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

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

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

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

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

SIMONE SARMENTO
ROZANE REBECHI
MARINE LAÍSA MATTE
(ORG.)

Porto Alegre • 2024 • 1ª edição

editora ZOUK

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

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

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

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

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

CDD 410
2024-175 CDU 81'1

Editora Zouk
r. Cristóvão Colombo, 1343 sl. 203
90560-004 – Floresta – Porto Alegre – RS – Brasil
f. 51. 3024.7554

www.editorazouk.com.br

Investigating Brazilian English Learners' Use of Academic Collocations: A Corpus-Based Study

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

Introduction

Writing has a special role in academic contexts as it is one of the main skills students have to master in order to achieve academic success (Biber & Gray, 2016). In addition, at the higher education (HE) level, academic literacies are being learned and tested all the time. New ways of constructing knowledge are constantly being discovered (Lea & Street, 1998), and these practices are necessarily dependent upon academic writing. In spite of the importance writing plays in academic contexts, it is usually assumed that students should already know the rules and conventions of this practice. However, these rules are not transparent, forming what Lillis (2001) called "practices of mystery". For Lillis, students who are not familiar with academic writing conventions may have their participation in HE impaired. Thus, these conventions should be explicitly taught since we cannot depend on incidental learning or on a hidden curriculum, as students "must now gain fluency in the conventions of English language academic discourses to understand their disciplines and to successfully navigate their learning." (Hamp-Lyons, 2002: 1)

Academic language is a specific subset of general language, and differs considerably from the type of language used in daily life situations, not only in terms of formality but also in terms of language features (Simpson-Vlach & Ellis, 2010). The language features specific to academic contexts may range from, for instance, choice of verb to combination of words, i.e. collocations. Collocations are also important in general language, but

gain even more importance in academic registers. According to Sinclair (1991: 110), any language user will have a repertoire of "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." In other words, proficient language users resort to collocations to convey meaning. Therefore, mastering collocations is imperative for guaranteeing fluency in a text, as writing proper academic English goes beyond knowing isolated words. When it comes to judging text quality, one of the criteria a reader has in mind, however unconsciously, is how conventionalized language is. This conventionality is partly guaranteed by the appropriate use of collocations.

Bearing the importance of collocations for academic texts in mind, the main goal of this study is to analyze how Brazilian students produce collocations in academic texts written in English by comparing two corpora of unpublished texts: one with texts produced by Brazilians studying in British universities (BrAWE) whose grades are unknown, and the reference corpus with texts written by students from multiple nationalities studying in British universities but which were graded with merit or distinction (BAWE). The latter will be used as baseline data. The following research questions will be addressed:

a) Is there a statistically significant difference in the frequency of the noun nodes and their respective collocates in BrAWE and BAWE?
b) Are there differences in syntactic structures of collocations between the two corpora?

Collocations

"You shall know a word by the company it keeps" is probably a sentence that immediately comes to mind of anyone acquainted with collocational studies. This sentence formulated by J. Firth (1957: 11) has inspired a great deal of research in the field, as it summarizes the core meaning of collocations, i.e., the likelihood of two or more words occurring together ((Sinclair, 1991; Hill, 1999; Durrant, 2009). Sinclair (1991) proposed the idea that language operates according to the open-choice principle and the idiom principle. The former considers language as the result of complex

choices to complete each unit (word, phrase and clause) that composes a text, i.e., all slots of a text can be filled with any word as long as grammaticality is preserved. The latter assumes that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments." (Sinclair, 1991: 110).

Regarding language learners, evidence shows that they do make use of collocations but tend to have a more limited repertoire of conventional combinations (Granger, 1998; Lorenz, 1999; Nesselhauf, 2005). The comparison between native (NS) and non-native (NNS)[1] collocational performance is presented in Howarth (1998), who analyzes adult learners of English writing academically in Social Sciences postgraduate courses and focuses on the use of collocations composed of verb + noun. The study reveals that the NNS "produced, on average, a much lower density of conventional combinations (25%), suggesting either a generally lower level of knowledge of collocations, or a lack of awareness of how to deploy them appropriately, or both." (Howarth, 1998: 36).

Granger (1998) analyzes intensifying adverbs ending in –ly that function as amplifiers and modifiers as the nodes of the collocations. By comparing a corpus of native English writers to a similar corpus of advanced French-speaking learners of English, the data revealed a statistically significant overall underuse of amplifiers in the learner corpus. However, when looking at some amplifiers individually, *completely* and *totally* were overused by the learners, while *highly* was underused. Granger suggests that this overuse can possibly be explained by the fact that these adverbs have direct equivalents in French and, consequently, students choose to translate them from French into English. Additionally, some combinations with amplifiers such as *acutely aware*, *bitterly disillusioned*, *gravely disorganised*, and *steeply dipping* are used exclusively by native speakers.

1 It is important to point out that most studies related to proper use of collocations rely on a contrastive analysis between native speakers (NS) and non-native speakers (NNS). However, in this study the comparison was not based on a NS vs. NNS dichotomy.

