UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

MARINE LAÍSA MATTE

A CORPUS-BASED STUDY ON THE USE OF ACADEMIC COLLOCATIONS IN ENGLISH BY BRAZILIAN STUDENTS

Porto Alegre
2019

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

MARINE LAÍSA MATTE

A CORPUS-BASED STUDY ON THE USE OF ACADEMIC COLLOCATIONS IN ENGLISH BY BRAZILIAN STUDENTS

Dissertação de Mestrado em Linguística Aplicada apresentada como requisito parcial para a obtenção do título de Mestre em Letras pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul

Orientadora: Profª Drª Simone Sarmento

Porto Alegre
2019

CIP - Catalogação na Publicação

Matte, Marine Laísa
A corpus-based study on the use of academic collocations in English by Brazilian students / Marine Laísa Matte. -- 2019.
130 f.
Orientador: Simone Sarmento.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, BR-RS, 2019.

1. Academic collocations. 2. English for Academic Purposes. 3. Academic writing. 4. Corpus Linguistics. I. Sarmento, Simone, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Marine Laísa Matte

A CORPUS-BASED STUDY ON THE USE OF ACADEMIC COLLOCATIONS IN ENGLISH BY BRAZILIAN STUDENTS

Dissertação submetida ao Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Mestre em Letras.

Porto Alegre, 27 de agosto de 2019.

Resultado:

BANCA EXAMINADORA:

Antonio Paulo Berber Sardinha
Programa de Pós-Graduaçaõ em Linguística Aplicada e Estudos da Linguagem
Pontifícia Universidade Católica de São Paulo (PUCSP)

Cristina Lopes Perna
Programa de Pós-Graduação em Letras
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Rozane Rodrigues Rebechi
Programa de Pós-Graduação em Letras
Universidade Federal do Rio Grande do Sul (UFRGS)

Resumo

Nas últimas décadas, a língua inglesa tem se tornado a língua de produção e disseminação de conhecimento (Altbach & Knight, 2007; Ammon, 2006; Baumvol, 2018), e assim, dominar as convenções de escrita acadêmica em língua inglesa se faz de extrema importância. Dentre elas, destacam-se as colocações, palavras que frequentemente ocorrem juntas devido ao seu grau de atratividade (Durrant, 2011; Hill, 2000; Nesselhauf, 2005; Sinclair, 1991). O uso adequado de colocações pode ser determinante para a qualidade de um texto acadêmico, uma vez que elas conferem fluidez e precisão à escrita (Durrant & Schmitt, 2009). O objetivo deste trabalho é investigar como alunos brasileiros utilizam colocações em inglês acadêmico formadas por um substantivo (nódulo) e seus colocados em uma análise contrastiva com produções acadêmicas escritas por alunos de excelência provenientes de universidades britânicas. Para tanto, foram estudados 125 nódulos, conforme a classificação de Frankenberg-Garcia et al. (2018). Como metodologia, optou-se pela Linguística de Corpus, visto que ela opera com dados de linguagem autêntica e encara a língua como um sistema de probabilidades. Os corpora utilizados para o presente trabalho são o British Academic Written English (BAWE; Alsop & Nesi, 2009) e o Brazilian Academic Written English (BrAWE; Goulart, 2017). A ferramenta Word Sketch do software Sketch Engine foi utilizada para o levantamento e posterior análise das colocações. Recorreu-se à calculadora estatística Log-Likelihood para verificar se as diferenças de usos nos dois corpora são estatisticamente significativas. Os resultados apontam para um subuso dos nódulos no corpus dos brasileiros em relação ao corpus britânico. Além disso, a análise qualitativa revelou diferenças significativas nas escolhas colocacionais, indicando pouca riqueza lexical nos textos produzidos pelos alunos brasileiros. Os resultados obtidos podem fornecer subsídios para a elaboração de material didático, para o ensino de Inglês para Fins Acadêmicos, bem como para discussões mais amplas de produção e disseminação de conhecimento que se dão majoritariamente em língua inglesa.

Palavras-chave: colocações acadêmicas – inglês acadêmico – linguagem formulaica – BAWE – BrAWE

Abstract

Over the last decades, English has become the language of knowledge production and dissemination (Altbach & Knight, 2007; Ammon, 2006; Baumvol, 2018) and thus, mastering academic writing conventions in English is extremely important. Among them, we find collocations, words that are frequently used together due to their attraction (Durrant, 2011; Hill, 2000; Nesselhauf, 2005; Sinclair, 1991). The appropriate use of collocations is indispensable for the quality of academic texts, since they guarantee fluency and accuracy to the text (Durrant & Schmitt, 2009). The objective of this study is to investigate how Brazilian students use collocations in English composed of one noun (node) and its collocates in a contrastive analysis with academic assignments written by British universities' outstanding students. 125 nodes were analyzed, according to Frankenberg-Garcia et al.'s (2018) classification. Corpus Linguistics was chosen as the research methodology since it operates with authentic data and understands the language as a probabilistic system. The corpora used in the study are the British Academic Written English (BAWE; Alsop & Nesi, 2009) and the Brazilian Academic Written English (BrAWE; Goulart, 2017). The Word Sketch tool from the software Sketch Engine was chosen for both collection and analysis of the collocations. The Log-Likelihood calculator was used to verify whether the differences of uses in both corpora are statistically significant. The outcomes show that the nodes are underused by Brazilians in comparison with the British corpus. Furthermore, the qualitative analysis revealed significant differences in the collocational choices, indicating a low lexical density in the texts produced by Brazilian students. The findings of this study can provide resources for material design, for English for Academic Purposes teaching, as well as for broader discussions about practices of knowledge production and dissemination that happen mainly in English.

Keywords: academic collocations – academic English – formulaic language – BAWE - BrAWE

Acknowledgements

One of the most difficult yet delightful parts of this journey is to thank every person that stood by my side from the beginning of this research, back in 2017.

I would like to express my infinite gratitude to my amazing advisor Dr. Simone Sarmento, who inspires me with her knowledge and is a role model for me to become a real Applied Linguist. Her guidance throughout this journey was indispensable and will be of great importance for the next four years. Also, I have to thank her for sharing the passion for Grêmio, and the 14th of July, the best day of the year.

Further appreciation goes to the Postgraduate Program in Languages at the Federal University of Rio Grande do Sul, specially to the professors Anamaria Kurtz de Souza Welp, Juliana Roquele Schoffen, Margarete Schlatter, Pedro de Moraes Garcez and Rozane Rodrigues Rebechi, To the Coordination for the Improvement of Higher Education Personnel (CAPES) for funding my MA studies. And for the committee members for their contributions.

My gratitude to *Simonetes*, especially to Larissa, whose help goes beyond sharing her corpus with me and she is aware of that; to Laura, a great inspiration that I had the pleasure to get to know better last year; to Paula, my internship partner; to Letícia, who shared all the craziness during IF's public selection process; to Lucas, for supporting me and for understanding my astrology birth chart so well. To all my colleagues (Laura, Bruna Álvaro, … and the list is huge!), my gratitude too.

I also have to thank professors Ana Frankenberg-Garcia, Ana Bocorny and Paula Tavares Pinto for the amazing partnership during this year. It was great to get the chance to join the workshops both as a student and as one of the tutors. I have learned a lot with you and

I wish our cooperation continues longer and longer. Finally, I genuinely hope ColloCaid's team reaps the fruits of their intense and inspiring work.

A special thanks to my best friends that share this crazy academic life: Kaiane, Gabrielle ("the sophisticated"), Ellen and Manuela, thank you for every moment of shared knowledge, motivation and kind words. Our support is mutual and this dissertation would not exist without your unconditional help.

I have to thank my cousin and roommate Lívia for understanding my absences and the times I had to lock the door to focus. Finally, I would like to express my gratitude to my parents Beatriz and Carlos, and my sister Marília who truly support me in every academic decision I take. Thank you, thank you, thank you!

List of Figures

List of Tables

List of abbreviations

ACL – Academic Collocation List

AFL – Academic Formulas List

AH – Arts and Humanities

AKL – Academic Keyword List

AVL – Academic Vocabulary List

BAWE – British Academic Written English

BrAWE – Brazilian Academic Written English

CL – Corpus Linguistics

COCA – Corpus of Contemporary American English

EAL – English as an Additional Language

EAP – English for Academic Purposes

ERPP - English for Research and Publication Purposes

ESP – English for Specific Purposes

GE – General English

HE – Higher Education

LL – Log-Likelihood

LS – Life Sciences

LwB – Languages without Borders

MI – Mutual Information

NNS – Non-native speaker

NS – Native speaker

PoS – Part of Speech

PS – Physical Sciences

SKELL – Sketch Engine for Language Learning

SS – Social Sciences

SSD – Statistically significant different

SwB – Science without Borders

FINANCIAL SUPPORT OF CAPES

This study was financed in part by the Coordenação de Aperfeiçoamento de PEssoal de Nível Superiror – Brasil (CAPES) – Finance Code 001

Table of contents

Chapter 1: Introduction

Academic English plays an important role in the Higher Education (HE) scenario due to “the dominance of English for global knowledge production and dissemination.” (Baumvol, 2018, p. 33). Hence, mastering academic writing parameters in English is essential for somebody to be inserted in contexts where it is the preferred language. And one of these parameters are collocations, which in turn, do not have a single definition, as scholars understand them in different ways. Therefore, considering the importance of collocational competence to academic written English, this investigation focuses on the use of collocations by Brazilians studying in British universities. The research questions which guide this investigation are:

1. Is there a statistically significant difference in the frequency of the nodes in BAWE and BrAWE?

2. Is there a statistically significant difference in the frequency of the collocates of these nodes in BAWE and BrAWE? If so, does this difference indicate overuse or underuse? Is it possible to identify the motivations for such differences?

The main motivation for conducting a research of this kind derives from my previous experience as an English for Academic Purposes (EAP) tutor in a Brazilian program, the Language without Borders (LwB), that aimed at providing free EAP classes to the academic communities of public universities. (Abreu-e-Lima et al., 2016; Brasil, 2014). Through this teaching experience, I could observe some difficulties students had when facing the challenge of writing academic English. Some struggles observed were related to not mastering academic general vocabulary, connectors (Matte

& Sarmento, 2018) and combinations of words, i.e., collocations. Hence, it has always been an interest of mine to better understand how Brazilians write academic English in order to be able to help them improve their writing skills.

Considering that writing proper academic English goes beyond knowing isolated words, due to the fact that language is formulaic in nature (Durrant & Schmitt, 2009), mastering collocations is imperative to guarantee fluency in a text. The challenges faced by non-native speakers of English, nonetheless, are enormous and must be tackled in EAP teaching environments (Howarth; Laufer & Waldman, 2011; Lorenz, 1999). In this context, Frankenberg-et al. (2018), aware of the importance of collocations, developed the ColloCaid tool aiming at improving collocational knowledge of students in a text editor program. As the user types the text, suggestions of collocations appear on the screen to be automatically incorporated into the writing pieces. This project, along with my personal motivation described before, and the lack of studies regarding how Brazilians use academic English collocations, made me realize the relevance of conducting the present investigation.

Adopting a Corpus Linguistics approach (CL) (Biber; Conrad & Reppen, 1998; McEnery & Wilson, 1996; McEnery & Hardie, 2011; Berber Sardinha, 2004), two academic corpora will be compared in order to describe the collocations chosen by Brazilian students. Alongside the analysis of the collocations used by Brazilians, this study also seeks to offer solutions for possible struggles in academic writing by providing suggestions of collocations that will hopefully be adopted by the ColloCaid team.

This study is composed of six chapters, the first one being this introduction. In chapter 2, I will present the literature review including notions regarding academic writing, as well as formulaic language with a specific focus on collocations. In chapter 3, I will explain the methodological procedures adopted to conduct the analysis. Then, chapter 4 will be dedicated to the results and discussion of both quantitative and qualitative findings. Finally, some final remarks will be given in chapter 5.

Chapter 2: Literature Review

In this chapter, I will present the literature review that based this study. First of all, I will introduce the importance of academic English in HE, as well as characteristics and challenges of mastering this language. Then, I will provide the reader with what corpus linguistics is, then with what is understood by formulaic language by focusing on collocations. After that, I will discuss some studies that analyzed collocations.

2.1 Academic language: importance, characteristics and challenges of academic English

Internationalization has been one of the indicators of quality development of higher education (HE) in the last few years. Thus, the demands to engage in situations where English is the primary language are also increasing day by day as it is "the lingua franca for scientific communication" (Altbach & Knight, 2007, p. 291). English is the language used for knowledge production and dissemination internationally, after all, "publications in English are widely read and quoted while publications in other languages hardly reach the international sphere, let alone the global arena" (Ammon, 2006, p. 18).

In this scenario, academic English is written by both L1[1] and English as an additional language (EAL) writers, the latter with a clear linguistic disadvantage in this scenario (Flowerdew, 2019). Hyland (2016a, 2016b), however, does not share this same view, insofar as he understands that L1 writers and EAL scholars encounter the same hurdles since "academic English is no one's first language". In response to this controversy, Flowerdew (2019) claims that besides having a restricted vocabulary,

[1] In this case, L1 is used to refer to English as a L1.

collocational competence and a lack of other features of academic language, EAL writers learn them while they are making an additional effort to learn other aspects of the language system, which has been naturally learned by L1 writers. In order to counterbalance this L1 writers' advantage, universities should provide courses in English for Academic Purposes (EAP) and in English for research and publication purposes (ERPP).

Coffin et al. (2003) developed a toolkit to help lecturers and tutors teach academic writing to HE students. For the authors, writing might be one of the most important skills at university level, as

> disciplinary knowledge and understanding are largely exhibited and valued through the medium of writing. Students can begin to understand the significance of writing by becoming aware that writing takes particular conventional forms in different contexts. (Coffin et al., 2003, p. 19)

In addition, writing has a special role in academic contexts because, according to Biber and Gray (2016), it is the first skill that students must master in order to achieve academic success. It is at the university level, also, that academic literacies are being tested all the time, as learning in HE is related to new ways of constructing knowledge. In other words, new ways of knowing and understanding are constantly being discovered (Lea & Street, 1998), and these practices are necessarily intertwined with academic writing. Although academic writing plays an important role in academic contexts, it is usually assumed that students already know the rules or conventions of what is considered academic writing. Therefore, by assuming that this is part of their "common sense" knowledge, these rules or conventions are not part of the curriculum. (Coffin et al, 2003). This 'common sense' knowledge mentioned by the authors above are related to what Lillis (2001) called the “practices of mystery”. According to the author, the members of the academic community must comprehend writing conventions to be able to use them properly. However, these rules are not transparent and must be taught, in the sense that the immersion in settings where academic English is used does not guarantee the mastery of these rules, although it might help. Regarding the “practice of mystery”, Lillis (2001), claims that it “is ideologically inscribed in that it works against those least familiar with the conventions surrounding academic writing, thus limiting their participation in HE as currently configured”. (p. 137). Hence, the necessity of having the teaching of English for Academic Purposes (EAP) as an important element in HE curriculum is paramount.

EAP arouse from the wider area of English for Specific Purposes (ESP), and developed as a field with the expansion of universities around the world and, consequently, with international students using English in their studies. EAP, as a legitimate aspect of English Language Teaching (Hamp-Lyons, 2001), can be understood as “an educational approach and a set of beliefs about TESOL[2] that is unlike that taken in general English courses and textbooks.” (Hamp-Lyons, 2001, p. 126). Based on this argument, general English (GE) and academic English have specific characteristics, according to the purpose they are used for.

Following this idea that academic English has peculiarities that differentiates it from GE or English for Specific Purposes (ESP), Hyland and Hamp-Lyons (2002) point out that EAP refers to the language that fulfills the needs of groups that circulate in academic contexts. Therefore, it is not about learning language by itself, but developing other kinds of literacy that involve specific skills required by academic disciplines.

[2] TESOL stands for Teaching English to Speakers of Other Languages.

Other authors, such as Hyland (2006) and Charles (2013) agree that EAP, as a broad term, covers the uses of English in academic communicative practices, such as pre-tertiary, undergraduate and postgraduate teaching, classroom interactions, research genres, student writing, and administrative practice.

Biber (2006) explains that EAP covers a wide range of registers, both in written and oral language that students must understand in order to succeed in the university. Nonetheless, despite the variety of registers, commonly students are not ready to navigate the genres, because universities do not offer enough linguistic assistance to write academic prose. (Biber, 2006). Having said that English is a demand not ideally attended, it is imperative to conceptualize what is understood as academic English.

According to Scarcella (2003), academic English is a variety of English used in professional books that contains particular linguistic features usually employed in academic disciplines. This definition restricts the understanding of the term, leading us to expand the notion of academic English into academic discourse which, in turn,

> refers to the ways of thinking and using language which exist in the academy. [...] Textbooks, essays, conference presentations, dissertations, lectures and research articles are central to the academic enterprise and are the very stuff of education and knowledge creation (Hyland, 2009, p. 1).

Thus, academic language/discourse/text is different from the type of language used in daily life situations, not only in terms of formality but also in terms of linguistic choices made for the purposes of each communicative situation. Simpson-Vlach and Ellis (2010) agree that the language needed in academic contexts differs from the one appropriate to more basic communicative situations. Therefore, given this difference in proficiency, knowing how to manage academic language is an extra demand upon students.

Based on Scarcella (2003), there is an important linguistic dimension in academic English. This dimension encompasses the four skills (reading, writing, listening and speaking) and includes five components: phonological, lexical, grammatical, sociolinguistic and discourse. Within each of these components, there are specific features we choose whenever we are using English in everyday situations or in academic situations. Table 1 summarizes these characteristics:

Table 1
Components of ordinary English and academic English

	Ordinary English[3]	**Academic English**
Phonological component	Combination of sounds, stress and intonation, graphemes, and spelling	Phonological features, stress, intonation, and sound patterns
Lexical component	Forms and meanings of words used in everyday situations, prefixes, roots, suffixes, parts of speech. Example: find out	Forms and meanings of words used across academic disciplines, prefixes, roots, suffixes, parts of speech. Example: investigate
Grammatical component	Morphemes entailing semantic; syntactic, relation, phonological and distributional properties; simple rules of punctuation	Grammatical features associated with argumentative composition, procedural description, analysis, definition, and procedural description;

[3] Ordinary English is equivalent to general English.

		grammatical co-occurrence restrictions governing words; grammatical metaphor; more complex rules of punctuation
Sociolinguistic component	Production of sentences, frequently occurring functions and genres	Increased number of language functions and genres
Discourse	Basic discourse devices to talk or write informally	Devices such as transitions and organizational signs.

