Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1229-1252, 2021

Analysing the behaviour of academic collocations in a corpus of research-papers: a data-driven study

Analisando o comportamento de colocações acadêmicas em um corpus de artigos científicos: um estudo dirigido por dados

Paula Tavares Pinto
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
paula.pinto@unesp.br
http://orcid.org/0000-0001-9783-2724

Diva Cardoso de Camargo
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
divaccamargo@gmail.com
http://orcid.org/0000-0001-6924-4757

Talita Serpa
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
talita.serpa@unesp.br
https://orcid.org/0000-0003-3324-9593

Luciano Franco da Silva
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
luciano.francco@gmail.com
https://orcid.org/0000-0001-7485-8657

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1229-1252

Abstract: Authors from different countries have published their papers in English, aiming to promote their research results widely and to become internationally known by their peers. It is also true that, although they are aware of the English terminology used in their respective field, some authors still struggle with some features of academic writing such as collocations. Thus, this paper presents a discussion on the underuse and overuse traces of academic collocations by Brazilian authors who had their articles published in English on an open electronic library of scientific journals. In order to analyse the collocations used by these researchers, we compiled a 906,035-word corpus from eight different academic areas. The collocations observed were statistically compared to those from an academic corpus of English writings which contains texts produced by English-speaking authors. Results showed that there are more collocations underused than overused by the authors. The analysis proved that the collocation repertoire of researchers could be broadened by being pointed out during academic writing workshops.

Keywords: academic collocations; research paper writing; corpus linguistics.

Resumo: Autores de vários países têm publicado seus artigos científicos em inglês com o intuito de promover amplamente os resultados de suas pesquisas dentre a comunidade científica internacional. É verdade que, embora estejam cientes da terminologia utilizada no respectivo campo de pesquisa, alguns autores ainda apresentam dificuldade em lidar com certas características da escrita acadêmica, como o uso das colocações. Este artigo apresenta uma discussão sobre traços de sobreuso e subuso de colocações acadêmicas utilizadas por autores brasileiros que têm seus artigos publicados em inglês numa plataforma eletrônica aberta de artigos científicos. Para analisar as colocações utilizadas por estes pesquisadores, compilamos um corpus de 906.000 palavras a partir de oito áreas científicas. As colocações analisadas foram comparadas estatisticamente com as colocações de um corpus acadêmico de inglês que contém textos escritos por autores anglófonos. Os resultados mostraram que há mais traços de subuso que sobreuso de colocações acadêmicas utilizadas pelos pesquisadores e este repertório poderia ser ampliado se fossem destacadas durante cursos de escrita acadêmica em língua inglesa.

Palavras-chave: colocações acadêmicas; escrita de artigos científicos; linguística de corpus.

Submitted on October 9th, 2020
Accepted on December 16th, 2020

1 Introduction

Authors worldwide recognise the importance of publishing academic articles in English. Although there may be some debate over the relevance of publishing in one's native language, researchers must publish in English if they want their study results to be read by members of international scientific communities. In that sense, Brazilian authors, who wish to have their studies internationally acknowledged, need to have their articles publicised on online databases, such as *The Scientific Electronic Library Online* (*SciELO*). This platform is an electronic library for Brazilian scientific journals written in Portuguese, Spanish and English.

Taking that into account, several studies (HYLAND, 2008; NESSELHAUF, 2003; PAQUOT, 2010) have already highlighted the fact that non-native speakers may lack the necessary linguistic knowledge to use adequate academic collocations when writing in English. Haswell (1991) has claimed that the underuse of collocations in scientific papers will reveal one's "apprentice writing" which can compromise the acceptance of papers by scientific journals. On the other hand, the proper use of academic collocations would demonstrate how linguistically competent the authors are.

The definition of collocation by the *Oxford Collocations Dictionary for students of English* (LEA; CROWTHER; DIGNEN, 2002, p. vii) is the following: "collocation is the way words combine in a language to produce natural-sounding speech and writing". As examples, the authors state that, in English, it is common to say *strong wind* and *heavy rain*, but not **heavy wind* or **strong rain*.

According to Frankenberg-Garcia *et al.* (2019a), some writers are not aware of collocations or do not use them, which may lead to readers' estrangement caused by combinations such as **depend of* something, instead of **depend on* something. For this reason, the researchers developed the Collocaid Project. The main objective of this tool is to create "a lexicographic resource that is accessed from within digital writing environments to help learners write more idiomatically" (FRANKENBERG-GARCIA *et al.*, 2019a, p. 24).

Another topic to be addressed is whether academic collocations stand out to non-native authors as terms and idioms do. According to Nesselhauf (2003), English collocations can be fuzzy for students, academic authors and even native speakers who are not familiar with some commonly patterned combinations. The *Oxford Collocations Dictionary for students of English* states that "collocation runs through the whole of the English language. No piece of natural spoken or written

English is free of collocation" (LEA; CROWTHER; DIGNEN, 2002, p. vii). If collocations in general English are already challenging to be noticed by non-native speakers, we wonder how it would be with academic collocations such as 'rates fell', 'the percentage dropped', 'gather information', 'funding research', among others. Consequently, we question if international researchers, who are non-native speakers of English, can proficiently combine words to produce natural collocations and, more specifically, we want to know how it happens among Brazilian researchers.