Collocations composed of adjective + noun or noun + noun are analyzed by Durrant and Schmitt (2009). The authors analyze a total of 96 texts organized in two sets: one containing NNS texts and the other NS texts. By classifying collocations into low-frequency and high-frequency and establishing collocational strength with t-score and Mutual Information measures[2], they came to three main findings: Firstly, native writers use more low-frequency combinations than non-natives. […] Secondly, non-native writers make at least as much use of collocations with very high t-scores as do natives. […] Thirdly, non-native writers significantly underuse collocations with high mutual information (MI)[3] scores in comparison with native norms (Durrant & Schmitt, 2009). These findings suggest that learners have a tendency to repeat favored items, as they quickly pick up frequent collocations because the less frequent and strongly associated items take longer to acquire (Durrant & Schmitt, 2009). Ellis, Simpson-Vlach and Maynard (2008) reinforce this idea that NS use a wider range of collocations, whereas NNS tend to use collocations they encounter more frequently. The issue of overusing collocations is discussed by Ackerman and Chen (2013: 3), who argue that "by using a less appropriate collocate, a non-native speaker will sound unnatural or may even become unintelligible among speakers of the target language."

Laufer and Waldman (2011) investigate *verb + noun* collocations produced by L1-Hebrew learners of English. Besides comparing the learner corpus to a NS one, the authors also compared the data within L1 Hebrew learners of English represented in the corpus. Results indicated that the NS produced almost twice as many collocations as the learners. Learners underused verb + noun collocations when compared to NS and produced significantly more deviant collocations. Advanced and intermediate learners

2 The t-score is an association measure that "highlights frequent combinations of words. [H]owever while all collocations identified by the t-score are frequent, not all frequent word combinations have a high t-score. [On the other hand], MI-score is negatively linked to frequency, meaning that the value is larger the more exclusively the two words are associated and the rarer the combination is." (Gablasova et al., 2017: 8-9).

3 MI is a measure of association between words. The higher the MI score, the stronger the relation between the items (Church & Hanks, 1990)

were the ones who produced more deviant collocations, probably because they feel more confident in relation to the English language when compared to basic students.

Chinese learners of English and their use of collocations in academic written texts were investigated by Wu (2016). The author analyzed verb + adverb and adverb + verb collocations comparing three academic English corpora, two of NS and one of NNs. Wu (2016) also shows that there are significant differences in terms of collocations chosen by Chinese learners of English who use, for instance, *develop quickly*, *widely use* and *abolish completely* more frequently than NS do. This difference regarding lexical competence and knowledge of collocations might be related to the fact that the teaching of collocations is not common in China, and that Mandarin and English have only few similarities.

Ohlrogge (2009) analyzed 170 written compositions written for an EFL proficiency test and found correlations between level of proficiency and collocations. Hence, the students who received higher grades presented a higher incidence of collocations. This follows what Crossley et al (2015) state regarding the relation between proficiency and collocations. After having investigated lexical proficiency in both oral and written texts produced by learners of three different levels (beginning, intermediate and advanced), raters judged the lexical proficiency according to analytical and holistic features, one of them being collocations. Results indicate that higher proficiency writers tend to use a wider range of collocations than lower proficiency writers, corroborating what was found in our study.

When it comes to the analysis of collocations used by Brazilian learners of English in academic genres, more specifically in argumentative essays, Guedes (2017) explored *verb + adverbs* ending in *-ly* collocations. The author found that the most common verbs used by the learners are action verbs (*apply* and *provide*). Also, there is a high frequency of verbs such as *improve*, *develop*, and *adopt* among learners of English. On the other hand, verbs such as *increase*, *include*, *occur*, *reduce*, and *require* are more frequent in BAWE. Due to the low frequency of *verb + adverbs* ending in *-ly* their collocational strength could not be statistically measured.

Matte and Rebechi (2019) analyzed the differences in the use of collocations of the Academic Collocation list (ACL)[4] (Ackermann & Chen, 2013) in the same corpora used in the present study. Their results show that only a few collocations of ACL are used differently in the comparative analysis of BAWE and BrAWE. Furthermore, the most frequent collocations in both corpora are not exactly the same presented in the list, which suggest a possible mismatch between what is presented in ACL and authentic language produced by students both in BrAWE and BAWE.

There are ready-made lists containing relevant collocations and formulas to be mastered, as those presented in the ACL (Ackermann & Chen, 2013) and the Academic Formulas List (AFL) (Simpson-Vlach & Ellis, 2010). However, despite the "progression in research from studies that provide evidence of the importance of collocations for L2 learners" (Boers & Webb, 2018), it is necessary to create pedagogical materials that fit students' needs. Thus, more than memorizing vocabulary and collocation lists, it is imperative to master collocations in terms of knowing their appropriate use, that is, collocational competence must be acquired in context. This argument is sustained by Frankenberg-Garcia (2018: 101), who points out that "the lexical knowledge is not just about understanding words, but also about employing words in context.".