As shown in Table 1, the same components exist in both ordinary and academic English, but a few differences are worth highlighting. The biggest disparity is related to the features of the grammatical component, which are responsible for the formality of academic language. So, according to this table, specific linguistic resources are required in typical writing styles of academic genres, such as argumentative composition, procedural description, analysis, definition, and procedural description. Another difference between academic and ordinary English lies in the fact that the first requires mastery of a wider range of linguistic features. Additionally, while ordinary conversation allows for inaccurate uses of words or phrases, written academic English does not. Therefore, the key word regarding the use of academic English is *mastery*. (Scarcella, 2003)

For many years, researchers have studied the differences between writing and speech. Hughes (1996, p. 33-34) elicits spoken and written features and differentiates them by claiming that in speech we use “simple and short clauses, with little elaborate embedding (particularly within noun phrases)" and "terms that depend on the context of production for their understanding", while in written discourse we prefer "longer and more complex clauses with embedded phrases and clauses, particularly in the form of densely informative noun phrases". The perception that academic writing is more elaborated and more explicit – with longer and complex sentences that must be written because the reader is not face-to-face with the writer -also holds, as discussed in Hyland (2002).

However, Biber et al. (1999) present a different perspective with respect to this issue, and point out that the great majority of finite dependent clauses are more commonly used in spoken mode than in written mode. Biber and Gray (2010), through a corpus-based study, conclude that academic written texts are drastically different from spoken ones. However, the perception mentioned above could not be confirmed as academic writing has developed particular characteristics, as the preference of relying on nominal/phrasal clauses than on clausal structures. (Biber & Gray, 2010).

Further features of academic texts can be found in Biber and Gray (2016, p. 79-82). Some of them are listed below:

1) The three most frequent parts-of-speech in written academic discourse are nouns, adjectives and prepositions;

2) The typical verb categories are copula *be*, passive forms (*be* + *made*/*given*/*taken*/*used*), derived verbs with prefix *re*- and suffix *–ize* (*reabsorb, itemize*), and lexical verbs: activity verbs (*use, produce, provide, apply, form, obtain, reduce*); Communication verbs (*describe, suggest*); Mental verbs (*consider, assume, determine*); Causative / Occurrence / Existence verbs (*follow,*

allow, require, include, involve, contain, exist, indicate, represent); Specific prepositional verbs (*lead to, result in, occur in, depend on, consist of, BE based on, BE associated with, BE related to*);

3) Adverbs and adverbials are usually more common in oral registers. However, there are some specific to written academic register: *often, usually, significantly, more, relatively, especially, particularly, generally, indeed.*

Whether in oral or written texts, it is possible to conclude that academic English has particular elements that differentiates it from other types of English used in various situations. Mastering academic English demands effort due to its challenges and inherent complexities. Hamp-Lyons (2002, p. 1) argues that students "must now gain fluency in the conventions of English language academic discourses to understand their disciplines and to successfully navigate their learning." If this is the current scenario in academic contexts, it is important that these students are well equipped with written academic English tools, mainly because it is not true that academic conventions are universal. In other words, mastering conventions in one language does not necessarily mean appropriate use of conventions in another one.

Considering that academic English is a big challenge for both L1 speakers of English and L2 learners of any language (Wray, 2000), it should have a mandatory space in the curriculum of HE courses. If not taught by professors, how will students be able to turn their weaknesses in academic English into strengths? If the goal is to produce texts that sound natural and close to what is expected from someone who is willing to participate in the academic community, the appropriate use of formulaic language - being collocations one of the elements that guarantee formulaicity to a text -, must be part of the curriculum. (Scarcella, 2003).

The next section discusses corpus linguistics and its main tools to the exploration of language.

2.2 Corpus Linguistics

Due to the fact that Corpus Linguistics (CL) is a research approach, this section aims at discussing CL's characteristics and some facilities to deal with language description.

CL allows for the study of authentic language, that is, language that occurs in real life. (Biber; Conrad & Reppen, 1998; McEnery & Wilson, 1996; McEnery & Hardie, 2011; Berber Sardinha, 2004). CL embraces a probabilistic perspective, meaning that it faces language in terms of how likely language items might occur in a certain context. This approach is different from Chomsky's one who believed frequency and probability were not relevant factors to be taken into account in order to describe a given language, and that whatever mattered in terms of language description should be accessed through native speakers' introspection.

A corpus, the object of study of CL, is understood as "a large, principled collection of naturally occurring texts that is stored in electronic form (accessible on computer)" (Conrad, 2002, p. 76). Additionally, the compilation of a corpus must follow some rules, as it is directly related to the language being depicted. Hence, there are different types of corpora (the plural form of 'corpus'), depending on the type of language someone is trying to represent. For instance, if someone is interested in conducting research on how sports news is reported, the corpus must necessarily contain sports news. If the goal is to study how abstracts are written in Biomedicine, then the researcher will have to deal with a specialized corpus of biomedical abstracts. That is the reason why there is a plethora of corpora, each one representative of a particular portion of language (a specific genre or a specific register, for example).

Besides the variety of corpora in terms of their nature, there is another typology in terms of format: monolingual, parallel and comparable. Monolingual corpora contain texts written in one language, i.e. in a corpus of Brazilian soap operas, the texts will necessarily be in Portuguese. Parallel corpora have the texts in a certain L1 aligned to the translation into L2, as it is the case of COMPARA[4]. Comparable corpora are created with the same criteria of text selection, that is, texts in both corpora are from the same area of expertise, have similar size, are written in the same textual genre; among other purposes, comparable corpora are useful to determine equivalents from one language to another.

Additionally, lemmatized and tagged corpora are two modalities. Lemmatization assigns the base form of a word (lemma) in a corpus with a tool called lemmatizer. Thus, in a corpus-based analysis, if the lemma is searched, all the derived forms of this word come up as a result. For instance, searching for *make*, will also find *makes*, *making*, or *made*. Tagged corpus, on the other hand, is a corpus whose words have been annotated syntactically, morphologically or semantically, among others. The most common way of tagging is according to the part-of-speech[5] (PoS tagger)

Biber, Conrad and Reppen (1998) distinguish between the studies of language in two areas: studies of structure (what they call the traditional way of studying language) and studies of use. CL fits in the latter, as it investigates "how speakers and writers

[4] "https://www.linguateca.pt/COMPARA/

[5] Part-of-speech (PoS) is a category associated to the word according to its syntactic function, i.e. a particular grammatical class of word (noun, pronoun, adjective, verb, adverb, preposition etc)

exploit the resources of their language. Rather than looking at what is theoretically possible in a language, [CL] studies the actual language used in naturally occurring texts." (p. 1). Moreover, the authors characterize corpus-based analyses by listing the four characteristics below:

- it is empirical, analyzing the actual patterns of use in natural texts;

- it utilizes a large and principled collection of natural texts (corpus), as the bases for analysis;

- it makes extensive use of computer for analysis, using both automatic and interactive techniques;

- it depends on both quantitative and qualitative analytical techniques. (Biber; Conrad & Reppen, 1998, p. 4)

Thus, as it is possible to observe, if corpus-based studies are empirical in the sense that patterns of uses of natural occurring texts are the object of study, this moves the study of language away from ideas of what is correct, towards what is typical or frequent (Sinclair, 1991, p. 17). Therefore, through a CL methodology, the researcher has access to what is actually being produced – in written or oral texts – and is able to observe recurrent patterns of language.

In order to analyze these patterns and describe the language, it is possible to use the free access corpora available online[6] or to compile a new one. In case of compiling a corpus to further explore it, a software is needed. The most common offline software

[6] The Corpus of Contemporary American English (COCA) (https://www.english-corpora.org/coca)) and the British National Corpus (BNC) (https://www.english-corpora.org/bnc/) have free online access.

tools are the AntConc[7] and WordSmith Tools[8] that require a quick download on the computer. These corpus analysis toolkits present basically the same functions, such as concordance, in which a word is analyzed within the context of use; wordlists that organize the words according to frequency or according to the PoS[9]; keywords extraction that determines the typical words of the corpus being analyzed in comparison to a reference corpus; and collocations, in which it is possible to observe combinations of words.

Besides AntConc and WordSmith Tools, Sketch Engine[10] is another online tool to explore language. It contains the same functions as the two mentioned above, but there is an extra tool, the Word Sketch, that is especially valuable for the analysis of collocations. It is a "one-page automatic, corpus-based summary of a word's grammatical and collocational behavior" (Kilgariff et al., 2004, p. 105). Therefore, Word Sketch shows collocates organized according to syntactic criteria of different types. Alongside the collocates of the specific word being searched, the collocational strength is provided with the Mutual Information (MI) score. This score indicates how strong the link between two items is. The higher the MI score, the stronger the relation between the items (Church & Hanks, 1990). Based on the facilities of Word Sketch to the exploration of collocations, Sketch Engine was chosen as the software for the analysis of the current corpora.

[7] https://www.laurenceanthony.net/software/antconc/

[8] https://www.lexically.net/wordsmith/

[9] Corpora uploaded in AntConc and WordSmith Tools are not tagged. Thus, only Sketch Engine allows for the wordlists organized according to the PoS.

[10] One of the negative aspects of Sketch Engine is that it is free for only 30 days for each user account.

2.3 Formulaic language: the case of collocations

Sinclair (1991) developed the notion that language operates according to two principles: the open-choice principle and the idiom principle. The first one, which is the foundation of the construction of (nearly) all grammars, considers language production, i.e. text, as the result of complex choices to complete each unit (word, phrase, clause) that composes a text. Therefore, every time a unit is completed, there are many possibilities (the reason why this principle is called open-choice) to be chosen and the only restriction has to do with grammar limitations. According to this principle, any slot of text can be filled with any word. The idiom principle, on the other hand, indicates 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, p. 110). On the assumption that "language is largely formulaic in nature, and that the competent use of formulaic sequences is an important part of fluent and natural language use" (Durrant & Schmitt, 2009, p. 157), this section aims at presenting definitions of collocations as well as previous studies regarding the use of collocations in written academic English.

2.3.1. What are collocations?

"You shall know a word by the company it keeps" might be the sentence that anyone acquainted with collocational studies immediately remembers. This sentence was formulated by J. Firth (1957, p. 11) and has inspired all research in the field, as it summarizes the core meaning of collocations. Although there is a plethora of criteria[11]

[11] Besides focusing only on the frequency of the words, syntactic and semantic criteria could be taken into consideration, as not every high frequency co-occurrences of words could be interpreted as collocations.

to define the extent to which two or more words can be considered a collocation, the statistical perspective will be adopted in this dissertation, i.e, the likelihood of two or more words occurring together (Sinclair, 1991).

Three different types of collocations will be investigated: *modifier* + *noun*, *noun (subject)* + *verb,* and *verb* + *noun (object).* Collocations can range from word level (1) to sentence level (2). Shimohata et al. (1997, p. 476) state that (1) refers to an "uninterrupted collocation which consists of a sequence of words" while (2) has to do with "an interrupted collocation which consists of words containing one or several gaps filled in by substitutable words or phrases which belong to the same category." Nevertheless, these two types share the same features, such as "collocations are recurrent; collocations consist of one or several lexical units; and order of units is rigid in a collocation." (Shimohata et. al., 1997, p. 476).

Wray (2000) subsumed collocations under an overarching term, 'formulaic sequence', which is a sequence of words that seems to be prefabricated, and are, thus, restored as a whole from our memory. Therefore, instead of retrieving isolated words, whenever needed, we recover strings of items that have a particular meaning. Nesselhauf (2005) is also aware of the variety of terms to name collocations and define them as being composed of two or more words with a lexical and/or syntactic fixity to a certain degree.

Hyland (2006) refers to collocations as the occurrence of two or more words in a text, meaning that words do not happen independently. On the contrary, words collocate with each other and their meaning is conveyed by association (CHOI, 2016). Furthermore, for Durrant (2011) and Hill (2000), one of the possibilities of defining collocations has to do with frequency. The latter associates with the frequency-based approach by arguing that collocations are multi-word combinations that make up a significant part of a text.

Based on these definitions of collocations, our own understanding of this linguistic element that provide fluency and accuracy to the language is:

> a combination of two words that are associated due to statistical probabilities of occurring together

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

The next section aims at discussing previous studies that analyzed collocations produced by learners of English.

2.3.2 Previous studies on collocations

Throughout the years, scholars have been investigating collocations mainly through a corpus linguistics perspective. As this dissertation is related to the use of collocations in a learner corpus, studies related to the use of collocations by nonnative speakers will be reviewed. Overall, learners of English do use formulaic language, but they tend to choose some expressions in detriment of others (De Cock et al., 1998; Foster 2001; Granger 1998; Lorenz 1999; Nesselhauf, 2005). Moreover, the choice for using idiosyncratic collocations can draw the attention away from the message being conveyed (Cowie & Howarth 1996; Nesselhauf, 2005).

The comparison between native (NS) and non-native (NNS)[12] collocational performance is presented in Howarth (1998), who conducted both quantitative and qualitative research. The author analyzed adult learners of English writing academically in Social Sciences postgraduate courses and focused on the use of collocations composed of *verb + noun*. The conclusions show that when learners try to vary their writing, even though the collocations are grammatically appropriate, they produce uncommon ones that sound unfamiliar to the proficient reader. Thus, their competence for producing collocations is usually ruled by some strategies, such as L1[13] transfer into L2 and “repeated use of a limited number of known collocations” (Howarth, 1998, p. 41). Moreover, 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, p. 36). Thus, the research points to the importance of mastering collocations, as “native speaker linguistic competence has a large and significant phraseological component […]” (ibid, p. 29).

Granger (1998) analyzed 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, her “initial hypothesis was that learners would make less use of prefabs, or conventionalised language, in their writing than their native speaker counterparts, given that the use of such language is universally presented as typically native-like” (Granger,

[12] In this study, NS stands for native speakers of English, whereas NNS refers to the speakers whose first language is not English.

[13] In this case, L1 is not the same as English as L1. Here, L1 means the learners’ first language.

1998, p. 146). As for the results of the study, 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 recall them and choose to translate from French into English. Additionally, some amplifiers are used exclusively by native speakers.

Collocations composed of *adjective + noun* or *noun + noun* were analyzed by Durrant and Schmitt (2009). The authors analyzed a total of 96 texts organized in two big sets of texts: one containing NNS texts and the other comparable set with NS texts. Both sets have a second organization in which there are long (research assignments and projects) and short texts (essays). By classifying collocations into low-frequency and high-frequency and establishing collocational strength with *t-score* and Mutual Information measures, 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 scores in comparison with native norms. (Durrant & Schmitt, 2009, p. 174). These findings suggest that learners have a tendency to repeat favored items, as they quickly pick up frequent collocations because the less common and strongly associated items take longer to acquire (Durrant & Schmitt, 2009, p. 175). Simpson-Vlach, Ellis and Maynard (2008) reinforce this idea that NS use a wider range of collocations, as NNS tend to use collocations they encounter more frequently (with a lower MI score). The issue of overusing collocations is discussed by Ackerman and Chen (2013, when they 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.” (p.3).

Laufer and Waldman (2011) investigated *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 the scope of the leaner corpus, as it was composed of three *subcorpora* according to three levels of proficiency. Thus, the comparison was between each level of proficiency with the NSs and among the three levels of L2. The learner corpus contained 759 texts. Results indicated that the NS produced almost twice as many collocations as the learners. Learners underused *verb + noun* collocations when compared to NS texts, at the same time that they produced significantly more deviant collocations. Interestingly, advanced and intermediate learners 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. Once again, Wu (2016) shows that there are significant differences in terms of collocations used 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 collocation might be related to the fact that the teaching of collocation is not a focus in China, and that Mandarin and English have only few similarities.

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 + adverb-ly* collocations. The author used AntConc to compare the frequencies and the uses of the collocations within different semantic domains in a learner corpus and in the British Academic Written English (BAWE[14]). Guedes found that the most common verbs used by the learners are action verbs. Moreover, 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. However, the study could not measure the statistically collocational strength in *verb +adverb-ly* because of their low frequency.

Regarding errors in collocation production, it is possible to affirm that they are usually interlingual (Gitsaki, 1999; Laufer & Waldmann, 2011; Selistre, 2010), suggesting that L1 influences L2, as learners try to produce L2 collocations based on the meaning of the sequences of words that convey the same message in their mother tongue. In the case of Portuguese language transfer, Selistre (2010) analyzes adjective, verb and noun transfer, and points out that sometimes L1 transfer might benefit or hinder L2 production, such as in the preference for using *common person* rather than *ordinary person,* because in Portuguese *pessoa comum* is a collocation closer in form to *common person*. *Food intoxication* rather than *food poisoning* is the chosen collocation due to *intoxicação alimentar* in Portuguese. An influence in a collocation with a verb as

[14] This academic corpus contains academic written assignments from students of four British universities: Oxford, Brookes, Reading and Warwick. It is divided into four big areas of expertise: Physical Sciences, Life Sciences, Social Sciences, and Arts and Humanities. In the next chapter, further explanations will be given.

a node can be seen in *make an order* instead of *place an order* due to the influence of *fazer um pedido,* frequently produced in Portuguese.

In a recent study with Brazilian students learning how to write English, Orenha-Ottaiano (2015, p. 837)[15] observed that the difficulties in producing some collocations might occur "if he or she has not learned it explicitly or has not observed the usage before or if opportunities to maximize the learning in an implicit way were not created." Therefore, it is within a teaching environment that these difficulties can be remedied. The author, based on her study that showed how Brazilian students struggle with collocations, created the *Online English Collocations Workbook*, a tool that complements any task focused on the appropriate use of collocations.

Matte and Rebechi (2019) analyzed the quantitative differences in the use of collocations of the Academic Collocation list (ACL) (Ackermann & Chen, 2013) in two academic corpora[16]: the BAWE and the Brazilian Academic Written English (BrAWE) (Goulart, 2017). Surprisingly, only few collocations came up as statistically significant. Furthermore, the most frequent collocations in both corpora are not exactly the same presented in the list, which suggest a possible mismatch between the prescription and authentic language.

There is a proliferation of research that focuses on the importance of learning and teaching since the late 1990s (Boers & Webb, 2018). Moreover, there are ready-

[15] This quotation was translated by the author. Original text: "caso não a tenha aprendido de modo explícito ou não haja observado seu uso antes, ou se não tiver sido criadas oportunidades para maximizar oportunidades de aprendizado de modo implícito." (Orenha-Ottaiano, 2015, p. 837)

[16] Considering that these corpora are under analysis in the present study, they will be further described in the methodology chapter.

made lists containing important collocations and formulas to be mastered, as 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), we will not reach the ideal scenario without pedagogic intervention that fit students’ needs.

Having said that, more than memorizing vocabulary and collocation lists, it is imperative to master collocations in terms of knowing their appropriate use; considering that words mean together with other words, collocational competence must be acquired in context. This argument is sustained by Frankenberg-Garcia (2018) who points out that “the lexical knowledge is not just about understanding words, but also about employing words in context.” (p. 101). This way, Frankenberg-Garcia et al. (2018) started the project called ColloCaid[17], which aims at providing support on academic English collocations. Developed by the Universities of Surrey, Bangor and Poznan, ColloCaid works as a text editor that suggests common academic collocation while the user is typing the node word of a specific collocation. It is a user-friendly tool as the users, when facing questions regarding certain words, do not need to leave the text editor and open new tabs on their computers. Therefore, users are not distracted[18] during their writing processes.