Despite the relevance of academic vocabulary and collocations in scientific texts, there are still few studies (DAYRELL 2007; PAIVA, 2009; SILVA *et al*., 2017; SILVA *et al*., 2018) that report their use in the writing of Brazilian authors. Dayrell compared collocational patterns in translated and non-translated texts. The author shows that translations from Portuguese into English draw on a small number of collocates (DAYRELL, 2007, p. 377). Paiva (2009) found evidence of overuse of specific verbs in research papers translated by Brazilian professional translators which are not frequent in articles published in high-impact journals. Babini and Silva (2012) showed that Brazilian researchers produce texts with overuse or underuse of specific lexical items which are generally expected in research papers in English. Silva *et al*. (2018) investigated the use of academic vocabulary by Brazilian (under)graduate students. They concluded that although students use a similar number of academic words compared to the Academic Word List (AWL) and the General Service List (GSL), the word forms chosen by students differ as they underuse affixation processes.

Although the four previous studies refer to the academic vocabulary produced by Brazilians, there are still several issues to be dealt with, such as the use of academic collocations by senior Brazilian researchers who have longer published papers in English. Do they tend to overuse or underuse collocations in their research papers? Are those collocations repeated over the article? These are some of the issues to be discussed in this article.

Therefore, this study seeks to shed some light on the way senior Brazilian researchers use academic collocations in their publications by presenting an investigation of data extracted from a corpus of papers in the eight major areas of research at *The Scientific Electronic Library Online* (SciELO).

The guiding research questions of this study are the following:

1. To what extent do the collocations used by Brazilian authors differ from the ones in international journals?
2. Do Brazilian authors use collocations influenced by their native language (Portuguese)?
3. Are there traces of overuse or underuse of specific collocations?

To answer those questions, we present a brief review of studies that discuss the importance of academic vocabulary and collocations.

2 Academic collocations

Previous studies have revealed that clusters, lexical bundles and collocations have been investigated in different genres of academic writing such as Master's thesis, Doctorate dissertations and research articles (ACKERMANN, CHEN, 2013; CORTES, 2004; FRANKENBERG-GARCIA *et al.*, 2019a, 2019b; HYLAND, 2008; SILVA *et al.*, 2017). Hyland (2008, p. 42) states that clusters are "words which follow each other more frequently than expected by chance, helping to shape text meanings and contribute to our sense of distinctiveness in a register" such as *a result of* or *it should be noted that* in academic writing. According to the author, mastering the use of these group of words, or "clusters" (SCOTT, 1996) will help non-native writers to overcome linguistic barriers which prevent their papers from reaching other members of the international community. At the same time, Cortes (2004, p. 400) states that "lexical bundles are extended collocations, sequences of three or more words that statistically co-occur in a register. Some examples of these word combinations in academic prose are: *on the other hand*, *in the case of*, *the context of the*, and *it is likely to*."

Firth (1951), in turn, was responsible for making collocations well-known and for the famous quote "you shall judge a word by the company it keeps" (*apud* PARTINGTON, 1998, p. 15). Besides, according to Nation (2001), "the term 'collocation' is used to refer to a group of words that belong together, either because they commonly occur together like *take a chance*, or because the meaning of the group is not apparent from the meaning of the parts, as with *by the way* or *to take someone in*. A significant problem in the study of collocation is determining, in a consistent way, what should be classified as a collocation" (NATION, 2001, p. 317).

Ackermann and Chen (2013) state that another difficulty in dealing with collocations is that they "often contain inflective or positional variations (e.g., *results obtained*, *broader contexts*, *achieving objectives*) which poses the great challenge of how to collate these relevant forms and present them in a uniform and consistent way" (ACKERMANN; CHEN, 2013, p. 236). The authors believe that this challenge can only be overcome by human intervention since there is still no automation method to simplify this process. The researchers mentioned above define collocation as "word combinations which co-occur more frequently than by chance across academic disciplines (hence corpus-driven) and are pedagogically relevant in an EAP[1] context (hence expert-judged)" (ACKERMANN; CHEN, 2013, p. 246). They highlight the importance of compiling a list of academic collocations based on the idea proposed by Nation (2001, p. 189-191). The author stated that academic collocations might "neither be sufficiently frequent in the language as a whole to be learnt implicitly nor part of the technical lexicon which is likely to be explicitly taught as part of subject courses".

Contrary to Nesselhauf's hypothesis (2003), which defines collocation only in its phraseological sense, in this paper we chose to adopt a frequency-based approach, which takes into account co-occurrences of words within a specific span, as Sinclair (1991) did in his work.

As far as teaching collocations is concerned, Nesselhauf (2003) suggests it is a task for teachers to make learners aware of these word combinations. The author adds that teachers should explicitly teach collocations since they do not always stand out to the learners' eyes. The criteria to be followed would be teaching the most frequent and acceptable collocations in the register on focus, in this case, academic collocations (*conduct/do/carry out a study* or *make an analysis*). Comparison to native languages (L1) is also desirable, even by highlighting functional elements such as articles and prepositions. The scholar also suggests that they should give the focus to the verb, which seems to be the cause of most mistakes. Finally, in the Brazilian context, Tagnin (2013) discusses the convention of language and dedicates part of her study to the adjective, noun, verb and adverbial collocations in Portuguese compared to English, Italian, French and Spanish. She also shows their importance in teaching and translation practice.

[1] English for Academic Purpose (EAP).

3 Methodology

The methodology followed in this study was composed of two steps: 1) compilation of the *Brazilian Academic Corpus of English* (BrACE); 2) selection of the most frequent academic collocations used by Brazilian researchers in comparison to frequent academic collocations in native English speakers' writings.