The corpora

The BAWE corpus (Alsop & Nesi, 2009) was compiled with the objective of gathering unpublished written assignments from students of multiple nationalities studying[5] in four different British universities: Warwick University, Reading University, Oxford Brookes University, and Coventry University. Unlike other academic corpora that are mostly composed of texts written by experts and edited by professionals, the BAWE is composed of discipline-specific learner texts. Despite containing students' writing, this corpus is different from those compiled with essays written under

4 https://www.eapfoundation.com/vocab/academic/acl/

5 BAWE contains texts of undergraduate and master's students.

examination conditions for analyzing non-native-speaker error and language acquisition, as it contains assignments written during undergraduate and master courses which were graded merit or distinction. The BAWE corpus was, thus, designed to enable the investigation of academic literacy and disciplinary knowledge development. BAWE has 6,968,089 words and it is balanced into four areas[6]: Life Sciences (LS), Social Sciences (SS), Physical Sciences (PS), and Arts and Humanities (AH). Each area encompasses a variety of disciplines. Moreover, the corpus is organized according to 13 different academic genre families proposed by Gardner and Nesi (2013). A total of 2,858 texts were compiled, being 1,953 written by L1 speakers of English and the remainder by highly proficient English as an additional language (EAL) students.

The Brazilian version of BAWE is BrAWE (the Brazilian Academic Written English corpus) compiled by Goulart (2017). The organization of the corpus is similar to the British one, as it covers the same areas of expertise and gathers assignments produced by undergraduate students. Therefore, BrAWE also follows Gardner and Nesi's (2013) classification of academic genre families, but only 12 categories were found. The final version of the corpus contains 380 assignments of students from 59 universities. The high number of universities involved is due to the fact that most of the students were participants of the Sciences without Borders (SwB) program, which partnered with over 80 universities in the United Kingdom alone. The SwB was a Brazilian scientific mobility program created in 2011 with the objective of strengthening and expanding the internationalization of Brazilian higher education by providing scholarships for both students and researchers.

Overall, engineering, natural sciences, and health sciences were the areas covered by the SwB. Areas such as arts and humanities were not contemplated by the program, but some texts from this area were included in the corpus as some students from other mobility programs were also contacted. Despite being comparable to BAWE, the corpus is unbalanced in terms of size of subcorpora. Considering that Life Sciences (LS), Social

6 Alsop and Nesi (2009) refer to these areas as disciplinary groups.

Sciences (SS) and Physical Sciences (PS) are the most representative areas in BrAWE, a subcorpus of BAWE was created in order to make it comparable to the BrAWE corpus. Thus, whenever BAWE is mentioned, we are referring to BAWE's subcorpora that contain only assignments in the fields of LS, SS, and PS.

	BAWE	**BrAWE**
Words	3,312,196	768,323[7]
Number of assignments	2,761	380
Quality of assignments	Merit and distinction	Passing (and higher)

Table 1. BAWE and BrAWE corpora

As stated above, the attested quality of assignments distinguishes BAWE and BrAWE. In BAWE, students were attributed merit and distinction, whereas in BrAWE students may have obtained a passing grade by the minimum requirement, which does not necessarily mean that no one wrote outstanding texts. Although grades were not given because of the quality of language, one can speculate that language may indeed play an important part in the quality of an assignment. According to Kumar and Rao (2018: p. 9), "poor academic writing skills and lack of command over the knowledge of English language" feature among the reasons why manuscripts are rejected. Therefore, due to the quality of texts, and to the high level of English language proficiency of participants, BAWE may be considered an adequate reference corpus to fulfill the purposes of a contrastive corpus analysis.

Methodological procedures

Collocations can be analyzed according to the frequency of the words or to the strength of association between the composing words using statistical measures, such as MI, t-score, Log Dice (Brezina, 2018). In this study,

7 The size of BrAWE in Sketch Engine is 768,323 rather than 670,314, as shown in Table 3, because this software counts punctuation marks as words.

we used Log Dice to calculate the strength of association between words since this is the default statistical measure of Sketch Engine, the software used to extract the collocations.

Three different types of collocations[8] were investigated: modifier + noun, noun (subject) + verb, and verb + noun (object). For example,

- Modifier: adjectives that come before the node
  Ex.: *difficult* + *task*, *advanced* + *technique*

- Verb (object of): used when the node is the object of the verb
  Ex.: *execute* + *task*, *apply* + *technique*

- Verb (subject of): used when the node is the subject of the verb
  Ex.: *task* + *require*, *technique* + *use*

These categories of collocates follow Frankenberg-Garcia et al.'s list (2018) composed of 187 collocational nodes which is a merge of three lists: the Academic Vocabulary List[9] (AVL, Gardner & Davies, 2014), the Academic Keyword List[10] (AKL, Paquot, 2010), and the Academic Collocations List (ACL, Ackermann & Chen, 2013). Of these 187 nodes 125 are nouns, 38 are verbs, and the remaining 24 are adjectives.

We focused on the identification of overused and underused academic noun-node collocations, through the comparison of two different corpora, the British Academic Written English corpus (BAWE) and the Brazilian Academic Written English corpus (BrAWE). The cut-off point to include a collocation in the study was a minimum frequency of four occurrences in BAWE in at least two out of the three remaining areas, i.e. Life Sciences, Health Sciences, and Social Sciences. Thus, collocations of one-single area were not included, as it is the case of *health need*, a collocation that only appears in LS assignments. The five methodological steps were:

8 The main word of a collocation is called node, and the ones associated to the node are the collocates. Thus, the basic structure of a collocation is node + collocate.