[17] The official website can be accessed at http://www.collocaid.uk/.

[18] There are a number of resources and tools that help EAP users of language with collocations, such as Longman Collocations Dictionary and Thesaurus (Mayor, 2013), the Louvain EAP Dictionary (Granger & Paquot, 2015), and Sketch Engine for English Language Learning, or SkELL (Baisa & Suchomel, 2014). However, in order to use them, EAP users have to stop their line of thought, leave their text editor, and try to solve their doubts with the help of one of these resources. ColloCaid, on the other hand, does not require this back and forth movement.

Finally, it is worth reinforcing the importance of mastering not only the whole academic English register, but collocations in particular, as they are indispensable for providing accuracy and fluency to the texts and, consequently, for making them more readable.

In the next chapter, the methodological procedures employed in this study will be delineated.

Chapter 3: Methodology

This chapter covers the methodology applied in this study: the two corpora, BAWE and BrAWE will be explained, followed by the presentation of the methodological procedures adopted in the analysis.

3.1. BAWE and BrAWE: the two academic corpora at stake

As already mentioned, this study focuses on the identification of overused and underused academic collocations produced by Brazilian university students through the comparison of two different corpora, the British Academic Written English corpus (BAWE) and the Brazilian Academic Written English corpus (BrAWE). Both corpora contain non-published texts written by university students who need to be contacted to make their texts available. Hence, as they are not published and require authors' contact and authorization, the kind of language produced in this type of corpora tends to be understudied, pointing to the relevance of this research. In this section, both corpora will be described.

3.1.1 BAWE corpus

The BAWE corpus (Alsop & Nesi, 2009) was compiled with the objective of gathering written assignments from students of multiple nationalities studying[19] 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, i.e. International Corpus of Learner English (Granger et al., 2002) and the Louvain Corpus of Native English Essays (Granger et al. n.d.), BAWE is composed of non-discipline-specific

[19] BAWE contains texts of undergraduate and master's students.

learner texts. Despite containing student writing, this corpus is different from those compiled with essays written under examination conditions for analyzing non-native-speaker error and language acquisition, as it contains assignments written for undergraduate and master disciplines. BAWE was, thus, designed to enable the investigation of academic literacy and disciplinary knowledge development.

BAWE has 6,968,089 words and it is balanced according to four areas of expertise[20]: Life Sciences (LS), Social Sciences (SS), Physical Sciences (PS), and Arts and Humanities (AH). Each area encompasses a variety of disciplines. Table 2 presents the number of students, assignments, texts and words for each areas of expertise. The numbers 1 to 4 in the first line refer to the year of study when the assignment was written by the student.[21].

Table 2
Number of texts and words by academic genre family in each area of expertise in BAWE

Disciplinary group		*1*	*2*	*3*	*4*	*Total*
Arts and Humanities	students	101	83	61	23	268
	assignments	239	228	160	78	705
	texts	259	231	161	83	734
	words	468,353	583,617	427,942	234,206	1,714,118
Life Sciences	students	74	71	42	46	233
	assignments	180	193	113	197	683
	texts	191	208	119	203	721
	words	299,370	408,070	263,668	441,283	1,412,391
Physical Sciences	students	73	60	56	36	225
	assignments	181	149	156	110	596
	texts	186	156	169	129	640
	words	300,989	314,331	426,431	339,605	1,381,356

[20] Alsop and Nesi (2009) refer to these areas as disciplinary groups.

[21] 1 = first year undergraduate; 2 = second year undergraduate; 3 = third year undergraduate; 4 = Masters level

	students	85	88	75	62	313[1]
Social	assignments	207	197	162	202	777[2]
Sciences	texts	218	202	169	204	802[3]
	words	371,473	475,668	440,674	688,921	1,999,130[4]
Total students		333	302	234	167	1039
Total assignments		807	767	591	587	2761
Total texts		854	797	618	619	2897
Total words		1,440,185	1,781,686	1,558,715	1,704,015	6,506,995

[1] Includes 3 of unknown level.
[2] Includes 9 of unknown level.
[3] Includes 9 of unknown level.
[4] Includes 22,394 in texts of unknown level.

Moreover, the corpus is organized according to 13 different academic genre families proposed by Gardner and Nesi (2013). Table 3 provides the number of academic genre families produced in each area of expertise:

Table 3
Distribution of academic genre families by areas of expertise (Heuboeck, Holmes & Nesi, 2010)

	AH	**LS**	**PS**	**SS**	**Total**
Case study	0	91	37	66	194
Critique	48	84	76	114	322
Design specification	1	2	87	3	93
Empathy writing	4	19	9	3	35
Essay	602	127	65	444	1238
Exercise	14	33	49	18	114
Explanation	9	117	65	23	214
Literature review	7	14	4	10	35
Methodology recount	18	158	170	16	362
Narrative recount	10	25	21	19	75
Problem question	0	2	6	32	40
Proposal	2	26	19	29	76
Research report	9	22	16	14	61
Total	724	720	624	791	2859

It should be noted that the number of texts is roughly balanced in each area of expertise.

3.1.2 BrAWE corpus

The Brazilian version of BAWE is the Brazilian Academic Written English corpus compiled by Goulart (2017). With the intention of designing a corpus that was comparable to BAWE, Goulart (2017) organized the corpus similarly to the British one in terms of covering the same areas of expertise and gathering assignments produced by undergraduate students. Therefore, BrAWE also follows Gardner and Nesi's (2013) classification of academic genre families into 12 categories[22]. The final version of the corpus accounts for the assignments of students from 59 universities. The number of universities is high due to the fact that the students were mainly part of the Sciences without Borders (SwB) program, which partnered with more than 80 universities only in the United Kingdom. SwB was a Brazilian scientific mobility program created in 2011 with the objective of strengthening and expanding the internationalization of Brazilian higher education through the provision of scholarships for both students and researchers. Overall, engineering, natural sciences and health sciences were the areas covered by the SwB. Thus, areas such as arts and humanities were not contemplated by the program, but some texts from this area were included in the corpus, because some students from other mobility programs sent their texts as well. Despite being comparable to BAWE, the corpus is unbalanced in terms of subcorpora, since SwB does not cover AH area. Considering that LS, SS and PS are the most representative areas in BrAWE, a subcorpus of BAWE was created in order to make the comparison with BrAWE corpus. Nevertheless, whenever BAWE is mentioned, we are referring to BAWE's subcorpus that contained only assignments in the fields of LS, SS, and PS. Table 4 contains BrAWE's data:

[22] Different from BAWE, empathy writing has zero texts in BrAWE.

Table 4
Number of texts and words by academic genre family in each area of expertise in BrAWE

	AH		SS		LS		PS		TOTAL	
	Texts	Words	Texts	Words	Texts	Words	Texts	Words	Texts	Words
Case study	-	0	5	15,326	9	20,908	18	41,866	32	78,100
Critique	-	-	7	15,341	16	25,782	19	36,053	42	77,176
Design specification	-	-	-	-	-	-	18	36,093	18	36,093
Empathy writing			-		-	-	-	-	-	-
Essay	4	7,887	13		46	82,975	31	48,041	94	160,169
Exercise	-		1		7	6,829	28	35,236	36	43,659
Explanation	-		7		11	14,976	29	54,266	47	80,613
Literature review	-		-		5	11,923	1	3,418	6	15,341
Methodology recount	-		-		19	24,790	31	41,593	50	66,383
Narrative recount	-		-		1	1,457	3	2,375	4	3,832
Problem question	-		2		3	4,506	3	3,602	8	11,277
Proposal	-		-		2	4,078	12	17,554	14	21,632
Research report	-		-		11	26,955	18	49,084	29	76,039
TOTAL	**4**	**7,887**	**35**	**67,907**	**130**	**224,979**	**211**	**369,541**	**380**	**670,314**[23]

[23] The mean number of words per text in each area of expertise is: 1.971,75 (AH); 1,940.2 (SS); 1,730.6 (LS); 1,751.3 (PS).

Below, Table 5 summarizes the main differences regarding BAWE and BrAWE corpora:

Table 5
BAWE and BrAWE corpora

	BAWE	BrAWE
Words	3,312,196[24]	768,323[25]
Number of assignments	2,761[26]	380
Quality of assignments	Merit and distinction	Passing

Alongside the differences in size, the quality of assignments also distinguishes BAWE and BrAWE. While in the first corpus students were graded merit and distinction, in the second, students received at least passing grades, which does not necessarily mean that no one wrote excellence texts. Therefore, due to the quality of texts, BAWE is an adequate reference corpus to fulfill the purposes of a contrastive analysis.

3.2 Methodological procedures

This study aims at shedding light on the importance of mastering collocations in academic writing in order to provide fluency to the text. Therefore, the goal of this corpus-based analysis is to compare academic collocations in two corpora – BAWE and BrAWE –to

[24] This is the total number of words of the subcorpus containing only assignments of LS, SS, and PS. Also, this subcorpus is representative of assignments written exclusively by authors whose first language is English. Thus, 3,312,196 is not the sum of the words from LS, SS, and PS in Table 2.

[25] Due to the fact that Sketch Engine is used for the analysis, the size of BrAWE is 768,323 – rather than 670,314 as shown in Table 3 because this software considers punctuation marks as words.

[26] This number is different from the one in Table 2 probably because some assignments were not categorized into one of the four areas of expertise.

determine whether Brazilian students overuse or underuse collocations when writing academic English. The research questions are:

1. Is there a statistically significant difference in the frequency of the nodes in BAWE and BrAWE?

2. Is there a statistically significant difference in the frequency of the collocates of these nodes in BAWE and BrAWE? If so, does this difference indicate overuse or underuse? Is it possible to identify the motivations for such differences?

In order to have these questions answered, the definition of collocations is the starting point:

> a combination of two words[27] that are associated due to statistical probabilities of occurring together

As already mentioned, node is the word being analyzed, and collocate is the word combining to the node (Sinclair, 1991). The collocations analyzed are of three different types, i.e., the nodes of the collocations are accompanied by three kinds of collocates as shown below:

- ✓ 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*

[27] The combination can have up to three words in between. For instance, in the sentence "Mott MacDonald presented the most stable values", the collocation is "present + value" and there are three other words separating the node *value* and the collocate *present*.

These categories of collocates follow Frankenberg-Garcia's et al (2018), in which the authors came up with a list of 187 collocational nodes. These nodes derived from the amalgamation of three lists: the words in the Academic Vocabulary List (AVL, Gardner and Davies, 2014) from the BAWE corpus, the Academic Keyword List[28] (AKL, Paquot, 2010), and the node words of the Academic Collocations List[29] (ACL, Ackermann & Chen, 2013). AVL comprises 3,000 top lemmas[30] occurring in all academic domains of the COCA corpus[31]; AKL contains 930 keywords extracted from two EAP corpora and a corpus of British and American student writings using a fiction corpus as a reference for contrastive purposes. ACL, which is different from the previous lists, presents collocation units instead of lemmas.

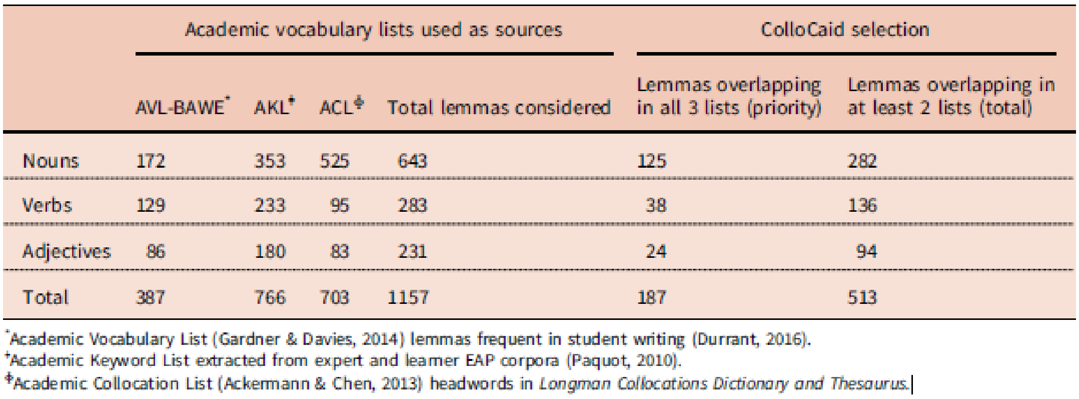
Thus, after gathering the nodes from these three different sources, Frankenberg-Garcia et al. (2018) came up with a total number of 187 nodes overlapping in the three lists, in which 125 are nouns, 38 are verbs, and the remaining 24 are adjectives. The authors also counted the nodes that overlap in at least two lists, with 513 as the total number of nodes, from which 282 are nouns, 136 are verbs, and 94 are adjectives, as can be seen in the figure below:

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

[29] https://www.eapfoundation.com/vocab/academic/acl/

[30] Lemma is the base form of a word, i.e. the lemma covers all the inflections of a word. For instance, *taking*, *takes*, *took*, *taken* are under the same lemma *take.*

[31] https://www.english-corpora.org/coca/. The COCA corpus is a large, genre-balanced corpus of American English. It contains spoken, fiction, popular magazines, newspapers, and academic texts, all equally divided.



	Academic vocabulary lists used as sources				ColloCaid selection	
	AVL-BAWE*	AKL†	ACL‡	Total lemmas considered	Lemmas overlapping in all 3 lists (priority)	Lemmas overlapping in at least 2 lists (total)
Nouns	172	353	525	643	125	282
Verbs	129	233	95	283	38	136
Adjectives	86	180	83	231	24	94
Total	387	766	703	1157	187	513

*Academic Vocabulary List (Gardner & Davies, 2014) lemmas frequent in student writing (Durrant, 2016).
†Academic Keyword List extracted from expert and learner EAP corpora (Paquot, 2010).
‡Academic Collocation List (Ackermann & Chen, 2013) headwords in *Longman Collocations Dictionary and Thesaurus.*

Figure 1. Collocation node selection in ColloCaid[32] (Frankenberg-Garcia et al., 2018)

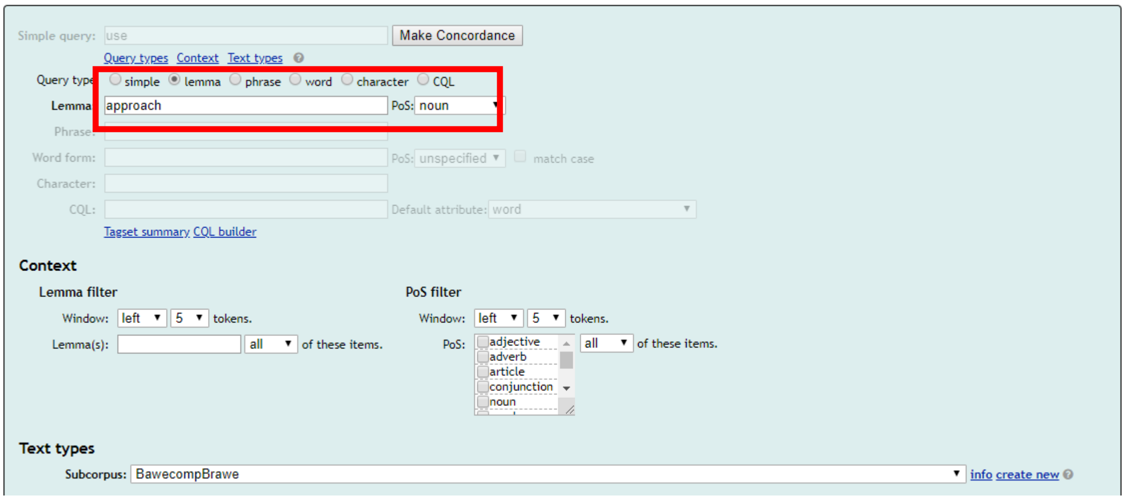
This study focuses on the 125 nouns that overlap in all three lists. For this, both BAWE and BrAWE were analyzed. The software Sketch Engine was chosen because it contains the Word Sketch tool which is particularly useful for the analysis of collocations (Kilgariff et al., 2004), as explained in the Literature Review. Besides the five main steps presented below, some other procedures were taken. The cut-off point was that the collocate must appear in at least two different areas and with a minimum frequency of four occurrences. Thus, collocations of specific areas were excluded from the analysis, as it is the case of *health need*, a collocation that only appears in LS assignments.

1st: the 125 nouns from the Frankenberg et al. (2018) list were ordered from the most to the least frequent in BAWE, as it is a much larger corpus than BrAWE, increasing the likelihood of having more variety of academic words. This ranking was determined based on the frequency of each node by using the "search" tool in Sketch Engine[33]. In Figure 2, it is possible to observe that the node was typed in the "lemma" box and the PoS – noun was

[32] The explanation on ColloCaid is given in the Literature Review chapter.

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

selected.[34] The choice for searching for lemmas – the base form of a word - is justified because when a lemma is searched all the words that derive from this base form come up as a result, i.e in the case of *approach*, the plural form – approaches – comes as a result as well.



Simple query: use

Query type: ○ simple ● lemma ○ phrase ○ word ○ character ○ CQL

Lemma: approach PoS: noun

Word form: PoS: unspecified □ match case

CQL: Default attribute: word

Context

Lemma filter

Window: left 5 tokens.

Lemma(s): all of these items.

PoS filter

Window: left 5 tokens.

PoS: □ adjective □ adverb □ article □ conjunction □ noun all of these items.

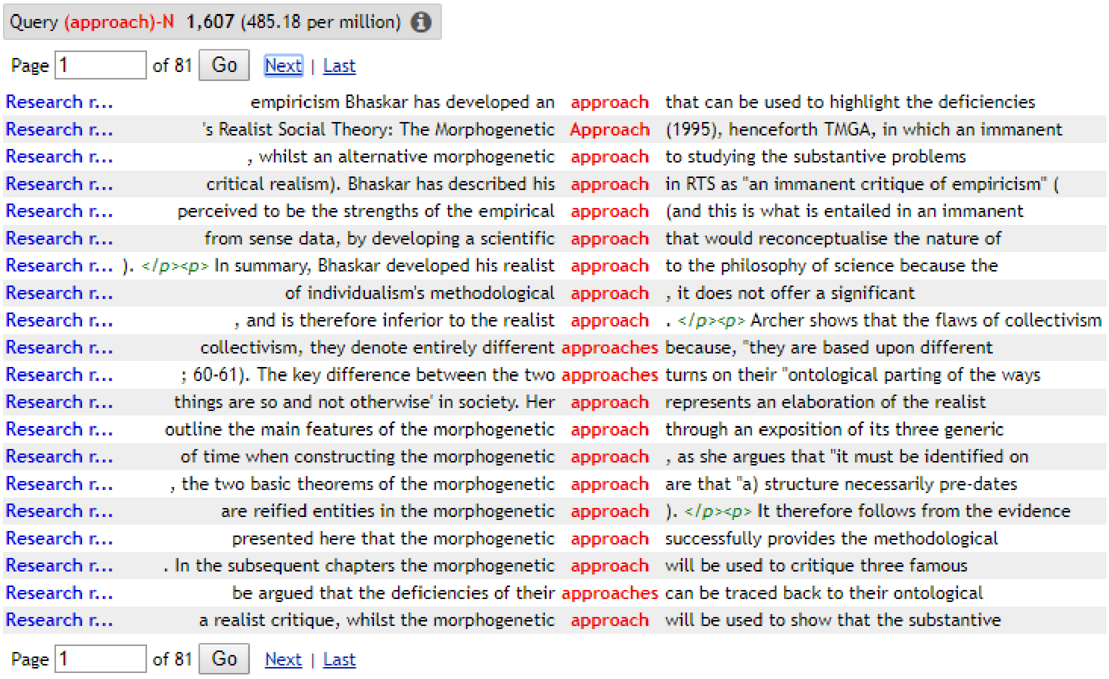
Text types

Subcorpus: BawecompBrawe

Figure 2. Search screen for the node *approach*

After that, the results came up as shown in Figure 3.