We present these steps in the following sections:

3.1 The Brazilian Academic Corpus of English (BrACE)

In order to identify the most frequent academic collocations used by Brazilian researchers in their writings, we selected papers from SciELO (an open cooperative database of journals originated in Brazil that currently features papers from several countries such as Argentina, Bolivia, Brazil, Chile, among others). This selection aimed to gather information from journals which could represent Brazilian authors' writing in all areas of research designated in Brazil. According to SciELO website:

> The Scientific Electronic Library Online - SciELO is an electronic library covering a selected collection of Brazilian scientific journals. The library is an integral part of a project being developed by FAPESP – *Fundação de Amparo à Pesquisa do Estado de São Paulo*, in partnership with BIREME – the Latin American and Caribbean Center on Health Sciences Information. Since 2002, the project is also funded by CNPq – *Conselho Nacional de Desenvolvimento Científico e Tecnológico.*[2]

The choice of SciELO as the source for our corpus is supported by the works of Neves *et al*. (2016), and Kuhn (2017). These authors also based their studies on the reliability of this procedure, since the selection of papers for SciELO lies on strict criteria and policy, based on "peer-review process, journal usage and impact factor" (KUHN, 2017, p. 194).

The website displays the main areas of domain with respective sub-areas: 1. Agricultural Sciences (*Ciências Agrárias*), with six sub-areas; 2. Biological Sciences (*Ciências Biológicas*), with 14 sub-areas; 3. Health Sciences (*Ciências da Saúde*), with nine sub-areas; 4. Physical

[2] SciELO.org – Scientific Electronic Library Online. Available from: www.scielo.org. Retrieved: May 23, 2018.

and Earth Sciences (*Ciências Exatas e da Terra*), with seven sub-areas; 5. Humanities (*Ciências Humanas*), with ten sub-areas; 6. Applied Social Sciences (*Ciências Sociais Aplicadas*), with 12 sub-areas; 7. Engineering (*Engenharias*), with 12 sub-areas; 8. Languages, Linguistics and Arts (*Linguística*, *Letras* e *Artes*), with three sub-areas.

Since some journals belonged to two or more different areas of SciELO, some of the articles were stored under the concept of interdisciplinary studies. There were some overlapping between some areas, for example, Agricultural Sciences overlapped with Chemical Engineering because some of the articles discussed soil use as well as chemical components used in agriculture. The same happened to Physical and Earth Sciences when some articles discussed topics related to Agriculture and Archaeology, which were also present in other interrelated areas. Our criterion was to follow the distinction made by the SciELO platform since they certainly had a reason for separating the publications under specific areas, as well as choose the ones with higher impact within each particular area.

The impact is based on the Qualis of journals, which is a Brazilian ranking used by the Coordination for the Improvement of Higher Education Personnel (CAPES- *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) to evaluate the quality of scientific journals in Brazil. In order not to have a random sample, we selected papers from, at least, two different journals from the same scientific area, starting in 2018 so as to have the most recent issues. We also observed some articles from previous years, which had been ranked as B2, B1, A2 or A1, corresponding to the highest journal impact for Qualis. This procedure would guarantee the excellent quality of these papers in each scientific community. One example was *Acta Botanica Brasilica* which had been ranked as B2 for Biodiversity and as B5 for Biology. Then, in this case, we selected five articles whose theme had to do with Biodiversity (B2) and looked for other journals that would discuss other areas of Biology related to animals, whose score for Qualis would be, at least, B2.

Following the areas of SciELO, we selected twenty (20) articles from each journal whose writings had been published in English by Brazilian authors or teams. The journals published most of the chosen papers between 2017 and 2018. However, in some areas, such as Physics and Humanities, the most recent papers were published in 2010. We decided to keep these papers to maintain the broadest range of subareas within each domain. We accessed the electronic versions of the journals, and the texts were downloaded and saved according to criteria based on their specific areas. In a different document, we held the references for all articles used with the same tag they would have in the corpus.

Since the primary goal of compiling this corpus was to have texts that would display academic collocations and clusters, we selected the complete articles with tables, abstracts and references. The tables and figures were not a problem since the program used for analysis, *Sketch Engine*® (KILGARRIFF *et al*., 2014), does not read them.

After following the criteria previously described, we compiled a 906,035-word corpus using *Sketch Engine*. At the end of this process, the BrACE corpus data was organised as follows:

TABLE 1 – Brazilian Academic Corpus of English (BrACE)

Areas	Journals and Years	Papers	Words
1. Agricultural Sciences	1. Acta Scientiarum. Agronomy, 2018; 2. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 2018 e 2017.	20	88,740
2. Biological Sciences	1. Acta Botanica Brasilica, 2018, 2017; 2. Memórias do Instituto Oswaldo Cruz, 2018.	20	92,220
3. Health Sciences	1. Jornal Brasileiro de Pneumologia, 2018; 2.Arquivos de Neuro-Psiquiatria, 2018; 3. Brazilian Dental Journal, 2018; 4. Brazilian Journal of Pharmaceutical Sciences, 2018, 2017.	20	74,254
4. Physical and Earth Sciences	1. Brazilian Journal of Physics, 2010; 2.Revista Brasileira de Meteorologia, 2017; 3. Brazilian Journal of Oceanography, 2017; 4. Boletim de Ciências Geodésicas, 2017	20	82,440
5. Humanities	1. Ambiente & Sociedade, 2017; 2. Brazilian Journal of Political Economy, 2017; 3.Cadernos Pagu, 2010	20	151,952
6. Applied Social Sciences	1. Ambiente & Sociedade, 2017.	20	142,930
7. Engineering	1. Journal of Aerospace Technology and Management, 2018; 2. Journal of Microwaves, Optoelectronics and Electromagnetic Applications, 2017; 3. Latin American Journal of Solids and Structures, 2017; 4. Revista IBRACON de Estruturas e Materiais, 2017.	20	109,236
8. Languages, Linguistics and Arts	1. Alfa: Revista de Linguística, 2017; 2. Revista Brasileira de Estudos da Presença, 2018; 3. Ilha do Desterro, 2018, 2017.	20	164,263
TOTAL			**906,035**

Source: BrACE corpus

In the following section, we explain how we analysed the collocations in BrACE.