9 Derived from BAWE.

10 https://uclouvain.be/en/research-institutes/ilc/cecl/academic-keyword-list.html

1st: Listing in descending order the 125 nouns from the Frankenberg-Garcia et al.'s list (2018) from the most to the least frequent in BAWE by using the "search" tool in Sketch Engine[11]. The node was typed in the "lemma" box and the PoS noun was selected. All the words that derive from the base form of the node came up as a result, for example for *approach*, the plural form – *approaches* – was also selected. This procedure was repeated for every noun, i.e., for the 125 nodes.

2nd: Extracting the collocates of the 125 nodes in both corpora using the "Word Sketch" tool. The following syntactic structures mattered to this study: *(different + approach)*, *object of* (verb) (*use + approach)*, and *subject of* (verb) (*approach + involve)*. Again, the node was typed in the "lemma" box in "word sketch", and the PoS – noun was selected.

3rd: Calculating the Log Likelihood (LL) value (Rayson, 2002) for the different frequencies of each one of the 125 nodes in both corpora. If the outcome of the statistical test is 6.63 or higher, there is a 99% chance that the results are not random ($p<0.01$).

4th: Calculating the statistical significance of the collocates using LL to determine whether the comparison of frequencies of the collocates of each individual noun in both corpora was statistically significant ($p<0.01$). The frequencies of each collocate were verified in both corpora, and the LL value was calculated.

5th: Verifying the syntactic structure of the collocates that go together with each of the 125 nodes in order to check if different patterns emerge in the comparison between both corpora.

Results and discussion

The most frequent of the 125 nodes in both corpora is *system* (1.38 per thousand words in BAWE and 1.60 per thousand words in BrAWE) and the node with the lowest frequency is *exception* (0.03 in BAWE and 0.02 in

11 "The Sketch Engine is a corpus query system which allows the user to view word sketches, thesaurally similar words, and 'sketch differences'" (Kilgarriff et al., 2004). Word sketches, the products of the "Word Sketch" tool, are summaries of the grammatical and collocational behavior of a word.

BrAWE) in both corpora too. From these 125 nodes, 36 are used with a similar frequency in both corpora, whereas 89 are used in a statistically different fashion based on the LL ratio. From these, 48 were underused in BrAWE (marked with **), while the remaining 41 were overused (marked with *) when compared to the reference corpus, BAWE. The complete data can be found in Table 2.

Table 2. Raw frequency and normalized values of the 125 nodes in both corpora

According to the table presented in appendix 1, we can observe that there are 2,679 collocates for the 125 nodes in BAWE. One exception is the node *contrast*, that does not have any collocate according to our cut off point. In BrAWE, there are only 1,015 collocates for the same 125 nodes, and there are no collocates for six of the 125 nodes (*contrast*, *exception*, *reference*, *attempt*, *tendency*, and *alternative*). Thus, there is a difference of 1,664 between the total number of collocates in BAWE as compared to BrAWE, showing a low density of conventional combinations in the corpus of Brazilian students.

The 125 nouns portray 287 collocates which show a statistically significant different use when comparing both corpora, being 190 underused and 97 overused, as shown below:

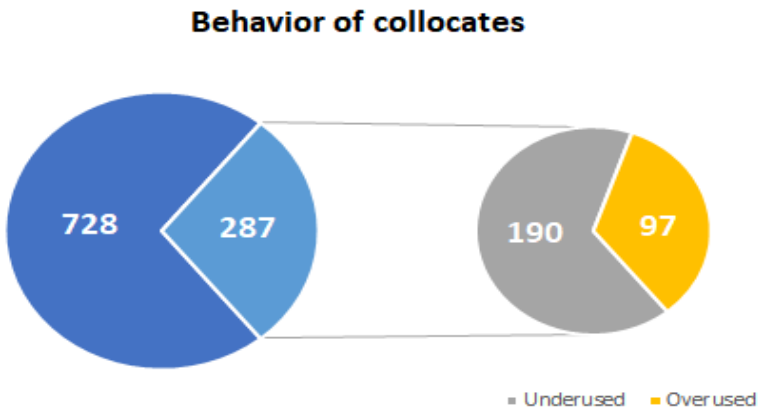

Figure 2. Behavior of collocates

Out of these 287 collocates, 202 are modifiers, 76 are verbs that collocate with nodes in the object position, and the remaining nine are nodes in the subject position. The types of collocates that go along with the 125 nodes in each corpus are displayed in Table 3.

