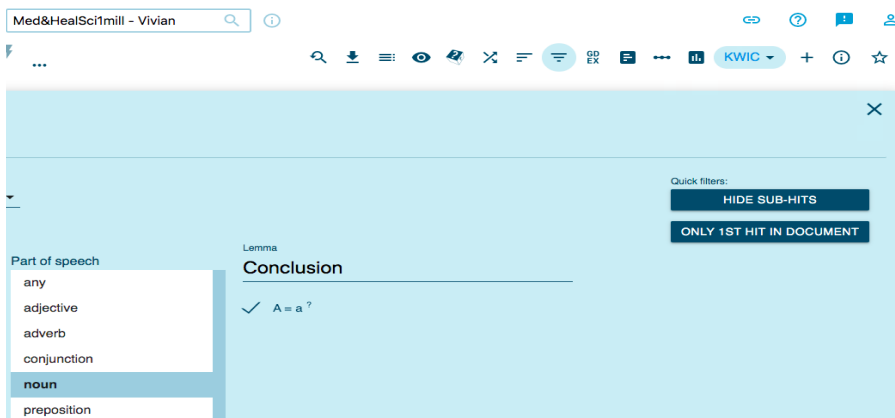
Figure 13. SE interface for PROCEDURE 4. From Kilgarriff et al. (2004)

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

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

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

the structured abstracts that have 'study' as a collocation node. In this case, the collocation node 'study' can be replaced by any of the collocation nodes identified in the wordlists extracted from the corpus. Figure 15 shows the results when using this CQL:
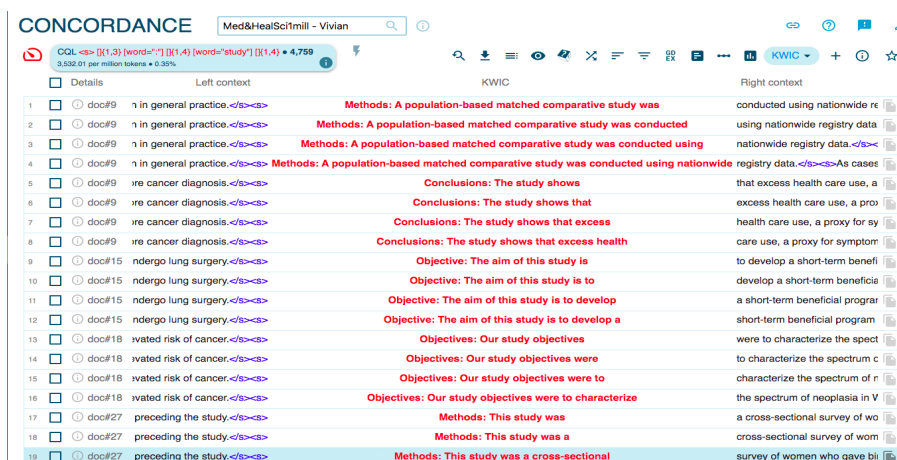
Figure 15. Results for the CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] [] {1,4}. From Kilgarriff et al. (2004)

PROCEDURE 5:
1) Open Sketch Engine
2) Select the corpus you want to work with
3) Go to Concordance
4) Select Advanced
5) Click on CQL, in Query type
6) Paste the CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4} under CQL
7) Press GO
8) Click on KWIC (to organize the results alphabetically)
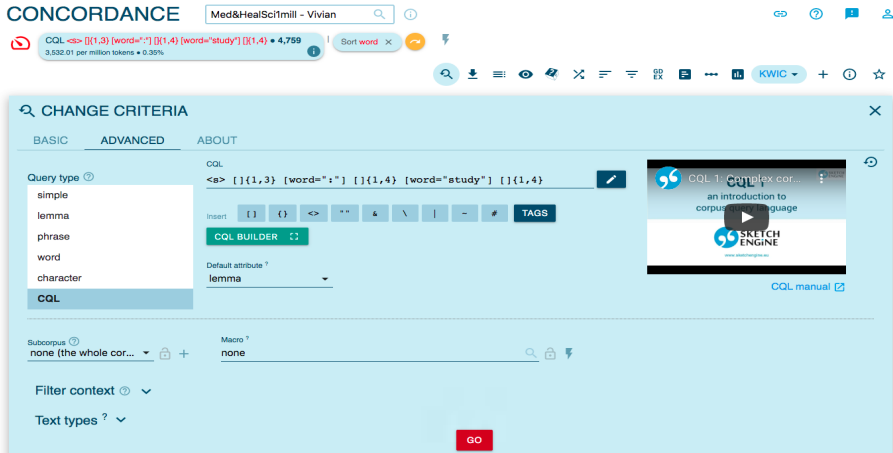
Figure 16. SE interface for extracting LBs from sections of the structured abstracts using CQL <s> []{1,3} [word=":"] []{1,4} [word="study"] []{1,4}. From Kilgarriff et al. (2004)

The results in Figure 16 indicate that collocations with 'study' occur across sections of these structured abstracts. These results can also be filtered for each section identified as part of the rhetorical structure of the abstracts under study. For example, as shown in Figure 17, collocations with the word 'study' occur 440 times in the section CONCLUSION in the corpus of Health Sciences:
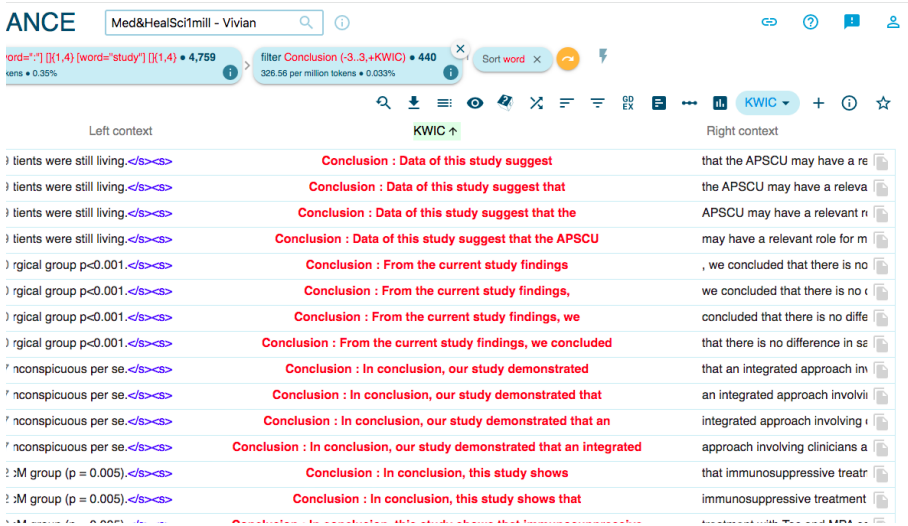
Figure 17. Collocations with the word 'study' filtered for the section CONCLUSION. From Kilgarriff et al. (2004)

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

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

Table 6. LF with the node 'study'

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

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

Concluding remarks

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

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

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

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

Acknowledgement

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for academic purposes*, *6*(4), 289-302.

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

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

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

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

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes*, *12*(1), 33-43.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kavanagh, B. (2021). Bridging the Gap from the Other Side: How Corpora Are Used by English Teachers in Norwegian Schools. *Nordic Journal of English Studies*, *20*(1), pp.1–35. DOI: http://doi.org/10.35360/njes.522

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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