3.2 Selection of the most frequent academic collocations used by Brazilian researchers in comparison to frequent academic collocations in English

In this study, we used semi-automatic retrieval of collocations, that is to say, statistical information and human judgement. We used a whitelist to generate a list of words that coincided with a combination of three well-known EAP vocabulary lists:

(i) the Academic Vocabulary List (AVL-BAWE), based on the Corpus of Contemporary American English (COCA) by Gardner and Davies (2014);

(ii) the Academic Keyword List (AKL), based on the list of keywords extracted by Paquot (2010), and

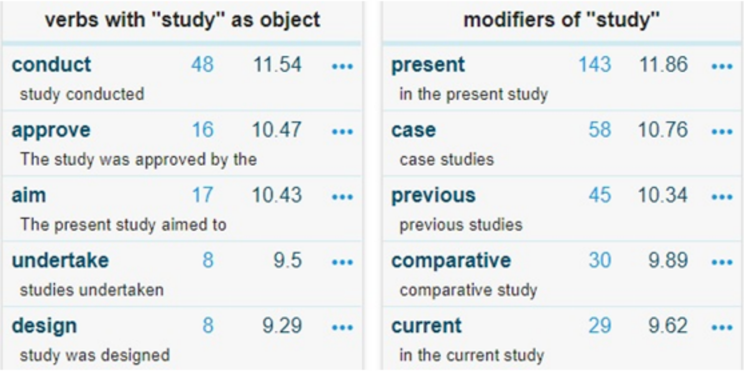
(iii) the Academic Collocations List (ACL) by Ackermann and Chen (2013).

We did this process during the time we had access to the database of the Collocaid project (FRANKENBERG-GARCIA *et al.*, 2019a) in which these lists had been used. The ColloCaid project is dedicated to developing a text-editing tool to help writers with collocations during the writing process. The research involves "investigating user needs, the visualisation of lexicographic data and human-computer interaction, and compiling an extensive database of collocation suggestions using state-of-the-art e-lexicography tools and resources".[3]

We started the selection of lexical words with nouns as base forms to observe how they would collocate most frequently in the BrACE corpus. The most frequent nouns in the list were studied. To illustrate the steps taken, we made a query with *study* as search word using a tool called WordSketch, which is a "one-page summary of a word's grammatical and collocational behaviour" (KILGARRIFF *et al.*, 2014, p. 9):

[3] Available from: https://www.collocaid.uk/.

FIGURE 1 – Screenshot of the query for "study" as a noun in the BrACE corpus



verbs with "study" as object		
conduct study conducted	48	11.54
approve The study was approved by the	16	10.47
aim The present study aimed to	17	10.43
undertake studies undertaken	8	9.5
design study was designed	8	9.29

modifiers of "study"		
present in the present study	143	11.86

case
case studies
	58	10.76

previous previous studies	45	10.34
comparative comparative study	30	9.89
current in the current study	29	9.62

Source: Sketch Engine®

In Figure 1, we see two different lists of words that are commonly combined with the search word "study". On the left, we have verbs that co-occur with the study as an object, such as "conduct + study", "approve + study", "aim + study". On the right we see modifiers of "study" as in "present + study", case +study" and "previous + study".

We selected single words that tended to co-occur in the span of three words from the reference word, coinciding at least five times in the corpus and having a LogDice score of, at least, 7. This kind of statistical data will "indicate how strong the collocation is. The higher the score, the stronger the combination of words is. A low score means that the words in the collocation also frequently combine with many other words". This decision was taken considering previous papers that reported the statistics used in the extraction of collocations from small and large corpora (CORTES, 2004, DAYRELL, 2007; ACKERMANN, CHEN, 2013; FRANKENBERG-GARCIA *et al*., 2019a).

The next step was analysing the list of (i) "modifiers" that collocated with the search word; (ii) verbs with the search word as "object" and (iii) verbs with the search word as "subject".

The search words and their collocates were saved in a list showing the frequency of each word combination to compare them to the reference list of common collocations in English.

We excluded terms (*translation/epidemiological/environmental* study; *discourse/ scientometric* analysis) and combinations with copular or auxiliaries (be – studies *were*…, have – *have* shown). The aim was to analyse general academic collocations instead of terms from specific areas. This way, we could retrieve collocations mostly used by Brazilian authors such as *the present study*, *case study*, *previous study* (modifier + study); *conduct a study*, *achieve/approve/aim a study* (verbs + study as object); *studies demonstrated*, this *study showed*, this *study suggests* (verbs + study as subject).

After selecting frequent collocations from BrACE, we looked for the ones that were not so frequently used by English authors in *The Oxford Corpus of Academic English* (OCAE), which is a 71,372,972-word corpus, to check if they were not used at all or if they were rarely used. The access to this corpus was possible during a period of a sabbatical break in which we worked with a research team who had this permission.

4 Results and Analysis

In this section, we present the results of our study concerning the academic collocations used by Brazilians in their papers published on SciELO, as well as characteristics of overuse and underuse.

4.1 Academic collocations overused by Brazilian researchers in comparison to frequent academic collocations in English

As presented in the methodology, we compared the wordlist of BrACE to the three academic vocabulary lists and selected the first twenty most frequent words, which were ranked from the most to the least frequent ones. We analysed them as candidates for academic collocations.

The first word class we observed from this list were nouns. We analysed collocations which had been frequently used by Brazilian authors with these nouns but were not as frequent in the three academic vocabulary lists commonly used by researchers who publish in English. We took this step to observe too frequent (overused) or uncommon (underused) collocations that had been chosen by Brazilian researchers and were not as frequent in papers originally written in English in the OCAE. As we will see, there were less overused collocations than the underused ones.