	BAWE	**BrAWE**
Modifiers *whole system, final result*	1,359 (50.7%)	506 (49.8%)
Verb (object) *make process, conduct research*	1,049 (39.1%)	444 (46.7%)
Verb (subject) *result show, strategy include*	271 (10.1%)	65 (6.4%)
TOTAL	***2,679***	***1,015***

Table 3. Types of collocates used in each corpus

The results above account for both the variety and types of collocates of the 125 nodes (nouns) under analysis. Among the three categories, modifiers, i. e. words that occupy a position before the node, account for roughly half of the occurrences in both corpora (50.7% and 49.8% in BAWE and BrAWE respectively). Some examples are *whole system* and *final result*, in which *whole* and *final* are the modifiers and *system* and *result* are the

nodes. Nodes as objects are preferred in BrAWE (46,7%) as compared to BAWE (39.1%), as in *make process* and *conduct research*, with *make* and *conduct* being the verbs when the nodes *process* and *research* are the objects. Conversely, nodes as subjects are more frequent in BAWE (10.1%) than in BrAWE (6.4%) as in *result shows* and *strategy includes*, in which *result* and *strategy* are the subjects and are followed by the verbs *show* and *include* respectively.

When comparing collocations composed of the nodes with statistically significant differences (the overused nodes and the ones composed of the underused nodes), it is possible to observe a balance in terms of syntactic structures in BAWE and in BrAWE, as shown below:

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
Overused nodes	562 (50.4%)	423 (37.9%)	130 (11.6%)	219 (48.2%)	196 (43.1%)	39 (8.6%)
TOTAL	*1115*			*454*		
Underused nodes	468 (51.03%)	355 (38.7%)	94 (10.2%)	147 (49.3%)	134 (44.9%)	17 (5.7%)
TOTAL	*917*			*298*		

Table 4. Syntactic structures of collocations in both corpora

Modifiers that precede the nodes are the most productive ones, with 50.4% and 48.2% of occurrences in BAWE and BrAWE with the overused nodes, and 51.03% and 49.3% in BAWE and BrAWE with the underused nodes. Subsequently, verb + node (object) collocations have the second highest percentage of occurrences, with 37.9% in BAWE with overused nodes and 43.1% in BrAWE with the same nodes. When it comes to the underused nodes, the percentages are 38.7% and 44.9% in BAWE and in BrAWE respectively. Node (subject) + verb collocations account for the lowest percentages with both overused and underused nodes: 11.6% and 10.2% in BAWE, and 8.6% and 5.7% in BrAWE.

When analyzing the LL values of the nodes, there is a bigger difference in the range of LL values of the underused nodes than with the overused

ones. Table 5 illustrates the LL values of the nodes with the most significant differences in the comparison between both corpora. Considering that BAWE is the reference corpus, the terms overused and underused refer to the uses in BrAWE:

	Overused	**Underused**
Lowest LL	*Factor* (7.06)	*Difficulty* (-6.84)
Highest LL	*Example* (370.55)	*Data* (-615.78)

Table 5. Lowest and highest LL

Higher LL values indicate that the differences between the frequency scores are more significant (Rayson, 2002). Table 6 shows collocations with the node *data* (the underused node with the highest LL) in both corpora. Differences can be observed not only in the total number of collocations (55 in BAWE vs. 10 in BrAWE), but also in the syntactic patterns, since 90% of the words that collocate with *data* in BrAWE are verbs, as compared to 63,6% in BAWE.

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
data	20 (36.3%)	23 (41.8%)	12 (21.8%)	1 (10%)	7 (70%)	2 (20%)
TOTAL	55			10		

Table 6. Collocations with the node *data*

While 20 different modifiers[12] collocate with *data* in BAWE, in BrAWE the only modifier is "raw". A possible explanation is that the assignments which compose the BAWE corpus are mostly evidence-based studies, justifying the higher use of *data*. We can also speculate that Brazilian students prefer not to characterize the type of data under analysis by using the word individually rather than as part of a collocation. When it comes

12 *experimental, empirical, quantitative, historical, available, raw, recent, sample, past, primary, following, financial, other, survey, character, relevant, personal, important, actual, old.*

to the verbs that combine with *data*, regardless of whether the node is the object or the subject, the differences continue to be significant. Table 7 demonstrates the different behaviors:

	BAWE	**BrAWE**
Verb (object)	use, collect, obtain, show, analyse, contain, provide, give, record, gather, transmit, compare, present, produce, take, require, store, plot, interpret, send, receive, need, fit	collect, obtain, show, transmit, store, plot, need
Verb (subject)	show, suggest, use, collect, follow, gather, link, seem, demonstrate, support, indicate, exist	show, seem

Table 7. Verbs that collocate with *data*

Among the nodes with statistically significant differences, *difficulty* is the underused node with the lowest LL (-6.84). This means that overall *difficulty* is underused in BrAWE in comparison to BAWE. Table 8 portrays the syntactic structures of the collocations with this node.

	BAWE			**BrAWE**		
	Modifier	**Verb (object)**	**Verb (subject)**	**Modifier**	**Verb (object)**	**Verb (subject)**
difficulty	7 (46.6%)	8 (53.3%)	0	3 (37.5%)	5 (62.5%)	0
TOTAL	15			8		

Table 8. Collocations with the node *difficulty*

In total, there are 15 different collocations in BAWE and eight in BrAWE, with collocates in the modifier and verb (object) categories. While seven different modifiers collocate before the node in BAWE, only three are produced by Brazilians. As for the verbs that accompany the node when it is the object, eight go together with *difficulty* in BAWE whereas five are used in BrAWE, as shown in table 9:

	BAWE	BrAWE
Modifier	Great, technical, financial, main, economic, other	Great, main, other
Verb (object)	Face, cause, encounter, experience, pose, highlight, create	Face, cause, highlight, create

Table 9. Types of collocates with *difficulty*

Conclusion

This corpus-based study aimed to unveil the use of collocations by Brazilians studying in British universities. To that end, a comparative analysis of collocations of the Brazilian Academic Written English Corpus (BrAWE; Goulart, 2017) and the British Academic Written English (BAWE; Alsop & Nesi, 2009) was conducted.