[34] Another possible search tool would be to upload a list containing all 125 nouns as a whitelist, which in turn would allow for the results regarding only the words of that list. However, as the whitelist does not contain the option of selecting the PoS, individual searches as shown in Figure 2 were conducted.



Query (approach)-N 1,607 (485.18 per million)

Page 1 of 81 Go Next | Last

Research r...	empiricism Bhaskar has developed an	approach	that can be used to highlight the deficiencies
Research r...	's Realist Social Theory: The Morphogenetic	Approach	(1995), henceforth TMGA, in which an immanent
Research r...	, whilst an alternative morphogenetic	approach	to studying the substantive problems
Research r...	critical realism). Bhaskar has described his	approach	in RTS as "an immanent critique of empiricism" (
Research r...	perceived to be the strengths of the empirical	approach	(and this is what is entailed in an immanent
Research r...	from sense data, by developing a scientific	approach	that would reconceptualise the nature of
Research r...	). </p><p> In summary, Bhaskar developed his realist	approach	to the philosophy of science because the
Research r...	of individualism's methodological	approach	, it does not offer a significant
Research r...	, and is therefore inferior to the realist	approach	. </p><p> Archer shows that the flaws of collectivism
Research r...	collectivism, they denote entirely different	approaches	because, "they are based upon different
Research r...	; 60-61). The key difference between the two	approaches	turns on their "ontological parting of the ways
Research r...	things are so and not otherwise' in society. Her	approach	represents an elaboration of the realist
Research r...	outline the main features of the morphogenetic	approach	through an exposition of its three generic
Research r...	of time when constructing the morphogenetic	approach	, as she argues that "it must be identified on
Research r...	, the two basic theorems of the morphogenetic	approach	are that "a) structure necessarily pre-dates
Research r...	are reified entities in the morphogenetic	approach	). </p><p> It therefore follows from the evidence
Research r...	presented here that the morphogenetic	approach	successfully provides the methodological
Research r...	. In the subsequent chapters the morphogenetic	approach	will be used to critique three famous
Research r...	be argued that the deficiencies of their	approaches	can be traced back to their ontological
Research r...	a realist critique, whilst the morphogenetic	approach	will be used to show that the substantive

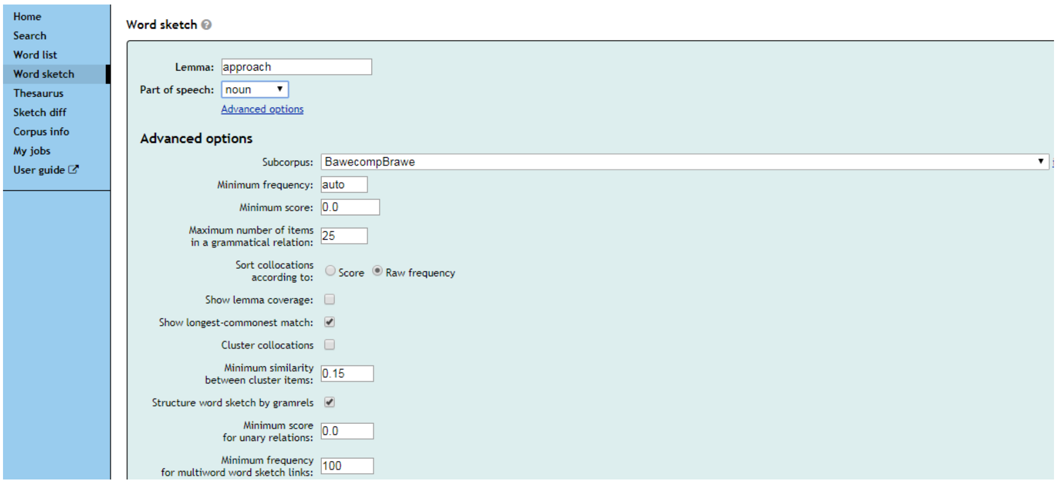
Page 1 of 81 Go Next | Last

Figure 3. Concordance screen for the node *approach* with frequency count

This procedure was repeated for every noun, i.e., for the 125 nodes.

2nd: After having a ranking with the 125 nodes organized from the most to the least frequent in the BAWE corpus, the frequency of the collocates[35] of the most frequent nodes was analyzed in both corpora by using the "Word Sketch" tool. Although this tool allows for the analysis of a variety of syntactic structures, the ones that matter for the purposes of this study are *modifier*, *object of* (verb), and *subject of* (verb). Again, the node was typed in the "lemma" box in "word sketch", and the PoS – noun was selected.

[35] The collocates were all lemmas. Hence, the goal is not to analyze inflections (different forms of the same lemma) of the lemmas (for instance, the plural form or verbs conjugated according to the subject).



Word sketch

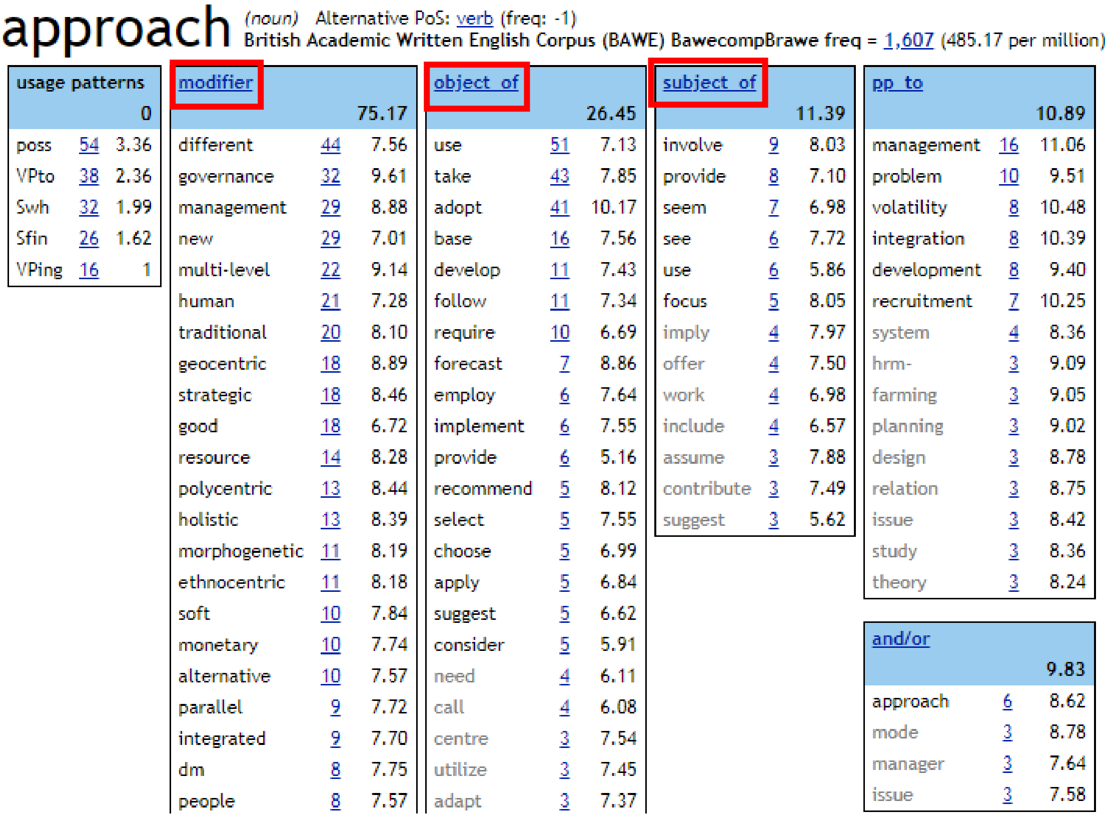
Lemma: approach
Part of speech: noun
Advanced options

Advanced options

Subcorpus: BawecompBrawe
Minimum frequency: auto
Minimum score: 0.0
Maximum number of items in a grammatical relation: 25
Sort collocations according to: ○ Score ● Raw frequency
Show lemma coverage: □
Show longest-commonest match: ☑
Cluster collocations □
Minimum similarity between cluster items: 0.15
Structure word sketch by gramrels ☑
Minimum score for unary relations: 0.0
Minimum frequency for multiword word sketch links: 100

Figure 4. Search of the node *approach* in the Word Sketch tool

The results of the collocates can be seen in Figure 5:



approach (*noun*) Alternative PoS: verb (freq: -1)
British Academic Written English Corpus (BAWE) BawecompBrawe freq = 1,607 (485.17 per million)

usage patterns		0
poss	54	3.36
VPto	38	2.36
Swh	32	1.99
Sfin	26	1.62
VPing	16	1

modifier		75.17
different	44	7.56
governance	32	9.61
management	29	8.88
new	29	7.01
multi-level	22	9.14
human	21	7.28
traditional	20	8.10
geocentric	18	8.89
strategic	18	8.46
good	18	6.72
resource	14	8.28
polycentric	13	8.44
holistic	13	8.39
morphogenetic	11	8.19
ethnocentric	11	8.18
soft	10	7.84
monetary	10	7.74
alternative	10	7.57
parallel	9	7.72
integrated	9	7.70
dm	8	7.75
people	8	7.57

object_of		26.45
use	51	7.13
take	43	7.85
adopt	41	10.17
base	16	7.56
develop	11	7.43
follow	11	7.34
require	10	6.69
forecast	7	8.86
employ	6	7.64
implement	6	7.55
provide	6	5.16
recommend	5	8.12
select	5	7.55
choose	5	6.99
apply	5	6.84
suggest	5	6.62
consider	5	5.91
need	4	6.11
call	4	6.08
centre	3	7.54
utilize	3	7.45
adapt	3	7.37

subject_of		11.39
involve	9	8.03
provide	8	7.10
seem	7	6.98
see	6	7.72
use	6	5.86
focus	5	8.05
imply	4	7.97
offer	4	7.50
work	4	6.98
include	4	6.57
assume	3	7.88
contribute	3	7.49
suggest	3	5.62

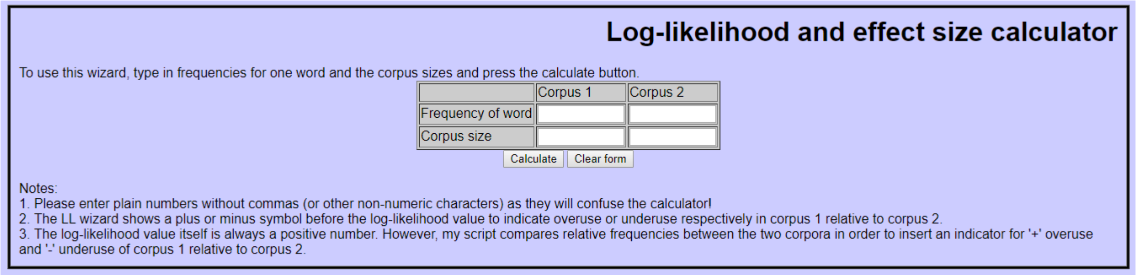
pp_to		10.89
management	16	11.06
problem	10	9.51
volatility	8	10.48
integration	8	10.39
development	8	9.40
recruitment	7	10.25
system	4	8.36
hrm-	3	9.09
farming	3	9.05
planning	3	9.02
design	3	8.78
relation	3	8.75
issue	3	8.42
study	3	8.36
theory	3	8.24

and/or		9.83
approach	6	8.62
mode	3	8.78
manager	3	7.64
issue	3	7.58

Figure 5. Collocates of the node word

As mentioned, only the frequencies of the collocates of the categories *modifier*, *object of*, and *subject of* were taken into account. With that in mind, regarding the outcomes in Figure 5, the collocates in the second, third and fourth columns were considered. Thus, collocations such as *different* + *approach*, *use* + *approach* and *approach* + *involve* are real examples with the node approach.

3rd: Log Likelihood (LL) is a test to compare frequencies of words or expressions between two corpora. (Rayson, 2002). 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). Furthermore, the outcome value can be positive (+) or negative (-). If it is positive, it means that the linguistic unit is overused in corpus 1 – in this case BrAWE -, while if the outcome is negative, the given collocate is underused in BrAWE. In order to obtain the LL outcome, the Log-Likelihood calculator[36] was used to determine whether the comparison of frequencies of the collocates of each individual noun in both corpora was statistically significant. Figure 6[37] shows the layout of LL calculator:



Log-likelihood and effect size calculator

To use this wizard, type in frequencies for one word and the corpus sizes and press the calculate button.

	Corpus 1	Corpus 2
Frequency of word		
Corpus size		

Notes:
1. Please enter plain numbers without commas (or other non-numeric characters) as they will confuse the calculator!
2. The LL wizard shows a plus or minus symbol before the log-likelihood value to indicate overuse or underuse respectively in corpus 1 relative to corpus 2.
3. The log-likelihood value itself is always a positive number. However, my script compares relative frequencies between the two corpora in order to insert an indicator for '+' overuse and '-' underuse of corpus 1 relative to corpus 2.

Figure 6. Log-Likelihood calculator

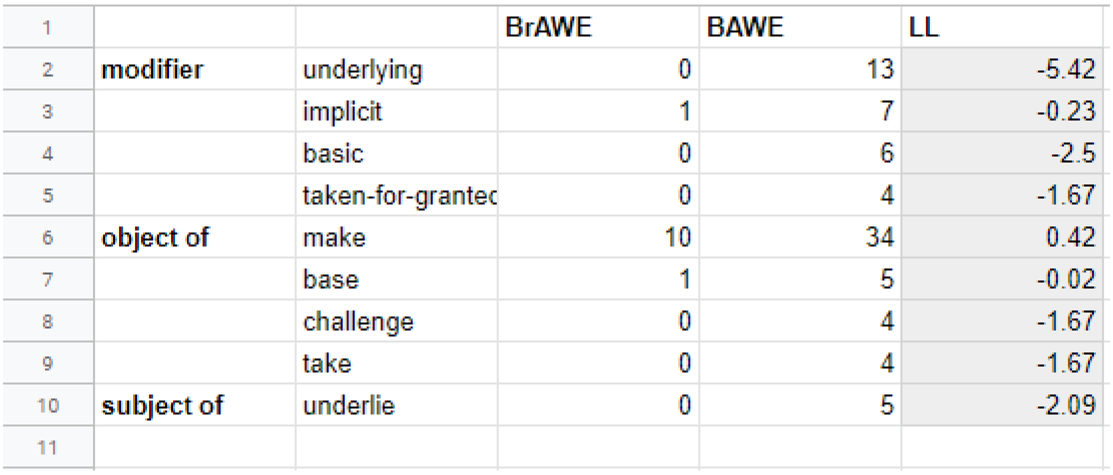
After typing the frequencies of the word and the corpora sizes, the LL was calculated by pressing the "calculate" button.

The data was organized in Excel spreadsheets to provide a better visualization of all the data. For each node, a separate spreadsheet was created to avoid missing important

[36] http://ucrel.lancs.ac.uk/llwizard.html

[37] The effect size calculator was not used, although it appears in Figure 6.

information when facing a great amount of data. Figure 7 shows an example of a spreadsheet for the node “assumption”:



1			BrAWE	BAWE	LL
2	modifier	underlying	0	13	-5.42
3		implicit	1	7	-0.23
4		basic	0	6	-2.5
5		taken-for-grantec	0	4	-1.67
6	object of	make	10	34	0.42
7		base	1	5	-0.02
8		challenge	0	4	-1.67
9		take	0	4	-1.67
10	subject of	underlie	0	5	-2.09
11					

Figure 7. Spreedsheet for the node *assumption*

The frequencies of each collocate were verified in both corpora, and the LL value was calculated. In this example, the LL values are all under 6.63 (positive and negative), meaning that the differences in frequencies of collocates composed of the node *assumption* + the collocates in the two corpora are not statistically significant. These procedures were carried out for each of the 125 nodes with the respective collocates in both corpora, resulting in 125 spreadsheets.

4th: The LL value was also calculated for each one of the 125 nodes to determine the statistically significant different nodes and how many are overused and how many are underused.

5th: Because some outcomes did not come up as expected, i.e. too many verbs with zero occurrences in BAWE, the collocational behavior of these verbs was individually analyzed through the use of the Word Sketch tool. Hence, similarly to the second step, instead of looking for the collocates of the node, the node itself had to be one of the collocates of the verb searched on Word Sketch.

3.3 Data analysis

In order to analyze the data, the analysis will be divided into two parts. First of all, the quantitative findings will be outlined containing the results regarding the nodes of the collocations in both corpora – BrAWE and BAWE. Next, the results regarding the collocates originated by the nodes with the highest frequencies will be presented.

The objective of the qualitative analysis is to check whether there are differences in the uses of collocations, and further try to explain the reasons why there are discrepancies in these uses. After defining the nodes with the highest frequencies in BAWE and BrAWE, thorough qualitative explanations on their collocates will be given. At this stage, the MI score[38] of some statistically significant collocates will be presented through the search of that specific collocate in the Word Sketch tool in Sketch Engine in a corpus of general English[39]. This stage is relevant because Brazilian students might be more prompt to use this collocate in broader contexts of general language and, therefore, overproduce it in contexts where academic English is used. Furthermore, the Word Sketch of the nodes will be explored in a

[38] Although MI scores are presented in the Word Sketch tool, they were not considered at the stage of defining whether the *node + collocate* is a collocation. In that specific step, only the raw frequencies of the collocates with the cut-off point of 4 occurrences were taken into consideration with the further calculation of the statistical significance with Log Likelihood calculator.

[39] English Web (enTenTen 2015). This corpus contains texts from Australian, Canadian, Indian, New Zealand, South African, UK and US domains of English. More information available at: https://www.sketchengine.eu/ententen-english-corpus

general Portuguese corpus as well, due to the fact that there might be L1 interference in the production of collocations in English.

The aim of this chapter was to outline the methodological procedures adopted to analyze the data. The next chapter presents the results and findings of this study.