The overused collocations in BrACE that were not as frequent in the three lists of comparison were: *corroborate + study (obj.) / study (subj.) + corroborate / study (subj.) + reinforce / analysis (adj.) + finite / analysis (adj) + correlation / make + analysis (obj.) / consider + analysis (subj.) / intensive (adj.) + use* and *present (adj.) + work*

After comparing the collocations from BrACE to the ones in the three academic lists, we analysed the specific examples in the OCAE. Although they were all part of the OCAE, we wanted to confirm whether their co-occurrence and LogDice scores were similar. The collocations are presented in the following table that shows the base word and its relation to the collocate, an example from BrACE, its co-occurrence, and LogDice in BrACE and OCAE respectively:

TABLE 2 – collocations overused in BrACE in comparison to OCAE

Base (relation) + collocate **Example from the BrACE**	**Co-oc. BrACE**	**LogDice**	**Co-oc. OCAE**	**LogDice**
1. Study (obj. of) corroborate These results *corroborate* previous biomechanical *studies* that found a lower stress concentration for wide diameter implants, especially in short implants. (|Health)	6 (4.89 per million)	9.06	4 (0.05 per million)	3.78
2. Study (subj. of) corroborate Several other *studies* have *corroborated* these findings, which indicate a change in cardiac autonomic modulation, demonstrating impairment of this activity in individuals with COPD. (Health)	5 (4.07 per million)	7.92	5 (0.06 per million)	4.29
3. Study (subj. of) reinforce This *study reinforces* the lines already traced out in recent research on the need to consider multidimensional approaches when analysing human-nature relationships. (Social Sciences)	5 (4.07 per million)	7.91	6 (0.07 per million)	4.44
4. Analysis (adj.) finite *Finite element analysis* on the influence of implant surface treatments, connection and bone types. (Health)	22 (17.91 per million)	9.26	50 (0.60 per million)	5.51
5. Analysis (adj) correlation The RDC results (Figure3) were similar to the results obtained for the *correlation analysis* for the two study years and all of the soil layers measured, with a correlation of -0.88 between the RDC and Pearson's correlation. (Agriculture)	10 (8.14 per million)	8.22	108 (1.28 per million)	6.6
6. Analysis (obj. of) make In the context described above, we *analys*ed the potential impacts from the installation and operation steps of this project, considering the understanding of the oceanographic processes and possible effects on the human well-being caused by the undermining of provided ecosystem services. (Biological Sciences)	9 (7.33 per million)	8.68	161 (1.91 per million)	6.06

7. Analysis (subj. of) consider This *analysis considers* the average of 100 independent runs. (Engineering)	6 (4.89 per million)	9.12	18 (0.21 per million)	6.55
8. Use (adj.) intensive In American agriculture, the conversion of conventional tillage systems to no-till systems and the *intensive use* of glyphosate in transgenic cropping has significantly influenced the composition and populations of weeds. (Agriculture)	5 (4.07 per million)	8.97	48 (0.57 per million)	6.15
9. Work (adj. of) present In the *present work*, it was evident that cellular debris from both the uterine epithelium and the trophoblastic cells are phagocytosed and digested by active trophoblastic cells. (Biological Sciences) 27 (21.99 per million) 10.73			103 (1.22 per million)	6.57

Source: Authors

As shown in Table 2, some collocations did not have a LogDice higher than 7.0 in the OCAE despite the fact they had higher LogDice in the BrACE, which could be an indication of overuse. These collocations were: *corroborate study*; *study confirms*; *study reinforces*; *finite element analysis*; *correlation analysis, make analysis, the analysis considers* and *intensive use*.

Although we found these collocations in the OCAE, they are not so frequently used in papers written by native English authors. Besides, the same collocations were not frequent in the combination of academic lists as well.

We looked up for collocational options with the same nouns in the OCAE that could replace the ones used by Brazilians. To do so, we used the same nouns as search words in Word Sketch to look for collocations with similar meanings. However, we looked for combinations with higher frequency in the OCAE, which might sound more natural to international researchers. For the collocations with *study* + *corroborate*, the optional choices in the OCAE would be: *study* + *support* / *confirm*. So, the sentence below, taken from BrACE, could be written in the following way:

> These results [*support*] [*confirm*] previous biomechanical *studies* that found a lower stress concentration for wide diameter implants, especially in short implants.

> Several other *studies* have [*supported*] [*confirmed*] these findings, which indicate a change in cardiac autonomic modulation, demonstrating impairment of this activity in individuals with COPD.

As for the collocation *study* + *reinforce*, a similar meaning with more substantial LogDice score would be *study* + *highlight.* In this case, the sentence used by the Brazilian author would be:

> The findings of the present *study* [*highlight*] the importance of banning tobacco displays at the point of sale.

Although the collocation *finite element* + *analysis* did not show a high LogDice score (5.51) in the OCAE, we found it in the sub-corpus of Engineering, which could mean it is a discipline-specific collocation, as in the example below:

> The *finite element analysis* of any problem involves four steps: (a) discretising the solution region into a limited number of sub-regions or elements, (b) deriving governing equations for a typical feature, (c) assembling all the parts in the solution region, and (d) solving the system of equations obtained.

A similar case is the collocation *correlation* + *analysis*, which has a low LogDice score in the OCAE (6.6) but is used in the areas of Medicine, Education and Computer Sciences, as the examples below:

> To examine the role of parents and friends as sources of influence on girls' college aspirations and motivation to achieve their goals, we conducted a series of *correlation analyses* separately for girls who were sexually active and those who were not.