Regarding the first research question *Is there a statistically significant difference in the frequency of the noun nodes and their respective collocates in BAWE and BrAWE?*, it is possible to state that from the 125 nodes analyzed, 36 have a similar frequency in both corpora, 48 were underused and 41 were overused in BrAWE. When it comes to the collocates, the 125 nodes produced 2,679 collocates in BAWE that met our inclusion criteria. In BrAWE, only 1,015 collocates occur with the same 125 nodes. Out of these collocates, 287 came up as having a statistically significant difference in use while analyzing the behavior of the 125 nouns, being 190 underused by Brazilians and 97 overused.

As for the second research question, *Are there differences in syntactic structures of collocations between the two corpora*?, the data revealed that from the 287 collocates which presented significant differences, 202 are modifiers, 76 are verbs in the object position, and nine are verbs in the subject position. In both corpora modifiers account for half of the occurrences (50.7% and 49.8% in BAWE and BrAWE respectively). Nodes as objects are more frequent in BrAWE (46,7%) as compared to BAWE (39.1%), whereas nodes as subjects are more preferred in BAWE (10.1%) than in BrAWE (6.4%). This discrepancy might be related to the type of study conducted by Brazilian students and to how proficient they are to employ different types

of verbs when nodes are used as subjects. For instance, studies conducted by students who wrote texts that compose BrAWE may be of different nature, thus the need to use a verb that best combines with the studies itself (make process, conduct research). On the other hand, when choosing verbs that are used after the node (subject of the sentence), their repertoire is narrower.

Based on the comparison of the two corpora used in this study – BAWE and BrAWE – we noted that academic collocations do not seem to be fully mastered by Brazilian students who write academic texts. For Sinclair (1991), learners operate more on the open choice principle than on the idiom principle, producing fewer collocations or collocations that do not sound natural. This lack of collocational competence was observed in the reduced number of collocations in BrAWE (1,015) when compared to BAWE (2,679) and in the number of outcomes that came up with statistically significant differences in the comparison between the data in the studied corpora. A node that illustrates this phenomenon is data, as displayed in Tables 6 and 7, in which it is possible to observe that the number of collocates used with data is significantly smaller in BrAWE than in BAWE.

The findings of this study suggest that Brazilian students have a limited variety of vocabulary as long as collocations are concerned. It is our belief that proper use of collocations is a major element in academic writing and should, thus, be treated as such in English teaching environments (AlHassan & Wood, 2015; Li & Schmitt, 2009; Martinez & Schmitt, 2012). For instance, the ones which are underused in BrAWE, such as *design* + *system*, *measured* + *value*, *good* + *value*, *decision-making* + *process*, *detailed* + *analysis*, *further* + *analysis*, *empirical* + *data*, and *quantitative* + *data* should be addressed with Brazilian students.

As pointed out by Hyland and Hamp-Lyons (2002: 10), "EAP offers the possibility of making even greater contributions to our understanding of the varied ways language is used in academic communities to provide even more strongly informed foundations for pedagogic materials." Some suggestions are given by Nesselhauf (2005: 253), for whom teaching collocations should begin with making students aware of this phenomenon. AlHassan and Wood (2005) also support the idea that a focus on formulaic

sequences in teaching reveals a development in L2 writing proficiency. Thus, a large repertoire of academic collocations improves students' writing, making it more formulaic and fluent, as formulaic sequences (such as collocations) provide fluency and conventionality to the language.

Considering that more information on the use of collocation by academic English learners would help us to establish a greater degree of accuracy on this matter, a natural progression of this work would be to thoroughly analyze and describe the collocates of all 125 nodes.

Acknowledgements

Marine Laísa Matte would like to thank CAPES for the financial support during her Masters. Simone Sarmento holds a CNPq research productivity scholarship level 1D.

References

Ackermann, K. & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235-247.

AlHassan, L. & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, 17, 51-62.

Alsop, S. & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1), 71-83.

Biber, D. & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Boers, F. & Webb, S. (2018). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, 51(1), 77-89.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), 22-29.

Crossley, S. A., Salsbury, T. & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36(5), 570-590.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, 47(2), 157-177.

Ellis, N. C., Simpson-Vlach, R. I. T. A. & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, 42(3), 375-396.

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (pp. 1-31). Special Volume of the Philological Society. Oxford: Blackwell. [Reprinted as Firth (1968)]Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, 35, 93-104.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P. & Sharma, N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), 23-39.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, 67(S1), 155-179.

Gardner, S. & Nesi, H. (2013) A classification of genre families in university student writing. *Applied Linguistics*, v. 34, n. 1, p. 25-52.