Chapter 4: Results and discussion

In this chapter, the outcomes of the comparative analysis of collocations in BAWE and BrAWE will be presented along with the discussion of the findings. As already mentioned, a test was run in order to determine whether the differences of frequencies of the collocates in both corpora are statistically significant. The results of the test can show whether a specific collocation is overused or underused by Brazilian students in comparison to outstanding international students represented in BAWE. This chapter is organized into three main sections: the (4.1) quantitative findings, in which relevant numbers regarding the nodes and collocations in both corpora will be presented; the (4.2) qualitative findings containing more detailed explanations on the collocates of the five most frequent nodes in both corpora alongside the discussion and the correlations with previous studies; and, finally, the answers for the research questions (4.3).

4.1 Quantitative findings

As previously stated, the 125 nodes (Frankenberg et al., 2018) were organized from the most to the least frequent in BAWE (Table 6). A node is the main word of a collocation to which other words (the collocates) are associated with. For instance, in *new concept*, *key concept*, *to develop a concept*, and *to define a concept*, the node is *concept* and, consequently *new*, *key*, *develop,* and *define* are the collocates. In the table below, the nodes are presented in the second column, the raw frequencies[40] of the nodes in BAWE are in the third column, followed by their normalized values[41]. The fifth and sixth columns, respectively, portray the raw frequencies and the normalized values of the nodes in BrAWE.

[40] Raw frequency is the arithmetic count of the number of a linguistic feature (a word, a structure etc)

[41] The normalized value is the frequency of the linguistic feature, i.e. node or collocate, per thousand words.

Table 6.
Raw frequency and normalized values of the 125 nodes in both corpora

	Node	BAWE (RF)	BAWE (NV)	BrAWE (RF)	BrAWE (NV)		Node	BAWE (RF)	BAWE (NV)	BrAWE (RF)	BrAWE (NV)
1	system*	4573	1.38	1234	1.60	64	example*	834	0.25	580	0.75
2	result*	3285	0.99	1217	1.58	65	conclusion**	830	0.25	98	0.12
3	value*	3267	0.98	1008	1.31	66	conflict**	814	0.24	17	0.02
4	figure**	3034	0.91	426	0.55	67	standard**	795	0.24	126	0.16
5	process*	2947	0.88	1118	1.45	68	reference**	789	0.23	142	0.18
6	group*	2928	0.88	453	0.58	69	aspect*	777	0.23	327	0.42
7	level	2897	0.86	655	0.85	70	error*	763	0.23	224	0.29
8	model	2828	0.85	585	0.76	71	movement	763	0.23	171	0.22
9	development**	2772	0.83	456	0.60	72	task*	715	0.21	226	0.29
10	data**	2553	0.77	80	0.10	73	measure	670	0.20	181	0.23
11	information**	2496	0.75	504	0.65	74	importance	665	0.20	170	0.22
12	research**	2404	0.72	379	0.49	75	support**	662	0.19	101	0.13
13	analysis*	2377	0.71	672	0.87	76	feature*	654	0.19	213	0.27

14	rate	2295	0.69	531	0.69	77	discussion**	608	0.18	99	0.12
15	effect	2226	0.67	524	0.68	78	perspective**	607	0.18	73	0.09
16	method*	2201	0.66	625	0.81	79	influence	602	0.18	165	0.21
17	change*	2159	0.65	622	0.80	80	requirement	601	0.18	159	0.20
18	strategy**	2097	0.63	287	0.37	81	extent**	595	0.17	36	0.04
19	factor*	2072	0.62	547	0.71	82	characteristic*	574	0.17	295	0.38
20	control**	2070	0.62	383	0.49	83	interaction*	573	0.17	213	0.27
21	use*	2037	0.61	588	0.76	84	author	566	0.17	143	0.18
22	policy**	2015	0.60	123	0.16	85	degree	563	0.16	103	0.13
23	theory**	1888	0.57	206	0.26	86	capacity	554	0.16	150	0.19
24	approach**	1607	0.48	289	0.37	87	understanding	551	0.16	115	0.14
25	structure	1596	0.48	336	0.43	88	concern	548	0.16	99	0.12
26	role**	1547	0.46	180	0.23	89	pattern*	543	0.16	208	0.27
27	quality*	1460	0.44	443	0.57	90	reduction*	542	0.16	188	0.24

28	difference*	1431	0.43	424	0.55	91	basis**	540	0.16	64	0.08
29	function	1396	0.42	367	0.47	92	definition**	536	0.16	88	0.11
30	activity*	1388	0.41	393	0.51	93	procedure*	534	0.16	180	0.23
31	organisation**	1383	0.41	124	0.16	94	trend**	523	0.15	64	0.08
32	environment*	1376	0.41	383	0.49	95	consideration**	507	0.15	80	0.10
33	resource**	1337	0.40	215	0.27	96	observation**	491	0.14	59	0.07
34	type*	1327	0.40	543	0.70	97	potential	483	0.14	92	0.11
35	society**	1311	0.39	132	0.17	98	improvement*	475	0.14	227	0.29
36	condition*	1306	0.39	472	0.61	99	purpose*	470	0.14	157	0.20
37	production*	1301	0.39	473	0.61	100	finding	462	0.13	90	0.11
38	form**	1297	0.39	235	0.30	101	assumption**	460	0.13	60	0.07
39	section**	1288	0.38	207	0.26	102	outcome	446	0.13	97	0.12
40	interest**	1280	0.38	115	0.14	103	aim*	437	0.13	134	0.17
41	relationship**	1242	0.37	201	0.26	104	presence*	407	0.12	164	0.21

42	source	1222	0.36	283	0.36	105	consequence*	403	0.12	197	0.25
43	impact	1219	0.36	259	0.33	106	explanation**	398	0.12	66	0.08
44	practice**	1206	0.36	192	0.24	107	implication**	388	0.11	23	0.02
45	need**	1203	0.36	226	0.29	108	variation*	386	0.11	178	0.23
46	growth	1195	0.36	255	0.33	109	category	383	0.11	97	0.12
47	material*	1166	0.35	653	0.84	110	difficulty**	372	0.11	61	0.07
48	period**	1159	0.34	209	0.27	111	description**	361	0.10	37	0.04
49	increase*	1122	0.33	329	0.42	112	link**	338	0.10	38	0.04
50	review**	1106	0.33	77	0.10	113	attempt	335	0.10	55	0.07
51	term**	1087	0.32	194	0.25	114	shift**	329	0.09	50	0.06
52	solution*	1074	0.32	442	0.57	115	significance**	288	0.08	32	0.04
53	individual**	1023	0.30	172	0.22	116	limitation**	251	0.07	93	0.12
54	concept	1000	0.30	267	0.34	117	proportion	246	0.07	41	0.05
55	demand**	976	0.29	170	0.22	118	phenomenon	238	0.07	61	0.07

56	population*	963	0.29	401	0.52	119	contrast	234	0.07	43	0.05
57	element*	963	0.29	285	0.37	120	recognition	234	0.07	63	0.08
58	knowledge	961	0.29	224	0.29	121	contribution	225	0.06	36	0.04
59	introduction**	938	0.28	36	0.04	122	alternative*	208	0.06	75	0.09
60	benefit	914	0.27	179	0.23	123	insight	168	0.05	30	0.03
61	experience	898	0.27	169	0.21	124	tendency	117	0.03	29	0.03
62	technique*	889	0.26	284	0.36	125	exception	101	0.03	16	0.02
63	range*	883	0.26	268	0.34						

Note. RF = raw frequency; NC = normalized value. This value indicates the frequency of the node per thousand words.
*Overused. **Underused

The most frequent node in both corpora is *system* and the node with the lowest frequency is *exception*. Considering that the sizes of the corpora are different, the raw frequencies do not indicate much, so the normalized values give a clearer picture of the real frequencies. On top of that, the LL value of each node was calculated, and the results indicated that 89 nouns came up as being statistically different. From these 89, 49 represent Brazilian students' underuse (marked with **), while the remaining 40 represent overuse (marked with *) of that specific academic word.

Some striking differences are worth mentioning, such as the node *data*. It occupies the 10th position in BAWE's ranking from the most to the least frequent node, whereas in BrAWE the same node appears in the 98th position. In BAWE, *data* appears 2553 times with a 0.77 normalized value, while in BrAWE the numbers are 80 and 0.10 for raw frequency and normalized values, respectively. Additionally, when analyzing the Word Sketch of this node, it is possible to observe that the variety of collocates in BAWE is much wider than in BrAWE. Thus, with collocations such as *quantitative data* and *available data*, it seems that students in the British corpus characterize the data being analyzed, whereas Brazilians use *data* as an isolated academic word, rather than the node of academic collocations.

Example is another intriguing case with discrepancies in both corpora. Interestingly, *example* occupies the 64th position in BAWE's ranking, while in BrAWE it appears as the 12th most frequent node. In order to speculate on this difference, Table 7 contains the normalized values of *for example* in BAWE and in BrAWE, as well as in a general corpus and in a Portuguese corpus:

Table 7.
Normalized values of 'for instance' and 'for example'

For example (BAWE)	For example (BrAWE)	For example (general English)	'Por exemplo'* (Portuguese corpus)

0.09	0.11	0.14	0.22

*Cognate for ‘for example’ in Portuguese

Based on the normalized values in Table 7, it looks like *example* occupies such a high position in BrAWE’s node ranking in comparison to BAWE’s, due to a L1 transfer from the Portuguese language into academic registers. This fact can be observed in the low normalized value of *for example* in BAWE, in comparison with the other three corpora: 0.11, 0.14 and 0.22 in BrAWE, in the general English corpus and in the Portuguese one, respectively.

The statistically significant differences are related to the Brazilian corpus, which means that the cases of overuse and underuse are in BrAWE in comparison with BAWE. Figure 8 shows the percentages of the overused and underused nodes among the statistically significant ones:

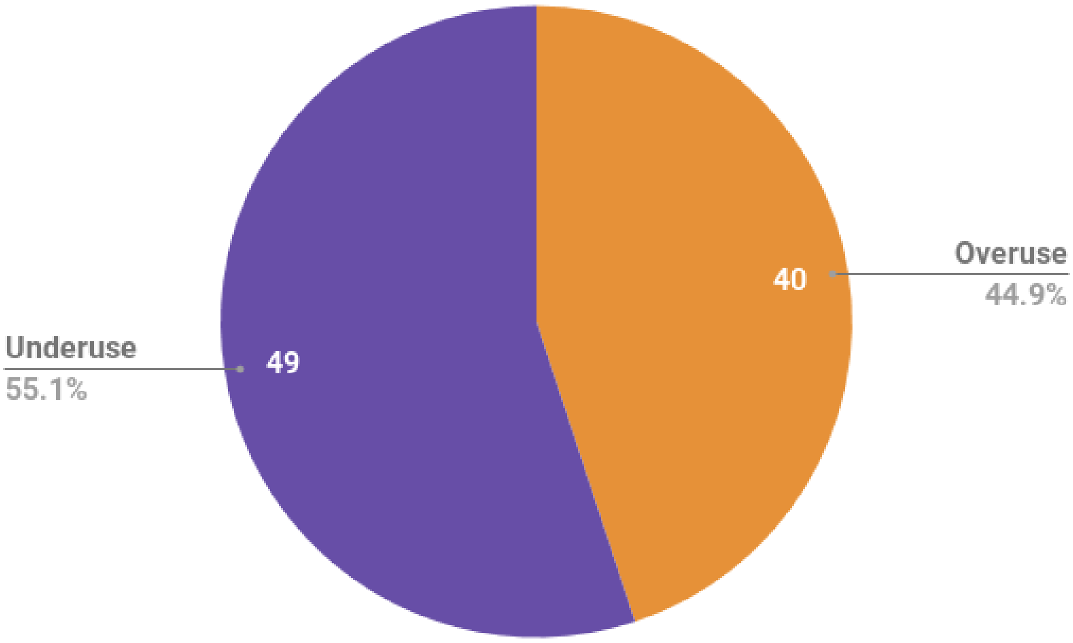


Figure 8. Overused and underused nodes

Based on the understanding that a collocation is a sequence of words that co-occur more often than would be expected by chance, that is, **a combination of two words that are associated due to statistical probabilities of occurring together**, from the 125 nodes, 2522 collocates were found. Also, the cut-off point established for considering the candidate collocates was a minimum frequency of four with a range of at least two areas of expertise. Thus, the collocate must appear at least four times and be used in two or more different areas – PS, LS, SS or AH.

Taking all the collocates into consideration[42], 323 showed a significant difference, which represents only 12.8% of the data. Out of the 323, Brazilian students overuse 132 (40.9% of the 323) and underuse 191 (59.1%). Therefore, it is possible to affirm that, in general terms, Brazilians underuse both academic nodes (Figure 8) and academic collocations when compared to students represented in the BAWE corpus, considering that the number of underused collocations is higher than the number of overused ones. Figure 9 pictures the representation of the collocates which are statistically different among the total data. The gray circle represents the total amount of collocates that emerged from the data; the yellow circle symbolizes the statistically different ones, of which the circles in purple and in orange come out, representing the underused and the overused collocates in the purple and orange circles, respectively.

[42] Appendix A contains the total amount of collocates that go together with the 125 nouns.

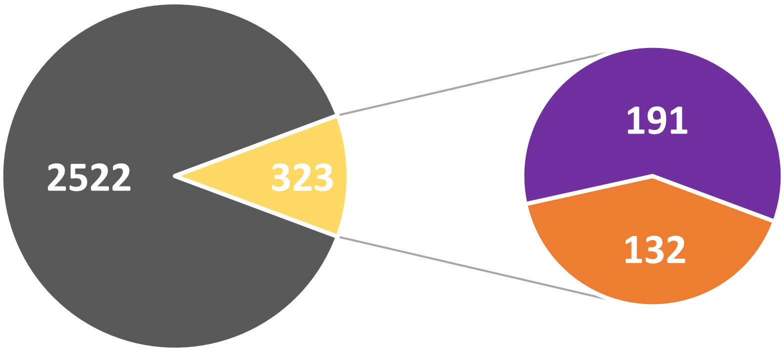


Figure 9. Statistically significant different collocates

For this study, the five most frequent nodes in BAWE as well as the five most frequent in BrAWE will be further analyzed. The top five nodes in BAWE are *system*, *result*, *value*, *figure* and *process*. In BrAWE, the five most frequent nodes are *system*, *result*, *process*, *value* and *analysis*. *Figure* is the 24th most frequent node in BrAWE, and *analysis* occupies the 13th position in BAWE. Table 8 provides information regarding the number of collocates in both corpora for each one of the five most frequent nodes in BAWE and BrAWE.

Table 8.
Number of collocates of the five most frequent nodes in BAWE and BrAWE

		Number of collocates
System	BAWE	48
	BrAWE	30
Result	BAWE	56
	BrAWE	44
Value	BAWE	52
	BrAWE	33

Figure*	BAWE	15
	BrAWE	3
Process	BAWE	56
	BrAWE	31
Analysis**	BAWE	35
	BrAWE	23
TOTAL	BAWE	262
	BrAWE	164

*Figure is among the top 5 nodes in BAWE only. **Analysis is among the top 5 nodes in BrAWE only.

Considering the data in Table 8, it can be observed that *system* has 48 collocates in BAWE and 30 collocates in BrAWE. *Result* contains 56 collocates in BAWE and 44 in BrAWE. *Value* presents 52 different collocates in BAWE and 33 in BrAWE. *Figure* appears with 15 collocates in BAWE, and only three in BrAWE. *Process* has 56 different collocates in BAWE and 31 in BrAWE. *Analysis* is accompanied by 35 collocates in BAWE and by 23 in BrAWE. The total number of different collocates of the top five nodes in BAWE and BrAWE sum up 262 and 164 in each corpus, respectively. Figure 10 illustrates the statistically significant different (SSD) collocates among the total amount of collocates that emerged from the analysis of the six nodes listed above:

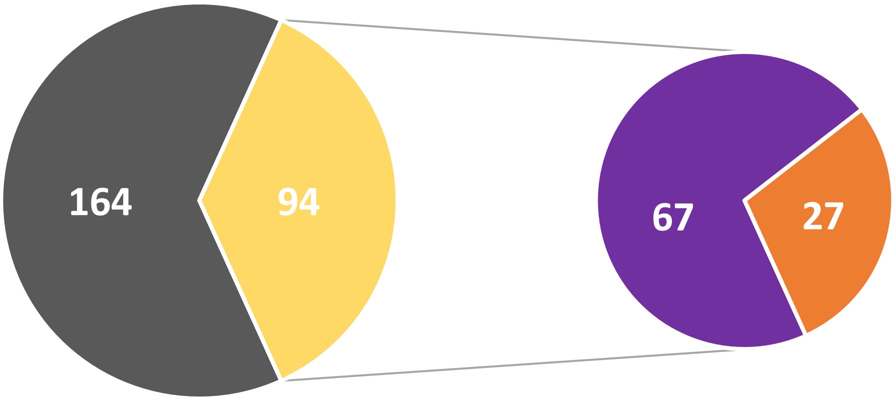


Figure 10: Statistically significant different collocates in the six nodes analyzed

From the 164 collocates composed of one of the six nodes under discussion – *system*, *result*, *value*, *figure*, *process* and *analysis* – 94 collocates came up as being SSD (yellow circle). Out of these 94, 67 are overused (purple circle) and 27 are underused (orange circle) by Brazilians represented in BrAWE.

In the next section, the collocations of the most frequent nodes both in BAWE and in BrAWE will be analyzed.

4.2 Qualitative findings

This section is focused on the collocations related to the nodes with the highest frequencies in both corpora. Thus, each node with their respective collocates will be analyzed in a different subsection, being 4.2.1 for *system*, 4.2.2 for *result*, 4.2.3 for *value*, 4.2.4 for *figure*, 4.2.5 for *process*, and 4.2.6 for *analysis*. The collocates that accompany each node can be of three different categories: 'modifier', that refers to the adjective that comes before the node as in *new* + *system* or *good* + *result*; 'verb (object of)’, that refers to the verb that is the object of the sentence, as in *design + system* and *show + result*, in which *system* and *result* are the object of the verbs *design* and *show*, respectively; or ‘verb (subject of)’, referring to the verb that collocates with the node as the subject, as in *system + have* and *result + show*, in which *system* and *result* are the subjects of the verbs *have* and *show*, respectively.

4.2.1 Collocations of *system*

In this subsection, the collocations with the node *system* will be analyzed. Among the 125 academic nodes (Frankenberg-Garcia et al., 2018), *system* is the most frequent one in both corpora (in overall frequency), and overused by Brazilians. The normalized values are 1.38 in BAWE, and 1.60 in BrAWE. Table 9 contains all the collocates that go together with this node.