Although the collocation *make* + *analysis* was high, other options found in the BrACE would be more aligned with the OCAE such as *perform/conduct/apply* + *analysis*. Therefore, the sentence below would sound more natural in the following way:

> In the context described above, we [*performed*] [*conducted*] [*applied*] an *analysis* of the potential impacts from the installation and operation steps of this project (…)

The collocation *analysis* + *consider* was not present among the most common collocations of OCAE. We believe the best option, in this case, would be the expression "the analysis takes into consideration" as in:

> This *analysis takes into consideration* the average of 100 independent runs.

The collocation *intensive use* could be substituted by *widespread/increased/unrestricted/extensive* + *use* as it is found in the OCAE.

In American agriculture, the conversion of conventional tillage systems to no-till systems and the [*widespread*] [*increased*] [*unrestricted*] [*extensive*] use of glyphosate in transgenic cropping has significantly influenced the composition and populations of weeds.

The next section of this article presents collocations that were underused by Brazilian authors.

4.2 Academic collocations underused by Brazilian researchers in comparison to frequent academic collocations in English

This time, we checked collocations that had not been so frequently used by Brazilian authors with the list of nouns we had, but were present in the EAP lists we had used in the first step. The underused collocations from BrACE that were not as frequent in the three lists of comparison were: *qualitative (adj.) + study / detail (adj.) + analysis / restrict + analysis (obj.) / extensive (adj.) + use / widespread (adj.) + use / increase (adj.) + use / support + use (obj.) /* encourage + *use (obj.) / design + system (obj.)/ system (subj.) + work / describe + process (obj.) / begin + process (obj.) / collect + data (obj.) / data (subj.) + suggest / data (subj.) + indicate / development (subj.) + occur / facilitate + development (obj.) / further (adj.) + development.*

Once again, we compared the co-occurrence of these collocations in BrACE to the OCAE, and we present the eighteen first ones below within their context in the OCAE:

TABLE 3 – Collocations underused in BrACE in comparison to OCAE

Base (relation) + collocate **Example from the OCAE**	**Co-oc. BrACE**	**LogDice**	**Co-oc. OCAE**	**LogDice**
1. Study (adj.) qualitative This *qualitative study* found that while knowledge about the TRiM system was not widespread, the majority of those personnel who were aware of TRiM viewed it positively and supported it being peer-delivered. (Medicine)	0	0	378 (4.48 per million)	7.82
2. Analysis (adj.) detail Poor exposures and a lack of reliable criteria prevent *detailed analysis* of the lower slope. (Earth Sciences)	2 (1.63 per million)	5.91	550 (6.51 per million)	8.73
3. Analysis (obj. of) restrict Therefore, we *restrict* our *analysis*, somewhat arbitrarily, to deflections of the form: w (x, y) = e wEu (x) + s wS (x) cos (kS y) + a wA (x) cos (kA y + A) (8.74). (Engineering)	0	0	103 (1.22 per million)	8.22
4. Use (adj.) extensive This propaganda campaign made *extensive use* of petitions as a device for expressing extra-parliamentary pressure on a public issue. (History)	0	0	188 (2.23 per million)	7.91
5. Use (adj.) widespread Although the genetics of many lower eukaryotic organisms had been studied in some detail, Beadle and Tatum's work initiated a much more *widespread use* of microbes. (Biochemistry)	1 (0.81 per million)	6.68	306 (3.62 per million)	8.74
6. Use (adj.) increase We are seeing the *increasing use* of computational modeling within historical linguistics and interdisciplinary research is not the exotic enterprise it used to be. (Linguistics)	0	0	383 (4.54 per million)	8.14
7. Use (obj.of) support The current analyses further *support* the *use* of extreme groups with an underpinning rationale. (Education)	0	0	160 (1.90 per million)	7.89
8. Use (obj.of) encourage The ease of storing, transmitting, and processing electrical data is *encouraging* the *use* of unmanned stations. (Earth Science)	0	0	110 (1.30 per million)	7.74
9. System (obj. of) design A small group of engineers (3 full-time and 4 part-time workers) used axiomatic design *to design* a *system* that can satisfy the requirements for crew survivability in a short time (5 months). (Engineering)	2 (1.63 per million)	7.29	460 (5.45 per million)	8.75

10. System (subj. of) work As I do not ever recollect an urgent request having been refused, this *system worked* very satisfactorily from our point of view. (Engineering)	0	0	133 (1.58 per million)	8.71
11. Process (obj. of) describe The *process described* there, by which lay people decide which action to take about symptoms of illness, is probably not greatly different from the general way that doctors diagnose illness. (Medicine)	3 (2.44 per million)	7.53	430 (5.09 per million)	8.84
12. Process (obj. of) begin Shot 7 initiates a new line of dramatic action that poses the question of what Lucy will do now, and also *begins a process* not exactly of rereading, but a search for a new reading of the meaning of the setups. (Media Cultural Studies)	0	0	164 (1.94 per million)	8.57
13. Data (obj. of) collect Lyons et al. (1998a) noted that geochemistry data from Lake Fryxell in the McMurdo Dry Valleys indicated an overall change in the ionic composition of the lakes when *data collected* in the mid-1990s are compared to older, but reliable, data obtained in the early 1960s. (Biological Sciences)	1 (0.81 per million)	7.83	1,786 (21.16 per million)	11.36
14. Data (subj. of) suggest We presented the 10th-grade classroom because our *data suggest* that Ms. Young fits Irvine's description of an "experienced and masterful pedagogue" who is "seeing with the cultural eye" (Irvine, 2001). (Education)	0	0	256 (3.03 per million)	9.47
15. Data (subj. of) indicate Together these *data indicated* a decline in grasslands and an increase in shrublands in the early Holocene. (Earth Science)	1 (0.81 per million)	7.21	132 (1.56 per million)	9.12
16. Development (subj. of) occur The next *development occurred* in the plateau country of Arizona, Utah, and Colorado. (Earth Science)	0	0	56 (0.66 per million)	7.24
17. Development (obj. of) facilitate It is surely in the interest of countries near and far away to *facilitate* the *development* of knowledge, skill, and freedom in these countries so they can become contributing, responsible members of the international community rather than breeding grounds for social pathology, infectious diseases, and terrorist violence. (Education)	0	0	126 (1.49 per million)	8.24
18. Further (adj. of) development The establishment and *further development* of this cascade provides us with a fertile research agenda. (Physical and Earth Sciences)	0	0	382 (4.52 per million)	8.16