Gardner, D. & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327

Goulart, L. (2017). Compilation of a Brazilian academic written English corpus. Revista e-scrita: *Revista do Curso de Letras da UNIABEU*, 8(2), 32-47.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, analysis, and applications*, 145 - 160.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds.). (2009). I*nternational corpus of learner English (Vol. 2)*. Louvain-la-Neuve: Presses universitaires de Louvain.

Guedes, A. D. S. (2017). *Verbos do inglês acadêmico escrito e suas colocações: um estudo baseado em um corpus de aprendizes brasileiros de inglês*. PhD Thesis. Universidade Federal de Minas Gerais

Hill, J. (1999). Collocational competence. *Readings in Methodology*, 162.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied linguistics,* 19(1), 24-44

Hyland, K. & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for academic purposes*, 1(1), 1-12.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). *The sketch engine. Information Technology*, 105, 116

Kumar, V. P. & Rao, C. S. (2018). A review of reasons for rejection of manuscripts. *Journal for research scholars and professionals of english language teaching*, 8(2), 1-11.

Laufer, B. & Waldman, T. (2011). Verb- noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 647-672.

Lea, M. R. & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in higher education*, 23(2), 157-172.

Li, J. & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, 18(2), 85-102.

Lillis, T. M. (2001). *Student Writing: Regulation, Access, Desire*. London: Routledge.

Lorenz, Gunter (1999). *Adjective Intensification – Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, 33(3), 299-320

Matte, M. L. & Rebechi, R. R. (2018). A quantitative analysis of collocations in Brazilian and British students' academic writing. *Entrepalavras*, 9(2), 195-213

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

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. *Formulaic language*, 2, 375-86.

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis.* London: Bloomsbury Publishing.

Rayson, P. (2002). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD Theses, Lancaster University.

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4), 487-512.

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

Wu, J. (2016). *A Corpus-Based Contrastive Study of Adverb + Verb Collocations in Chinese Learner English and Native Speaker English*. Master degree project. Stockholm University.

Appendix 01

Node	BAWE	BrAWE	0 occurrences in BrAWE	Node	BAWE	BrAWE	0 occurrences in BrAWE
system	48	23	25	example	24	8	16
result	53	31	22	conclusion	8	6	2
value	50	23	27	conflict	7	2	5
figure	15	3	12	standard	25	8	17
process	52	20	32	reference	1	1	0
group	50	16	34	aspect	22	11	11
level	49	14	35	error	15	7	8
model	59	17	42	movement	3	1	2
development	45	12	33	task	20	13	7
data	55	10	45	measure	25	0	25
information	51	21	30	importance	25	12	13
research	41	15	26	support	18	5	13
analysis	34	15	19	feature	23	5	18
rate	55	18	37	discussion	4	1	3
effect	53	22	31	perspective	6	1	5
method	51	19	32	influence	13	6	7
change	55	20	35	requirement	21	8	13
strategy	43	13	30	extent	8	5	3

factor	68	25	43	characteristic	23	3	20
control	31	7	24	interaction	6	2	4
use	45	21	24	author	2	1	1
policy	30	8	22	degree	10	5	5
theory	20	3	17	capacity	12	5	7
approach	32	13	19	understand-ing	13	7	6
structure	26	11	15	concern	15	8	7
role	32	12	20	pattern	17	8	9
quality	29	16	13	reduction	10	5	5
difference	41	18	23	basis	9	4	5
function	28	12	16	definition	11	5	6
activity	37	11	26	procedure	9	5	4
organisation	16	5	11	trend	25	5	20
environ-ment	31	6	25	consideration	12	2	10
resource	26	11	15	observation	5	3	2
type	34	11	23	potential	11	3	8
society	5	2	3	improvement	11	6	5
condition	46	16	30	purpose	7	2	5
production	34	7	27	finding	13	8	5
form	20	4	16	assumption	9	3	6
section	16	5	11	outcome	10	5	5
interest	23	7	16	aim	5	2	3
relationship	35	12	23	presence	6	3	3
source	25	13	12	consequence	9	3	6
impact	30	16	14	explanation	6	4	2
practice	18	5	13	implication	7	0	7
need	46	20	26	variation	9	4	5
growth	23	8	15	category	10	2	8
material	26	11	15	difficulty	14	8	6
period	14	5	9	description	6	3	3
increase	28	11	17	link	8	3	5
review	6	3	3	attempt	1	1	0
term	16	6	10	shift	5	2	3
solution	24	17	7	significance	1	0	1
individual	6	0	6	limitation	2	1	1
concept	18	10	8	proportion	7	5	2
demand	25	9	16	phenomenon	7	5	2
population	26	10	16	recognition	2	1	1
element	24	12	12	contrast	0	0	0

knowledge	23	8	15	contribution	5	3	2
introduc-tion	3	0	3	alternative	4	4	0
benefit	35	15	20	insight	7	5	2
experience	17	6	11	tendency	1	1	0
technique	30	10	20	exception	1	1	0
range	21	9	12				
TOTAL	BAWE		BrAWE		0 occurrences in BrAWE		
	2679		1015		1664		