Table 9.
Collocations of system in both corpora

SYSTEM				
		BrAWE	BAWE	LL
modifier	control	15	62	0.02
	new	0	45	-18.78**
	reward	0	35	-14.6**
	production	35	35	34.45*
	whole	11	35	0.73
	current	0	34	-14.19**
	computer	0	25	-10.43**
	communication	0	23	-9.6**
	complex	9	22	1.88
	health	17	0	56.77*
verb (object of)	use	23	58	4.35
	design	0	38	-15.85**
	base	7	24	0.27
	develop	12	22	5.1
	implement	9	19	2.82
	make	6	17	0.73

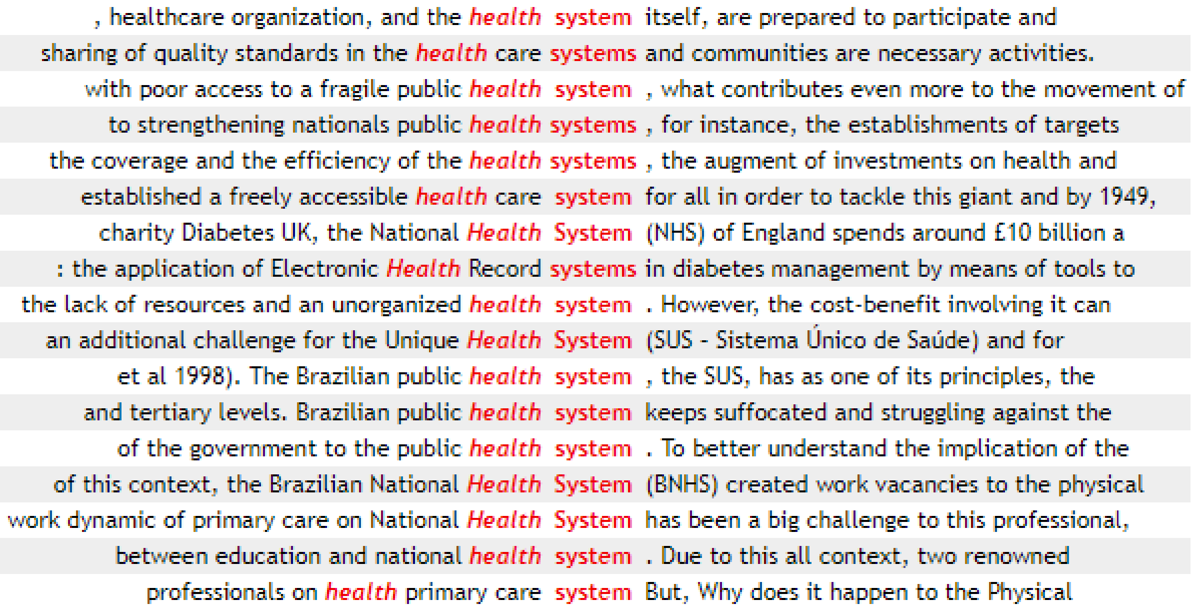
	show	4	14	0.13
	provide	4	14	0.13
	enter	0	13	-5.42
	introduce	0	12	-5.01
	consider	0	12	-5.01
	test	0	11	-4.59
	create	9	11	7.12*
	apply	8	0	26.72*
	present	5	0	16.70*
	help	4	0	13.36*
	choose	4	5	3.08
	install	0	9	-3.76
	improve	5	8	2.71
	compare	0	6	-2.5
	explain	0	6	-2.5
	need	4	6	2.4
	involve	0	6	-2.5
	affect	4	6	2.4
	see	0	6	-2.5
verb (subject of)	use	3	20	-0.55
	provide	3	12	0.01
	work	4	10	0.78
	contain	0	9	-3.76
	need	3	9	0.28
	make	3	9	0.28
	enable	2	8	0.01
	allow	5	7	3.32
	become	6	7	5.01
	mean	0	6	-2.5
	lead	0	6	-2.5
	show	2	6	0.19
	define	0	5	-2.09
	depend	0	5	-2.09
	play	0	5	-2.09
	create	0	5	-2.09
	include	0	5	-2.09
	have	24	0	45.23*
Number of collocates		30	48	

*Overused. **Underused

Based on Table 9, there are 30 collocates in BrAWE and 48 collocates in BAWE. Among these 48 in BAWE, there are some words with zero occurrences in BrAWE that did not came up as being SSD. This is the case of the verbs *enter, introduce, consider, test, install, compare, explain, involve, see, contain, mean, lead, define, depend, play, create,* and *include*. Moreover, 13 collocates are SSD, out of which seven are overused by Brazilians and six are underused based on the LL value, higher than 6.63, either positive or negatively. The SSD collocates are highlighted with * and ** to indicate the cases of overuse and underuse respectively.

Brazilian students overuse *production* and *health* in the 'modifier' category; *create, apply, present, help,* and *choose* in the 'object of (verb)' category; and *have* in the 'subject of (verb)' category. When it comes to the underused collocates, we have *new, reward, current, computer* and *communication* in the 'modifier' category; *design* is the underused verb in the 'object of' category.

Considering the cases of overuse, *health, apply, present, help* and *have* have zero occurrences in the BAWE corpus when coupled with *system*. The collocation *health* + *system,* for instance, has zero occurrences in BAWE and it accounts for 6.6% of the total amount of collocations with *system* as the node in BrAWE. In the general English corpus, this collocation appears 157,402 times, i.e. 8.56 occurrences per million. The concordance lines for *health system* in BrAWE are shown below:



, healthcare organization, and the health system itself, are prepared to participate and
sharing of quality standards in the health care systems and communities are necessary activities.
with poor access to a fragile public health system , what contributes even more to the movement of
to strengthening nationals public health systems , for instance, the establishments of targets
the coverage and the efficiency of the health systems , the augment of investments on health and
established a freely accessible health care system for all in order to tackle this giant and by 1949,
charity Diabetes UK, the National Health System (NHS) of England spends around £10 billion a
: the application of Electronic Health Record systems in diabetes management by means of tools to
the lack of resources and an unorganized health system . However, the cost-benefit involving it can
an additional challenge for the Unique Health System (SUS - Sistema Único de Saúde) and for
et al 1998). The Brazilian public health system , the SUS, has as one of its principles, the
and tertiary levels. Brazilian public health system keeps suffocated and struggling against the
of the government to the public health system . To better understand the implication of the
of this context, the Brazilian National Health System (BNHS) created work vacancies to the physical
work dynamic of primary care on National Health System has been a big challenge to this professional,
between education and national health system . Due to this all context, two renowned
professionals on health primary care system But, Why does it happen to the Physical

Figure 11. Concordance lines for *health system* in BrAWE.

Health system is used 17 times in BrAWE. The areas from which these concordance lines were taken are Physical Sciences, Life Sciences and Social Sciences. The collocation appears twice in PS texts (11.7% of the 17), four times in SS (23.5%) and 11 times in LS (64.8%) It is also worth highlighting that all the underused collocates – *new*, *reward*, *current*, *computer*, and *communication* in the ‘modifier’ category, and *design* in the ‘object of’ category - have zero occurrences in the BrAWE corpus. In the category ‘object of’, the verb *design* as a collocate of *system* is not used at all by Brazilian students who apparently prefer to use the verb *create* instead, which is an overused verb in the ‘object of’ category. Below, there is a Physical Science excerpt where *system* is the object of the verb *create* in BrAWE. The collocate is in italics (*create*) and the node is in bold and italics (***system***).

PS excerpt - ENUT03I65

After the charges were imposed, there was a significant increase on public transport use (Albalate and Germà, 2009). With the level of technology that is available nowadays, it is possible to *create* ***systems*** that follow closely the theories of congestion charging, efficiently internalizing the external costs of congestion. Congestion pricing has proved to be effective and capable of bringing several benefits to the entire society. However, the political and human element is above it, and that is why it was successful in some places and a failure in others.

In the case of the collocations *create* + *system* and *design* + *system*, the meaning is basically the same, as they both indicate 'the process of making something new that did not exist before'[43], as can be seen in the examples below. The collocates are highlighted in *italics* and the node ***system*** is featured in bold and in italics.

SS excerpt - POUT01I155 (BrAWE)

Moreover, Womack (1990) says that Lean Thinking grants flexibility and responsibilities to all involved in the production line. Further, managers must *create* a ***system*** that is able to deal with identified defects immediately and capable to solve completely.

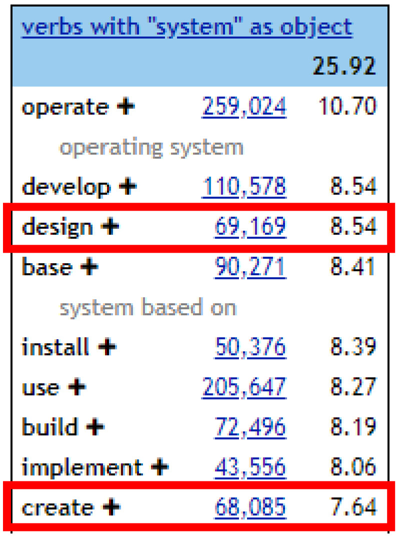
PS excerpt - 6107d (BAWE)

Holding this design philosophy, we will deal with more practical details in the following implementation session. </p> 4. Implementation <p> In the previous part, we have *designed* the whole ***system*** including the exact interfaces. According to the design philosophy, a figure about the whole buggy system has been drawn and given below.

[43] https://www.macmillandictionary.com/

In the first excerpt taken from BrAWE, it should be noted that the meaning of ‘creating a system’ is related to the fact that there is no available system with those characteristics yet. Therefore, it is necessary that managers devise a new one. In the BAWE excerpt, the need for having a new system is similar. Hence, students in the British corpus apparently prefer to use the collocation *design* + *system* rather than *create* + *system*.

When analyzing the collocates of *system* with the Word Sketch tool in a corpus of general English[44], the MI scores (the third column in Figure 12) of *design* and *create* are different, and reveal that one has a higher collocational strength. While *design* has a MI of 8.54, *create* has a MI of 7.64. The higher the MI score, the stronger the relation between the items. Thus, this difference in the MI score corroborates that *design* + *system* is a preferred choice by speakers of English as a first language and may help explain this choice in BAWE.

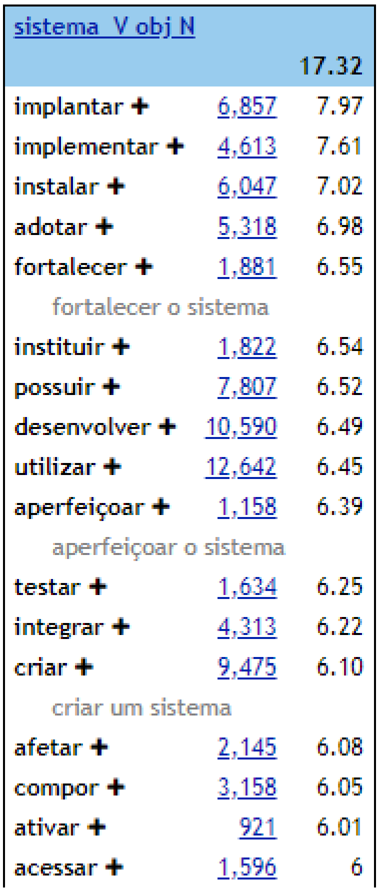


verbs with "system" as object		
		25.92
operate +	259,024	10.70
operating system		
develop +	110,578	8.54
design +	69,169	8.54
base +	90,271	8.41
system based on		
install +	50,376	8.39
use +	205,647	8.27
build +	72,496	8.19
implement +	43,556	8.06
create +	68,085	7.64

Figure 12: verbs that collocate with *system* as object in the general English corpus

[44] As explained in the previous chapter, the general English corpus used for the data analysis was the English Web (*enTenTen 2015).*

Another possible explanation for the overuse of *create* + *system* and the underuse of *design* + *system* in BrAWE is L1 influence, in this case, Portuguese. The verbs that collocate with ‘sistema’ (the cognate[45] for *system* in Portuguese) as an object are shown below:

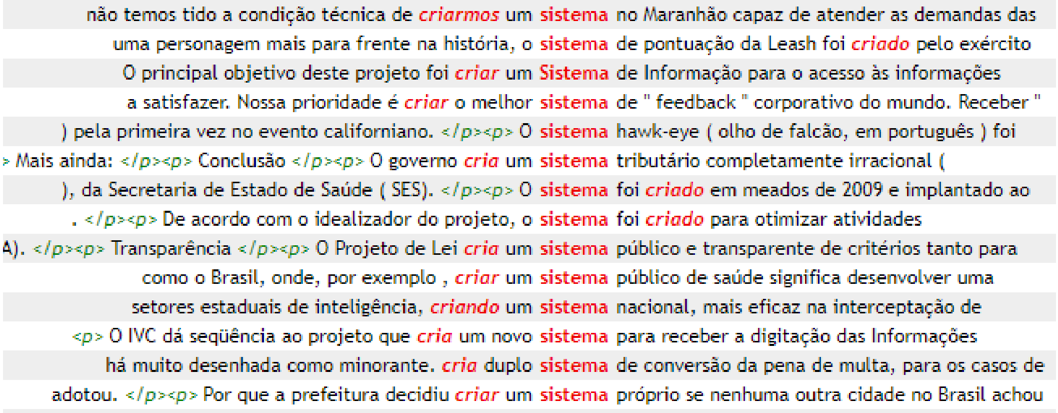


sistema V obj N		
		17.32
implantar +	6,857	7.97
implementar +	4,613	7.61
instalar +	6,047	7.02
adotar +	5,318	6.98
fortalecer +	1,881	6.55
fortalecer o sistema		
instituir +	1,822	6.54
possuir +	7,807	6.52
desenvolver +	10,590	6.49
utilizar +	12,642	6.45
aperfeiçoar +	1,158	6.39
aperfeiçoar o sistema		
testar +	1,634	6.25
integrar +	4,313	6.22
criar +	9,475	6.10
criar um sistema		
afetar +	2,145	6.08
compor +	3,158	6.05
ativar +	921	6.01
acessar +	1,596	6

Figure 13: verbs that collocate with the object ‘sistema’ in the Portuguese corpus

The verb ‘criar’ is the cognate for *create* in Portuguese, and it appears as one of the collocates of the node ‘sistema’. The equivalents for *design* (‘projetar’, ‘planejar’, or even ‘design’), however, do not figure among the possible verbs that best collocate with the node being discussed. Hence, another speculation for Brazilians overusing *create* + *system* might have to do with L1 interference. The concordance lines of ‘criar + sistema’ are presented in Figure 14:

[45] In this study, cognates are understood as equivalents *prima facie*.



```
       não temos tido a condição técnica de criarmos um sistema no Maranhão capaz de atender as demandas das
         uma personagem mais para frente na história, o sistema de pontuação da Leash foi criado pelo exército
           O principal objetivo deste projeto foi criar um Sistema de Informação para o acesso às informações
            a satisfazer. Nossa prioridade é criar o melhor sistema de " feedback " corporativo do mundo. Receber "
    ) pela primeira vez no evento californiano. </p><p> O sistema hawk-eye ( olho de falcão, em português ) foi
> Mais ainda: </p><p> Conclusão </p><p> O governo cria um sistema tributário completamente irracional (
    ), da Secretaria de Estado de Saúde ( SES). </p><p> O sistema foi criado em meados de 2009 e implantado ao
       . </p><p> De acordo com o idealizador do projeto, o sistema foi criado para otimizar atividades
A). </p><p> Transparência </p><p> O Projeto de Lei cria um sistema público e transparente de critérios tanto para
               como o Brasil, onde, por exemplo , criar um sistema público de saúde significa desenvolver uma
             setores estaduais de inteligência, criando um sistema nacional, mais eficaz na interceptação de
      <p> O IVC dá seqüência ao projeto que cria um novo sistema para receber a digitação das Informações
          há muito desenhada como minorante. cria duplo sistema de conversão da pena de multa, para os casos de
     adotou. </p><p> Por que a prefeitura decidiu criar um sistema próprio se nenhuma outra cidade no Brasil achou
```

Figure 14: concordance line of ‘criar + sistema’ in the Portuguese corpus

Another intriguing result is the overuse of the collocate *apply* in BrAWE with 0 occurrences in BAWE. Although ‘aplicar’ (the cognate of *apply* in Portuguese) does not collocate with ‘sistema’ in the Portuguese corpus, the overuse of the collocate *apply* might be explained by the fact that ‘adotar’, which is a verb similar in meaning to ‘aplicar’, figures as the verb with the fourth highest MI score in the Portuguese corpus (MI = 6.98 as shown in Figure 13). By observing the following excerpt retrieved from a PS text in BrAWE,

PS excerpt - BSUT02I89 (BrAWE)

> The challenges in *applying* these **systems**, however, are much higher, the developing countries many times do not have enough resources or they lack stability to successfully *apply* the **systems.**

it is possible to observe that the collocation under discussion is used twice with the meaning of adopting, using, putting into practice those specific systems. If it were in the BAWE corpus, the verb choice would be probably *implement* or *use*, which are among the five verbs with the highest frequencies that collocate with *system* as the object.