Source: Authors

The collocations presented above have not been frequently used by the Brazilian researchers in their texts represented in our corpus. The examples were all taken from *The Oxford Corpus of Academic English*, which means that to have a more natural text, it would be necessary for Brazilian researchers to be aware of this use and try to incorporate these collocations into their writings.

In the next section, we discuss the general results of this study based on the observation of overuse and underuse of academic collocations used by Brazilian researchers in their articles.

5 Discussion

The discussions presented in this section seek to answer the three research questions stated at the beginning of this paper. The first one was "To what extent do the collocations used by Brazilian authors differ from the ones in international journals?". Although Brazilian researchers have had their papers published in high-impact academic journals, we could see that there are significant differences regarding underused collocations, which outnumber the overused ones. This result shows that these writers were not aware of some of the collocations mostly used by scholars in international journals. These extracts are not so different to Brazilian Portuguese such as a *detailed (adj.) + analysis, restrict + analysis (obj.), extensive (adj.) + use, widespread (adj.) + use, describe + process (obj.)* and *begin + process (obj).* We did not expect some of the results such as the underuse of collocations as *collect + data* and *data + suggest* which are not so different from the Brazilian Portuguese. Because of that, further studies will be carried out as soon as we have more articles added to the BrACE corpus so that we can confirm or not the lack of some collocations in those articles.

The previous result leads us to the second and third questions, which are: "Do Brazilian authors use collocations influenced by their native language (Brazilian Portuguese)?" and "Are there traces of overuse or underuse of specific collocations?".

We could find evidence that indicates the influence of Brazilian Portuguese in the choice of collocations which called our attention. This is the case of s*tudy* (obj. of) + *corroborate* and s*tudy* (subj. of) + *corroborate* which were overused by the Brazilian researchers and have the equivalent in Portuguese "estudo (obj of) + corroborar" and "estudo (subj. of) + corroborate" which are very common in articles written in this language. This result pointed out to the trace of collocation overuse. Although this combination has been found in the OCAE, it is not as frequent in research papers initially written in English, which clearly shows the influence of Portuguese in those texts.

Another comparison we can make is that Brazilians *suggest the use of* whereas authors who commonly write in English *support the use* o*r encourage* it. At the same time, instead of *data points*, Brazilians most commonly write *data indicates that*.

Upon analysing different areas of research, the collocation *qualitative study* is present in areas such as Business, Medicine and Sociology in the OCAE. In contrast, in BrACE, we find *qualitative analysis*, but not a *qualitative study*. The same happens to *regression analysis*, which is the first most frequent collocation with research in the OCAE but is not present in the BrACE. In cases like this, it is necessary to consider that the BrACE is still a small corpus of 906,035 words and some collocations not found here may start to appear as the corpus grows. These limitations do not allow us to generalise the behaviour of academic collocations as a whole but show Brazilian researchers' preferences.

It would be desirable to compare the results shown here to international authors who frequently publish in renowned journals of different domains.

Regarding the methodology, as stressed by Dayrell (2011), it would be interesting to analyse a lemmatised corpus to see the behaviour of the same lemma in different contexts as well as different span values and strength of association between nodes and collocates. Another interesting perspective would be the investigation of an additional criterion of window-sizes of collocations that could range more four words to the right and the left. It would allow us to observe longer phraseologies in research papers written by Brazilian or international researchers.

6 Final Remarks

The primary aim of this study was to identify the most frequent collocations used by Brazilian authors who had their research papers published in the eight major areas of SciELO. After identifying these collocations, we compared them to the most frequent academic ones used by native English writers and international research groups so we could locate academic collocations that had been overused and underused by Brazilian researchers.

These results have led us to suggest further studies and actions to encourage Brazilian researchers to write more naturally in English academic style. By doing so, they will become aware of these differences in academic language that may not have been noticed in their writings.

As suggested by Nesselhauf (2003), teachers could point out the most relevant collocations through writing exercises in academic workshops or courses of academic English. The author argues that we should explicitly teach collocations since they do not always stand out to the learners' eyes. The main suggestion is to start by introducing the most frequent and acceptable collocations and, then, comparing them to native researchers' textual productions. Having these results, consequently, we could stress functional elements such as the difference between possible combinations in English and those that are more common in the students' native language. In this way, we believe that researchers would be more familiar with the language patterns used in research papers published in high-impact academic journals.

It would also be desirable to encourage students to write abstracts and papers when they are still in college so they become more and more familiar with the academic English. Also, teachers should encourage students to read as many quality papers written in English as possible so that students became aware of their specific research communities writing style. This practice would certainly enhance the use of academic collocations. Another way of stimulating the students to use more collocations would be explicitly showing them samples of sentences containing these structures.

As for senior researchers, it would be ideal to show them the collocations commonly used in their areas through writing crash courses and by teaching them to compile their corpora to be used as examples of writing in each area. By doing so, they would be acquainted not only with the language style and structure but also with genre constraints in each area.

Actions like these have already been taken as, for example, the writing masterclasses supported by the British Council in which Brazilian researchers and EAP tutors (FRANKENBERG-GARCIA *et al.*, 2019b) worked together to develop their writing autonomy through the use of specialised corpora and linguistic tools.