Appendix 02: Types of collocates for each node

NODE	Modifier	Object	Subject	Modifier	Object	Subject
	BAWE			BrAWE		
system	9	22	17	4	11	8
result	20	23	10	11	14	6
value	19	23	8	7	15	1
figure	9	5	1	0	2	1
process	14	24	14	5	9	6
analysis	12	14	8	4	10	1
group	18	20	12	7	5	4
level	21	25	3	6	8	0
model	14	24	21	4	11	2
development	25	16	4	10	3	0
data	20	23	12	1	7	2
information	24	25	2	10	11	0
research	22	9	10	9	3	3
rate	24	24	7	7	10	1
effect	25	25	3	11	10	1
method	17	23	11	11	6	2
change	24	24	7	10	8	2
strategy	18	19	6	6	6	1
factor	25	25	18	14	8	3
control	14	16	1	3	5	0
use	25	18	2	8	12	1
policy	9	16	5	1	5	2
theory	3	9	8	0	3	0
approach	10	15	7	3	9	
structure	10	15	1	3	8	0
role	18	14	0	8	4	0

quality	12	16	1	6	10	0
difference	24	15	2	9	9	0
function	11	16	1	5	7	0
activity	16	19	3	4	7	0
organisation	5	3	8	1	2	2
environment	23	7	1	4	2	0
resource	16	9	1	9	2	0
type	22	12	0	8	3	0
society	5	0	0	2	0	0
condition	25	17	4	11	4	1
production	22	9	3	4	3	0
form	16	4	0	3	1	0
section	10	1	5	3	1	1
interest	13	10	0	4	3	0
relationship	17	17	1	5	7	0
source	20	4	1	10	2	1
impact	21	8	1	12	4	0
practice	14	4	0	4	1	0
need	25	21	0	10	10	0
growth	13	10	0	5	3	0
material	14	9	3	7	4	0
period	11	2	1	5	0	0
increase	21	6	1	7	4	0
review	4	2	0	2	1	0
term	10	5	1	4	2	0
solution	9	13	3	5	9	3
individual	1	3	2	0	0	0
concept	9	9	0	2	8	0
demand	14	11	0	5	4	0
population	19	5	2	8	2	0
element	18	7	0	8	4	0
knowledge	13	10	0	3	5	0
introduction	3	0	0	0	0	0
benefit	2	19	2	8	6	1
experience	13	4	0	4	2	0
technique	18	10	3	4	5	1
range	14	7	0	6	3	0
example	14	8	2	5	3	0
conclusion	4	4	0	2	4	0
conflict	2	3	2	0	2	0
standard	12	12	1	5	2	1

reference	0	1	0	0	1	0
aspect	15	7	0	10	1	0
error	5	10	0	3	4	0
movement	2	1	0	1	0	0
task	11	8	1	8	4	1
measure	16	7	2	0	0	0
importance	11	14	0	6	6	0
support	12	6	0	1	4	0
feature	16	5	2	3	1	1
discussion	4	0	0	1	0	0
perspective	6	0	0	1	0	0
influence	11	2	0	6	0	0
requirement	15	6	0	5	3	0
extent	6	2	0	5	0	0
characteristic	17	5	1	2	1	0
interaction	4	1	1	2	0	0
author	2	0	0	1	0	0
degree	7	3	0	3	2	0
capacity	8	4	0	5	0	0
understanding	8	5	0	4	3	0
concern	12	3	0	6	2	0
pattern	12	5	0	5	3	0
reduction	4	6	0	2	3	0
basis	6	3	0	3	1	0
definition	8	3	0	2	3	0
procedure	5	3	1	1	3	1
trend	15	7	3	1	4	0
consideration	8	5	0	1	1	0
observation	2	3	0	1	2	0
potential	5	6	0	2	1	0
improvement	5	6	0	2	4	0
purpose	7	1	0	2	0	0
finding	5	4	4	2	3	3
assumption	4	4	1	1	2	0
outcome	5	6	0	4	1	0
aim	4	1	0	1	1	0
presence	2	4	1	0	2	1
consequence	9	0	0	3	0	0
explanation	3	3	0	2	2	0
implication	5	2	0	0	0	0
variation	6	3	0	2	2	0

category	8	1	1	2	0	0
difficulty	7	8	0	3	5	0
description	4	2	0	1	1	0
link	4	4	0	2	1	0
attempt	0	1	0	0	1	0
shift	1	4	0	0	2	0
significance	1	0	0	0	0	0
limitation	2	0	0	1	0	0
proportion	6	1	0	5	0	0
phenomenon	3	4	0	2	3	0
recognition	1	1	0	1	0	0
contrast*	0	0	0	0	0	0
contribution	4	4	0	2	1	0
alternative	2	2	0	2	2	0
insight	2	5	0	1	4	0
tendency	1	0	0	1	0	0
exception	1	0	0	1	0	0
TOTAL	1359 (50.7%)	1049 (39.1%)	271 (10.1%)	506 (49.8%)	444 (46.7%)	65 (6.4%)
	2679			1015		

**contrast* is an academic noun classified in Frankenberg-Garcia et al.'s (2018) study that does not have productivity in BAWE nor in BrAWE.