4.2.2 Collocations of *result*

In this subsection, the collocations with the node *result* will be analyzed. Among the 125 nodes (Frankenberg-Garcia et al., 2018), *result* is the second most frequent one in both corpora, however, it is proportionally much more frequent in BrAWE (1.58) than in BAWE (0.99). Even so, in BAWE, there are 56 collocates, while in BrAWE the total amount of collocates with *result* as the node is 44 (Table 7). The collocates that go together with this node are displayed in Table 10:

Table 10.
Collocations of result in both corpora

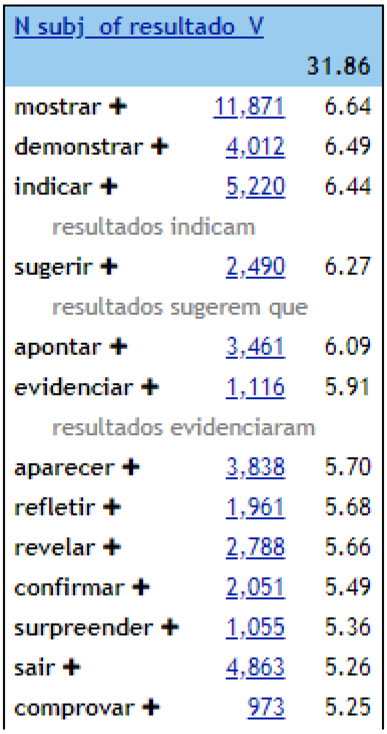
		RESULT		
		BrAWE	BAWE	LL
modifier	good	43	46	39.51*
	experimental	4	28	-0.93
	test	15	23	8.71*
	different	9	22	1.88
	accurate	9	18	3.19
	similar	4	17	0
	final	12	14	10.03*
	direct	0	12	-5.01
	end	0	11	-4.59
	search	0	11	-4.59
	following	0	11	-4.59
	analysis	0	10	-4.17
	expected	0	10	-4.17
	research	0	9	3.76
	negative	6	9	3.6
	positive	14	8	21.25*
	previous	0	7	-2.92
	above	0	6	-2.5
	poor	2	6	0.19
	financial	3	6	1.06

	same	12	0	40.07*
	more	9	0	30.06*
	reliable	6	0	20.04*
	successful	5	0	16.70*
object of	obtain	62	99	33.74*
	show	51	69	35.46*
	produce	7	57	-2.97
	give	15	47	1.09
	compare	15	30	5.32
	provide	11	25	2.85
	yield	0	23	-9.6**
	present	25	19	31.24*
	achieve	20	17	22.83*
	affect	0	15	-6.26
	find	19	14	24.31*
	get	8	12	4.8
	report	0	10	-4.17
	analyse	8	10	6.16
	gain	0	10	-4.17
	expect	19	9	32.04*
	summarise	0	8	-3.34
	take	0	8	-3.34
	interpret	4	7	1.86
	generate	5	7	3.32
	see	0	7	-2.92
	use	0	7	-2.92
	gather	0	6	-2.5
	have	20	0	66.79*
	improve	8	0	26.72*
	bring	6	0	20.04*
	explain	5	4	6.00
	influence	5	4	6.00
	collect	5	0	16.70*
	record	4	5	3.08
	confirm	4	0	13.36*
	plot	4	0	13.36*
subject of	show	44	96	12.7*
	indicate	11	25	2.85
	follow	4	10	0.78
	suggest	11	8	14.21*

	confirm	3	6	1.06
	support	0	6	-2.5
	find	0	6	-2.5
	give	0	4	-1.67
	seem	0	4	-1.67
	demonstrate	5	3	7.36*
Number of collocates		44	56	

*Overused. **Underused

There are 24 SSD collocates with the node *result*. From these 24, 23 are overused, and only one is underused in BrAWE. This is probably due to the fact that even though *result* is proportionally more used by Brazilians, the variety of collocations in BAWE is wider, thus, Brazilian collocates will be stronger. *Direct*, *end*, *search*, *following*, *analysis*, *expected*, *research*, *previous*, *above*, *affect*, *summarise, take*, *see*, *use, gather*, *support*, *find*, *give,* and *seem* have zero occurrences in BrAWE as collocates of *result*., whereas *same*, *more*, *reliable,* and *successful* in the ‘modifier’ category, *have, improve, bring, collect*, *confirm,* and *plot* in the ‘verb (object of)’ category do not happen in BAWE. In order to find an explanation about the differences from both corpora, the corpus of Portuguese was resorted for the verbs that best collocate with ‘resultado’ (the congnate for *result* in Portuguese) when it is the subject of the sentence are (Figure 15):



N subj_of resultado V		
		31.86
mostrar +	11,871	6.64
demonstrar +	4,012	6.49
indicar +	5,220	6.44
resultados indicam		
sugerir +	2,490	6.27
resultados sugerem que		
apontar +	3,461	6.09
evidenciar +	1,116	5.91
resultados evidenciaram		
aparecer +	3,838	5.70
refletir +	1,961	5.68
revelar +	2,788	5.66
confirmar +	2,051	5.49
surpreender +	1,055	5.36
sair +	4,863	5.26
comprovar +	973	5.25

Figure 15: verbs that collocate with ‘resultado’ as a subject in the Portuguese corpus

Hence, the overuse of *demonstrate* in BrAWE might be explained by L1 interference again, since ‘demonstrar’ appears as the second top collocate with the subject ‘resultado’.

In relation to the ‘modifier’ category, the SSD collocates are *good*, *test*, *final*, *positive*, *same*, *more*, *reliable*, and *successful*, the last four having zero occurrences in BAWE. Additionally, it is possible to observe that, although *good result* is frequent in both corpora, it is more frequent in BrAWE probably because *good* is a very frequent word in general English too.

With regards to the verbs that accompany the node under discussion, except for *yield*, they are all overused. Table 11 contains concordance lines with the SSD collocations composed of the ‘object of’ verbs in BrAWE that have at least four occurrences in BAWE. Each line of the column refers to an example extracted from BrAWE. The node word being analyzed – *result* – is centralized in the second column and highlighted in italics and in bold. What comes before the node word can be found in the first column, whereas the context after the node word comes right after it, in the third column. The verbs that collocate with the node *result* are in italics. An indication of the area from which the concordance line was taken can be found in the last column of the table.

Table 11
Statistically significant different collocations comprising the 'object of' category verbs in BrAWE with occurrences in BAWE

Considering the first day as March 21st, the date corresponding to this temperature is October 6th. This finding is in accordance with typical	***results***	*obtained* from the TML model, where the temperature peak for bottom water takes place at the overturn point.	PS excerpt - BAUT03I47 (BrAWE)
The bacteria susceptibility to antibiotics found on the	***results***	*shown* on table 2, as well as the inefficacy of penicillin against E. coli and P. aeruginosa; of ampicillin, chloramphenicol and tigecycline were supported by the results of CMS (2003) that present similar range of the length of zone of inhibition.	LS excerpt - MDUT04I80 (BrAWE)
After filtering only the target market responses ('upper-class people aged 18 to 34 in the Southern Brazil, more precisely in Porto Alegre'), the new	***results***	are *presented* as follows: The questionnaire provided information that should be considered when setting up the business.	SS excerpt - LOUT03I77 (BrAWE)
However, the	***results***	*achieved* in the calculation were not suitable for the expected, giving a high loop gain.	PS excerpt - REUT01I87 (BrAWE)
This, in turn, means that this	***result***	was *found* partially by the difference in expertise. In the correlation test between anticipation accuracy and recognition sensitivity, the authors reported that significance almost reached a significant level.	LS excerpt - BLUT03I58 (BrAWE)
On the other hand, the design team has to deliver a design that fits the requirements of the company and be clear when giving the construction company (Racional) what the design is like, how it should be executed and what	***results***	are *expected*.	PS excerpt - WEUT01 (BrAWE)

It is possible to observe that the examples in Table 11 are in the passive voice. The noun phrase acting as the subject in the passive voice usually corresponds to the direct object of the associated sentence in the active voice. Furthermore, these examples show that the auxiliary verb in the construction of the passive voice is not always used. When an auxiliary verb is used, however, it appears either in the present (is/are) or in the past (was/were). Table 12 shows when the verbs from the 'object of' category are accompanied by an auxiliary verb in the present or in the past. The raw frequency, i.e. number of hits, and the normalized values are indicated:

Table 12.
Number of hits vs. passive voice of the 'object of' category verbs that collocate with result in BrAWE with occurrences in BAWE

	Number of hits in BrAWE	Passive voice	Auxiliary verb	
			Present	Past
obtain	62 (0.080)	52 (0.067)	2 (0.002)	4 (0.005)
show	51 (0.066)	21 (0.027)	15 (0.019)	0 (0)
present	25 (0.032)	12 (0.015)	3 (0.003)	0 (0)
achieve	20 (0.026)	5 (0.006)	1 (0.001)	2 (0.002)
find	19 (0.024)	16 (0.020)	0 (0)	5 (0.006)
expect	19 (0.024)	1 (0.001)	1 (0.001)	0 (0)
TOTAL	196 (0.255)	107 (0.139)	22 (0.028)	11 (0.014)

In BrAWE, there are 196 occurrences of SSD verbs that appear in the BAWE corpus and go together with the node *result* when it is the object of the sentence. From these 196, 107 are used in the passive voice, out of which 33 contain the auxiliary verb, i.e. 22 are in the present whereas the remaining 11 are used in the past.

Table 13.
Number of hits vs. passive voice of the SSD verbs in the 'object of' category verbs that collocate with result in BAWE

	Number of hits in BAWE	Passive voice	Auxiliary verb	
			Present	Past
obtain	99 (0.029)	85 (0.025)	7 (0.002)	14 (0.004)
show	69 (0.020)	26 (0.007)	18 (0.005)	0 (0)
present	19 (0.005)	6 (0.001)	3 (0.0009)	0 (0)

achieve	17 (0.005)	2 (0.0006)	1 (0.0003)	0 (0)
find	17 (0.005)	8 (0.002)	1 (0.0003)	2 (0.0006)
expect	9 (0.002)	5 (0.001)	1 (0.0003)	0 (0)
TOTAL	230 (0.069)	132 (0.039)	31 (0.009)	16 (0.004)

In BAWE, there are 230 occurrences of the SSD verbs that go together with the node *result* when it is the object of the sentence. From these 230, 132 are used in the passive voice, out of which 47 contain the auxiliary verb, i.e. 31 are in the present whereas the remaining 16 are used in the past. While in BrAWE *achieve* is used twice in the passive voice in the past (Table 12), there are no occurrences of this verb being used in the passive voice in the past by students in BAWE.

When calculating the LL value of the uses of passive voice in both corpora (107 in BrAWE and 132 in BAWE), the result indicates an overuse (LL = 83.70) of the passive voice by Brazilians as far as the verbs *obtain*, *show*, *present*, *achieve*, *find*, and *expect* collocating with *result* are concerned. In order to check if there are statistically significant differences in the uses of the passive voice with the node *result* for each verb individually (Tables 12 and 13), the following LL values are found:

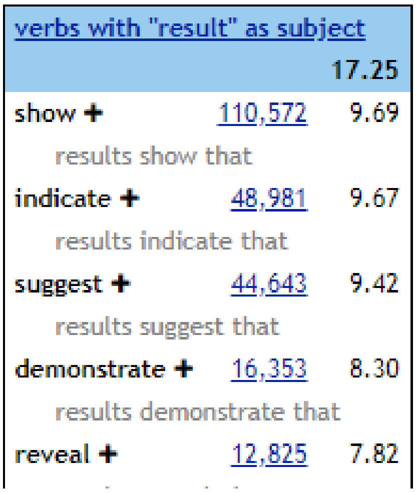
Table 14.
LL value of the passive voice with the node result in both corpora

	BrAWE	BAWE	LL	BrAWE	BAWE	LL	BrAWE	BAWE	LL
	Passive voice			Auxiliary verb (present)			Auxiliary verb (past)		
obtain	52 (0.067)	85 (0.025)	27.23	2 (0.002)	7 (0.002)	0.06	4 (0.005)	14 (0.004)	0.13
show	21 (0.027)	26 (0.007)	16.36	15 (0.019)	18 (0.005)	12.13	0 (0)	0 (0)	0.00
present	12 (0.015)	6 (0.001)	19.66	3 (0.003)	3 (0.0009)	2.95	0 (0)	0 (0)	0.00
achieve	5 (0.006)	2 (0.0006)	9.16	1 (0.001)	1 (0.0003)	0.98	2 (0.002)	0 (0)	6.68

find	16 (0.020)	8 (0.002)	26.22	0 (0)	1 (0.0003)	-0.42	5 (0.006)	2 (0.0006)	9.16
expect	1 (0.001)	5 (0.001)	-0.02	1 (0.001)	1 (0.0003)	0.98	0 (0)	0 (0)	0.0

Based on Table 14, the passive voice is overused in BrAWE with the verbs *obtain*, *show*, *present*, *achieve*, and *find*, whereas with the verb *expect*, the passive voice is not SSD. In relation to the use of the auxiliary verb in the present, the statistically significant difference is observed in the verb *show*, which is overused in BrAWE when accompanied by an auxiliary verb in the passive voice. In the past, the verbs *achieve* and *find* are SSD, being also overused by Brazilians.

When it comes to the 'subject of' category, *show*, *suggest*, and *demonstrate* collocate with *result* and are overused in BrAWE. These verbs are in the top four verbs that collocate with *result* in the general English corpus, with the MI scores of 9.69, 9.42 and 8.30 respectively, as illustrated in Figure 16:



verbs with "result" as subject

		17.25
show +	110,572	9.69
results show that		
indicate +	48,981	9.67
results indicate that		
suggest +	44,643	9.42
results suggest that		
demonstrate +	16,353	8.30
results demonstrate that		
reveal +	12,825	7.82

Figure 16: verbs that collocate with the subject *result* in the general English corpus.

Hence, besides L1 transfer as shown in Figure 15, the overuse of these verbs might also be explained by transfer from general English.

Regarding the voice of the sentences whenever *result* is the subject, Brazilians prefer the active voice with the collocations *result + demonstrate*. Another aspect to be highlighted is the verb tense. Table 15 shows the verb tenses of the SSD verbs collocating with *result* in the 'subject of' category:

Table 15.
Verb tenses of the SSD verbs that collocate with result as the subject

	BrAWE		BAWE	
Verb	Present	Past	Present	Past
show	24	20	64	32
suggest	8	3	8	0
demonstrate	2	2	3	0
Total	34	25	75	32

Based on the table above, it should be noted that out of the 44 occurrences of *result + show* in BrAWE, the present tense is used 24 times and the past tense is used 20 times. In BAWE, the 96 occurrences of these collocations are used 64 times in the present and 32 times in the past, representing half the frequency in the present, a phenomenon that is not observed in the BrAWE corpus. The concordance lines below illustrate an excerpt in the present and one in the past for each corpus:

1. *The **results show** that activities to guarantee supportable clean water supplies ought not to end with the development of a well (FILE PSEBTUT01I117) (BrAWE)*
2. *The **results showed** that in negative shots the success rate was 61,8%, in neutral shots they scored 73,7% and in positive shots the performance was 92,0% (FILE LSEBLUT04I58) (BrAWE)*
3. *These **results** clearly **show** that during the water deprivation test, there was failure to concentrate the urine resulting in a decreasing urine osmolality (FILE LS0047g) (BAWE)*

4. *The* ***results*** *also* ***showed*** *that there were no significant gains to be made from the installation of additional self-service machines (FILE SS 0232b) (BAWE)*

With the collocation *result + suggest,* there are eight occurrences in the present and three in the past in BrAWE. All occurrences of this collocation in BAWE (8) are used in the present. The concordance lines below help exemplify some uses of this collocation:

1. *The* ***results suggest*** *that the effectiveness of the harm reduction programs depends on the proper integration of different approaches (FILE LSEMDUT02I80) (BrAWE)*
2. *His* ***results suggested*** *that, but in my opinion, surface roughness indeed does decrease permeate flux, at least in the early stages of the fouling process, due mainly to the preferential ways over the surface of the membrane (FILE PSCRSWUT02I106) (BrAWE)*
3. *The* ***results suggest*** *that the Elliptical design offers the greatest strength, however it is expected that the large area at the tip would make insertion difficult, therefore the optimal designs were chosen from the remaining 5. (FILE PS 0250e) (BAWE)*

At last, *result + demonstrate* appears twice in the present and twice in the past tense in BrAWE, and three times in the present in BAWE, as in the examples below:

1. *The* ***results demonstrated*** *that the compressive strength of the concrete was increasing from the first test to the third test, as expected. (FILE PSMRLMUT01I73) (BrAWE)*

2. *Our **results demonstrate** that ACTB, DDX54 and PPIA were the most stable reference genes and TUBA, 18S rRNA and GAPDH were ranked as the least stable genes. (FILE LSRRNOUT01I153) (BrAWE)*
3. *The above **results demonstrate** impaired renal function, presence of increased urinary protein indicative of glomerular damage. (FILE LS0245f) (BAWE)*

The next subsection aims at discussing the collocations with the node *value.*

4.2.3 Collocations of *value*

This subsection covers the collocations composed of the node *value*. This noun is the third most frequent in BAWE out of the 125 nouns and the fourth most frequent in BrAWE. As shown in Table 6, the collocations with *value* as the node in BAWE are composed of 50 different collocates. In BrAWE, there are 32 different collocates. Out of these numbers, 25 collocates are statistically different, meaning that they are overused or underused in the Brazilian corpus. Table 16 contains the total amount of collocates of *value*.

Table 16.
Collocations of value in both corpora

		VALUE		
		BrAWE	BAWE	LL
modifier	high	27	68	5.13
	low	17	43	3.18
	final	0	40	-16.69**
	critical	12	39	0.7
	mean	12	32	1.86
	market	0	24	-10.01**
	extreme	0	22	-9.18**
	different	22	20	23.6*
	absolute	0	19	7.93*
	experimental	14	19	9.69*
	measured	0	18	-7.51**
	nutritional	0	18	-7.51**

	intrinsic	0	17	-7.09**
	negative	0	17	-7.09**
	net	0	16	-6.68**
	cultural	8	16	2.84
	good	0	16	-6.68**
	true	0	15	-6.26
	great	0	15	-6.26
	new	14	0	46.75*
	same	11	0	36.73*
	initial	9	0	30.06*
	normal	8	0	26.72*
	numeric	5	0	16.70*
object of	give	12	43	0.31
	use	19	43	4.98
	add	23	37	12.37*
	calculate	7	30	0
	find	22	28	16.56*
	obtain	19	26	13.01*
	determine	12	20	6.08
	show	16	20	12.32*
	compare	5	18	0.12
	create	6	18	0.56
	take	0	18	-7.51**
	provide	0	17	-7.09**
	reach	8	16	2.84
	affect	0	15	-6.26
	substitute	4	12	0.37
	increase	0	12	-5.01
	get	5	11	1.4
	represent	5	11	1.4
	estimate	5	10	1.77
	produce	0	10	-4.17
	reduce	0	9	-3.76
	set	0	7	-2.92
	exceed	1	6	-0.10
	have	27	0	90.17*
	present	13	0	43.41*
	consider	5	6	4.04
	put	4	4	3.94
subject of	indicate	3	12	0.01

	increase	0	8	-3.34
	use	0	6	-2.5
	match	0	4	-1.67
	fall	0	4	-1.67
	lie	0	4	-1.67
	appear	0	4	-1.67
	need	0	4	-1.67
Number of collocates		33	52	

*Overused. **Underused

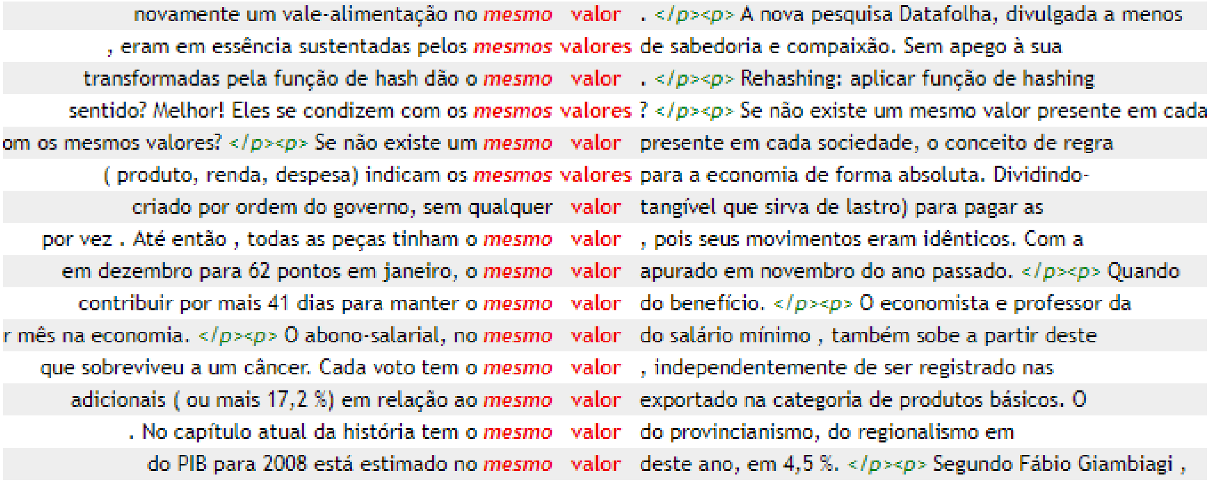
As can be seen in Table 16 *value* has 33 collocates in BrAWE, while in BAWE the number of collocates is 52. Among the 32 in BrAWE, *new*, *same*, *initial, normal, numeric, present, consider*, and *put* have zero occurrences in BAWE. Conversely, there are some collocates in the British corpus that are not used by Brazilians, as in the case of the verbs *increase*, *use*, *match*, *fall, lie*, *appear*, and *need* in the 'subject of' category. There are also other collocates with zero occurrences in BrAWE, but they are among the SSD ones, which are going to be addressed next.