Acknowledgements

We would like to thank Dr. Ana Frankenberg-Garcia and Dr. Geraint Rees for all valuable suggestions during the development of this research work at the University of Surrey. The authors would also like to acknowledge funding from the São Paulo Research Foundation (FAPESP/16/25198-6).

Authorship statement

This study reports on data from Dr. Paula Tavares Pinto's Post-Doctoral research at the University of Surrey. The first author was in charge of gathering data, transferring the data to spreadsheets for data analysis, and writing the first draft of the article. The four authors collaborated on interpreting results and revising the essay and the data analysis, including the statistics.

References

ACKERMANN, K.; CHEN, Y. H. Developing the Academic Collocation List (ACL). A Corpus-Driven and Expert-Judged Approach. *Journal of English for Academic Purposes*, [*S.l.*], v. 12, n. 4, p. 235-247, 2013. DOI: https://doi.org.10.1016/j.jeap.2013.08.002

BABINI, M.; SILVA, E. B. A terminologia acadêmica nos textos científicos em língua inglesa uma abordagem baseada em corpus. *In:* ISQUERDO, A. N.; SEABRA, M.C.T.C. (org.). *As ciências do léxico:* lexicologia, lexicografia, terminologia. Campo Grande: UFMS, 2012. p. 415-427.

CORTES, V. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for specific purposes*, [*S.l.*], v. 23, n. 4, p. 397-423, 2004. DOI: https://doi.org.10.1016/j.esp.2003.12.001

DAYRELL, C. A Quantitative Approach to Compare Collocational Patterns in Translated and Non-Translated Texts. *International Journal of Corpus Linguistics*, [*S.l.*], v. 12, n. 3, p. 375-414, 2007. DOI: https://doi.org.10.1075/ijcl.12.3.04day

DAYRELL, C. Corpora no ensino de inglês acadêmico: padrões léxico-gramaticais em abstracts de pós-graduandos brasileiros. *In*: VIANA, V.; TAGNIN, S. (org.). *Corpora no Ensino De Línguas Estrangeiras*. São Paulo: HUB Editorial, 2011. p. 131-172.

FIRTH, J. R. *Modes of Meaning*. v. 4: Essays and Studies (English Association). Indianapolis: Bobbs-Merril, 1951. p. 118-149.

FRANKENBERG-GARCIA, A. *et al*. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL*, Cambridge, v. 31, n. 1, p. 23-39, 2019a. DOI: https://doi.org.10.1017/S0958344018000150

FRANKENBERG-GARCIA, A. *et al*. *Supporting the Internationalisation of Brazilian Research*: Curso oferecido via financiamento *Capes: Print* para a Universidade Federal do Rio Grande do Sul e a Universidade Estadual Paulista, 4-06 de jun.de 2019. 30f. Notas de aula.

GARDNER, D.; DAVIES, M. A New Academic Vocabulary List. *Applied Linguistics*, Oxford, v. 35, n. 3, p. 305-327, 2014. DOI: https://doi.org.10.1093/applin/amt015

HASWELL, R. *Gaining Ground in College Writing:* Tales of Development and Interpretation. Dallas: Southern Methodist University Press, 1991.

HYLAND, K. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics*, [*S.l.*], v. 18, n. 1, p. 41-62, 2008. DOI: https://doi.org.10.1111/j.1473-4192.2008.00178.x

KILGARRIFF, A. *et al*. The Sketch Engine: Ten Years On. *Lexicography*, Sheffield, UK, v. 1, n. 1, p. 7-36, 2014. DOI: https://doi.org/10.1007/s40607-014-0009-9

KUHN, T. A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students. 2017. 421f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras, Universidade de Lisboa, Lisboa, 2017.

LEA, D.; CROWTHER, J.; DIGNEN, S. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press, 2002.

NATION, I. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press, 2001. (Cambridge Applied Linguistics).

NESSELHAUF, N. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, Oxford, v. 24, n. 2, p. 223-242, 2003. DOI: https://doi.org/10.1093/applin/24.2.223

NEVES, M. L.; JIMENO-YEPES, A.; NÉVÉOL, A. The SciELO Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUTATION – LREC, 10th., 2016, Portorož, Slovenia. *Proceedings* [...]. Portorož: LREC, 2016.

PAQUOT, M. *Academic Vocabulary in Learner Writing*: From Extraction to Analysis. London: Continuum, 2010.

PARTINGTON, A. *Patterns and Meanings:* Using Corpora for English Language Research and Teaching. Amsterdam: John Benjamins Publishing, 1998. DOI: https://doi.org/10.1075/scl.2

PAIVA, P. T. Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em corpus. 2009. 288f. Tese (Doutorado em Linguística Aplicada) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2009.

SCIENTIFIC ELECTRONIC LIBRARY ONLINE. Available from: https://scielo.org/. Access on: Jun. 13, 2020.

SCOTT, M. *WordSmith Tools 4*. Oxford: Oxford University Press, 1996.

SILVA, E. B.; BABINI, M.; OTTAIANO, A. O. Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora. *Acta Scientiarum*, Maringá, v. 39, p. 345-353, 2017. DOI: https://doi.org/10.4025/actascilangcult.v39i4.31811

SILVA, L. G.; MATTE, M. L.; SARMENTO, S. Brazilian Students's Use of English Academic Vocabulary: An Exploratory Study. *In:* FINATTO, M. J. *et al.* (org.). *Linguística de corpus:* perspectivas. Porto Alegre: Instituto de Letras, UFRGS, 2018. p. 509-526.

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

TAGNIN, S. E. O. *O jeito que a gente diz* - combinações consagradas em inglês e português. Barueri: Disal Editora, 2013.