From the 25 SSD collocates, 14 are overused and the remaining 11 are underused in BrAWE, either in the 'modifier' category of in the 'object of' category. Thus, there are no SSD verbs that collocate with the node *value* whenever it is used as a subject. In the 'modifier' category, *final, market, extreme*, *absolute, measured, nutritional, intrinsic, negative, net*, and *good* are the underused collocates, while *different*, *absolute*, *experimental, new, same, initial, normal,* and *numeric* are the overused ones. In relation to the underused collocates, they have zero occurrences in the Brazilian corpus.

When it comes to the overused collocates in the 'modifier' category, there are *different*, *experimental*, *new*, *same*, *initial*, *normal* and *numeric*, being the last five with zero occurrences in BAWE. Considering that these modifiers have zero occurrences in the general English corpus, they might either be related to the areas of expertise or, still, be due to language transfer.

In the 'object of' category, there are 8 SSD verbs that collocate with *value* as an object, eight being overused and two being underused in the BrAWE corpus. The overused verbs are: *add*, *find*, *obtain*, *show*, *have*, and *present*, while *take* and *provide* are underused by Brazilian students.

Taking all the overused collocates (14) into account, seven have zero occurrences in BAWE: *new, same, initial*, *normal* and *numeric* in the 'modifier' category, and *have,* and *present*, in the 'verb (object of)' category. Although it is not possible to prove that L1 interferes in the collocational choice for every overused collocates with zero occurrences in BAWE, 'mesmo' (RF = 11,267; NV = 0.00347) the cognate for *same* in Portuguese is used in the Portuguese corpus, as shown in the concordance lines below:



```
           novamente um vale-alimentação no mesmo  valor   . </p><p> A nova pesquisa Datafolha, divulgada a menos
       , eram em essência sustentadas pelos mesmos valores de sabedoria e compaixão. Sem apego à sua
        transformadas pela função de hash dão o mesmo  valor   . </p><p> Rehashing: aplicar função de hashing
     sentido? Melhor! Eles se condizem com os mesmos valores ? </p><p> Se não existe um mesmo valor presente em cada
om os mesmos valores? </p><p> Se não existe um mesmo  valor   presente em cada sociedade, o conceito de regra
      ( produto, renda, despesa) indicam os mesmos valores para a economia de forma absoluta. Dividindo-
          criado por ordem do governo, sem qualquer  valor   tangível que sirva de lastro) para pagar as
     por vez . Até então , todas as peças tinham o mesmo  valor   , pois seus movimentos eram idênticos. Com a
      em dezembro para 62 pontos em janeiro, o mesmo  valor   apurado em novembro do ano passado. </p><p> Quando
       contribuir por mais 41 dias para manter o mesmo  valor   do benefício. </p><p> O economista e professor da
r mês na economia. </p><p> O abono-salarial, no mesmo  valor   do salário mínimo , também sobe a partir deste
     que sobreviveu a um câncer. Cada voto tem o mesmo  valor   , independentemente de ser registrado nas
      adicionais ( ou mais 17,2 %) em relação ao mesmo  valor   exportado na categoria de produtos básicos. O
          . No capítulo atual da história tem o mesmo  valor   do provincianismo, do regionalismo em
           do PIB para 2008 está estimado no mesmo  valor   deste ano, em 4,5 %. </p><p> Segundo Fábio Giambiagi ,
```

Figure 17: Concordance lines of 'mesmo valor' in the Portuguese corpus

Considering the collocational range among the areas of expertise represented in BrAWE, the examples below illustrate that specific collocations are preferred by certain areas.

The collocation *add + value* is used in PS and SS areas of expertise, as shown in the excerpts below:

1. *In the model proposed is possible to check the activities that* ***add value*** *to the client and activities that does not add value. (FILE PSDSHUT06I98) (BrAWE)*
2. *The company has been developing a system called GMS (Global Manufacturing System) which intends to help the company to become more competitive and it uses the principles of Lean Manufacturing as eliminate wastes and* ***add value****. (FILE SSPOUT01I155) (BrAWE)*

The collocates *obtain*, *show*, *have*, *present* are employed by Brazilian students of the following areas: PS, LS and SS. The excerpts below exemplify one use of the collocation in each area of expertise:

<u>*obtain + value*</u>

1. *In order to* ***obtain a value*** *for this effort is crucial to develop a general strategy simulation. (FILE PSCRSHUT07I98) (BrAWE)*
2. *However, a subtraction of 180o from the* ***obtained value*** *was made, and this step is not predicted in the AKE test (FILE LSCR DUUT01I39) (BrAWE)*
3. *According to the* ***values obtained*** *in each indicator, improvements could be seen in all five categories of BPD. (FILE SSCRPOUT01I155) (BrAWE)*

<u>*show + value*</u>

1. *Since it is not an ideal differential amplifier, the output **shows** a small **value** (though not negligible) instead of zero due to the common-mode voltage amplified by a common-mode gain. (FILE PSDGCUT01I67) (BrAWE)*
2. *Although, the table 2, **shows** a Pearson correlation **value** of 0.46, so there is a moderate correlation between length of service and % cell damage and the p-value much lower than 0.05 indicates that the H1 of this test was accept, suggesting with a very evidence, that the two sets of data come from correlated population. (FILE LSPQMDUT11I80) (BrAWE)*
3. *... piled up into buns, feathers, classical stones from the continent and animal print/geometric patterns characteristic from the African culture, the brand chose to select mainly white models to **show** the **value** of the African black culture and this event might lead to a reflection about cultural appropriation. (FILE SSCRLLUT01I41) (BrAWE)*

<u>have + value</u>

1. *The Figure 1 shows that the semi-arid area in the Northeast Region **had** the highest irradiation **value** (brighter areas), reaching 6,5kWh/m2. (FILE PSCSMULT01I63) (BrAWE)*
2. *They found that teixobactin has a PD50 of 0.2 mg per kg that is better comparing with vancomycin (which is the choice of MRSA treatment) that **has** a **value** of 2.75 mg per kg. (FILE LSCRULUT01I183) (BrAWE)*
3. *Despite the criticism and limitations, the contingency approach **has** a practical **value** for the managers by providing a comprehensive framework that relates environment variables to the design of the organisational structure. (FILE SSCRLOUT02I77) (BrAWE)*

present + value

1. *The number of Basic elements (BELS) is exactly the same (1 see Figure 2(c)) and the timing results* ***present*** *the same* ***values****. (FILE PSEXQMLT02I91) (BrAWE)*
2. *Normally, the* ***values*** *are* ***presented*** *on tables, line graphs, histograms or figures in order to help the reader to verify the findings and understand them (Polit & Beck, 2014). (FILE LSCRSAUT01I163) (BrAWE)*
3. *From all the companies analysed, Mott MacDonald* ***presented*** *the most stable* ***values****, with ratios from all classes being inside medium parameters, neither high nor low. (FILE SSCRBLUT01I56) (BrAWE)*

The collocation *consider + value* is used five times and only by students from PS and LS areas of expertise, as shown below:

1. *Even considering the highest values of convergence, which would be the worst cases, both the deflection at the tip and the maximum equivalent stress satisfies the original design requirements (FILE PSDBSUT03I42) (BrAWE)*
2. *Lastly, after cash flow was calculated and accumulated, internal rates of return were found, considering cash flow values, to estimate the value of the company for both projects. (FILE LSPQNWUT05I184) (BrAWE)*

Put + value appears in PS, LS and AH, with only one occurrence in the latter. Below, the concordance lines help exemplify the uses:

1. *We can even confirm this by simply* ***putting*** *these* ***values*** *into Simulink and running a simulation, as described in the following figure. (FILE PSEXLLUT02I140)*

2. *The dataset of* ***values*** *was* ***put*** *into a table (Figure 16) and the resulting polynomial function (Figure 17) was used the get the values of V T for each step. (FILE LSMRGSUT02I40)*
3. *The king's personality apparently is one that* ***puts*** *a great* ***value*** *in one's own word (FILE AHERHUT04I45) (BrAWE)*

4.2.4 Collocations of *figure*

This subsection aims at analyzing the collocates that go with *figure* as the node. *Figure* is the fourth most frequent node in BAWE and only the 24th most frequent node in BrAWE. Table 17 contains the total amount of words that collocate with the node *figure*:

Table 17.
Collocations of figure in both corpora

		FIGURE		
		BrAWE	BAWE	LL
modifier	above	0	9	-3.76
	following	0	9	-3.76
	sales	0	7	2.92
	overall	0	6	-2.5
	actual	0	5	-2.09
	high	0	5	-2.09
	profit	0	4	-1.67
	performance	0	4	-1.67
	current	0	4	-1.67
object of	see	29	142	-0.4
	give	0	8	-3.34
	quote	0	5	-2.09
	show	0	5	-2.09
	use	6	5	6.97*
subject of	show	3	23	-1.02
Number of collocates		3	15	

* Overused

As shown in Table 17, there are 15 collocates in the British corpus, and only three in the Brazilian corpus. Among the collocates in BAWE, only *see, show,* and *use* also appear in BrAWE. Additionally, the modifiers are exclusively used in BAWE.

Interestingly, the verb *use* in the 'object of' category is the only statistically different, with a raw frequency of six in BrAWE and five in BAWE, respectively. Therefore, differently from the previous nodes analyzed, there are no underused collocates with this node. Further comments on the only SSD collocate are presented below:

All the six occurrences of *use + figure* in BrAWE are in PS assignments; four are used in the active voice, and two are used in the passive voice, as in:

1. ***Using** the **figure** 8 it was also possible to find the core length for k-? and k-e simulations. (FILE PSEXPSTUT01I167) (BrAWE)*
2. *if Zo > RL is going to be **used figure** (b), as can be seen above (FILE PSQMLT09I09) (BrAWE)*

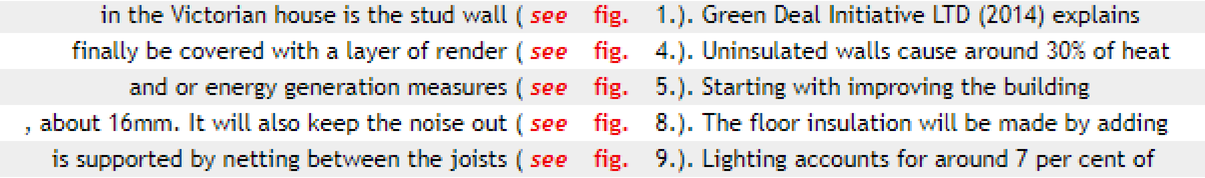
The five occurrences of *use + figure* in BAWE appear in PS and SS assignments and in the passive voice, such as:

1. *This **figure** is also **used** to assist the cash flow analysis. The improvement may be led by better credit control, fewer potential bad debts, or fewer occurrences of payment delays. (FILE PS0166b) (BAWE)*
2. ***Figures** are **used** well, supplementing the information written in the article. </p><p> The experiments undertaken for this journal are of a suitable number and investigate all possibilities of mechanism described by the article. (FILE PS0388g) (BAWE)*

Although *see* is not SSD, it will be analyzed as well because of its much higher frequency than the other collocates with the node *figure*. There are 29 occurrences of the collocation *see* + *figure* in BrAWE, and 142 occurrences in BAWE. In both corpora, this collocation is used between parentheses to indicate a figure containing the information the author is referring to. However, there are three occurrences in BrAWE that does not follow this pattern, two of them in the same text, as in

1. *Each alphabet, **see figure** 5, consisted of the uppercase letters A-Z, the numbers 1-9 and the punctuation comma and dot. (FILE PSSURT01I128)*
2. *if we also look at the minimum period of it we'll **see** the same **figure** for it (3.710ns) and the clock signal (clk), **see Figure** 5(e). (FILE PSQMLT02I91)*

Furthermore, another difference is that in BrAWE *figure* appears five times abbreviated as ‘fig’, all of them written by the same student:



in the Victorian house is the stud wall (	see	fig.	1.). Green Deal Initiative LTD (2014) explains
finally be covered with a layer of render (	see	fig.	4.). Uninsulated walls cause around 30% of heat
and or energy generation measures (	see	fig.	5.). Starting with improving the building
, about 16mm. It will also keep the noise out (	see	fig.	8.). The floor insulation will be made by adding
is supported by netting between the joists (	see	fig.	9.). Lighting accounts for around 7 per cent of

Figure 18. see + fig. in BrAWE.

4.2.5 Collocations of *process*

Process is the fifth most frequent noun in BAWE out of the 125 nodes, and the third most frequent one in BrAWE. It appears as the node accompanied by 56 different collocates in BAWE, and 31 collocates in BrAWE. The collocates are displayed in Table 18.

Table 18.
Collocations of process in both corpora

PROCESS				
		BrAWE	BAWE	LL
modifier	development	0	57	-23.78**
	business	8	35	0
	production	18	34	7.21*
	whole	8	31	0.07
	decision-making	0	29	-12.1**
	manufacturing	19	24	14.44*
	political	0	20	-8.34**
	new	0	19	-7.93**
	natural	5	17	0.21
	decision	0	15	-6.26
	stage	0	15	-6.26
	other	0	15	-6.26
	selection	0	13	-5.42
	internal	0	13	-5.42
	same	12	0	40.07*
	creative	7	0	23.38*
	building	6	0	20.04*
	complex	6	0	20.04*
	continuous	5	0	16.70*
verb (object of)	make	10	51	-0.25
	use	15	19	11.36*
	base	0	18	-7.51**
	influence	3	13	0
	explain	5	12	1.11
	know	0	11	-4.59
	apply	0	10	-4.17
	call	3	10	0.15
	facilitate	0	9	-3.76
	describe	5	9	2.2
	develop	5	9	2.2
	show	4	9	1.06
	improve	13	7	20.44*
	involve	0	7	-2.92
	find	0	7	-2.92
	speed	0	6	-2.5
	aid	0	6	-2.5
	discuss	0	6	-2.5

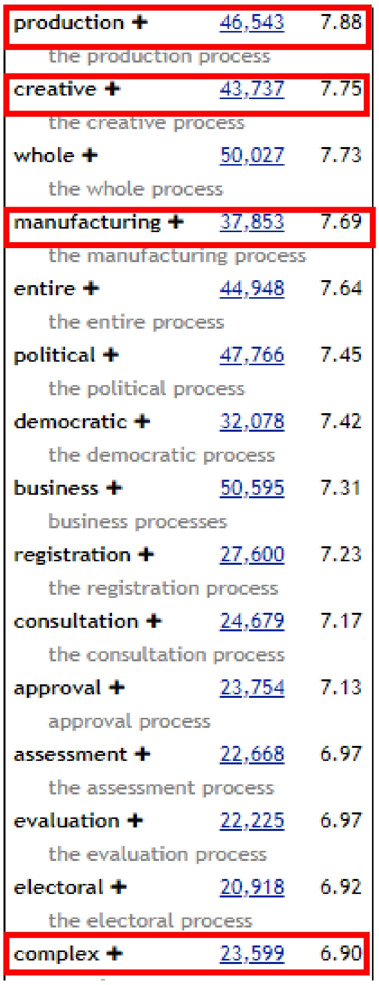
	indicate	0	6	-2.5
	determine	0	6	-2.5
	follow	4	6	2.4
	affect	0	6	-2.5
	provide	0	6	-2.5
	streamline	0	5	-2.09
	do	6	0	20.04*
	repeat	5	5	4.92
	start	5	5	4.92
	understand	5	4	6.00
verb (subject of)	involve	0	14	-5.84
	occur	8	13	4.23
	take	3	10	0.15
	include	4	8	1.42
	start	3	7	0.72
	need	3	6	1.06
	use	0	6	-2.5
	wait	0	5	-2.09
	work	0	5	-2.09
	repeat	0	4	-1.67
	depend	0	4	-1.67
	influence	0	4	-1.67
	require	7	4	10.63*
	affect	0	4	-1.67
	consist	5	3	7.36*
Number of collocates		31	56	

*Overused. **Underused

Among the collocates in BAWE, *development, decision-making, political, new, decision, stage, other, selection, internal, base, know, apply, facilitate, involve, find, speed, aid, discuss, indicate, determine, affect, provide, streamline, involve, use, wait, work, repeat, depend, influence,* and *affect* have zero occurrences in BrAWE. On the other hand, considering the collocates only used by Brazilians we have *same, creative, building, complex, continuous*, and *do.*

Moreover, there are 17 SSD collocates, five being underused and 12 being overused in BrAWE. In the ‘modifier’ category, four collocates are underused: *development*, *decision-making*, *political*, and *new*. The remaining SSD collocates in this category (7) are overused: *production*, *manufacturing*, *same*, *creative*, *building*, *complex*, and *continuous*. With the exception of *production* and *manufacturing* (two of the overused collocates), the cases of underuse have zero occurrences in BrAWE, while the cases of overuse have zero occurrences in BAWE.

When analyzing only the cases of overuse and comparing them to the collocates of *process* in the general English corpus, *production*, *manufacturing*, *creative*, and *complex* are the modifiers in common.

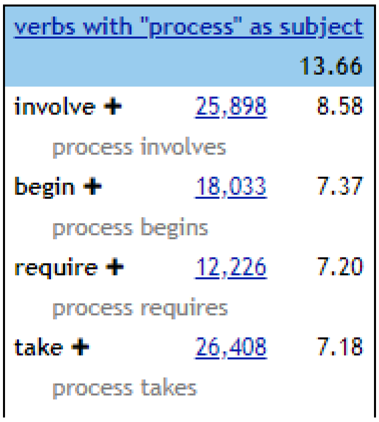


production +	46,543	7.88
the production process		
creative +	43,737	7.75
the creative process		
whole +	50,027	7.73
the whole process		
manufacturing +	37,853	7.69
the manufacturing process		
entire +	44,948	7.64
the entire process		
political +	47,766	7.45
the political process		
democratic +	32,078	7.42
the democratic process		
business +	50,595	7.31
business processes		
registration +	27,600	7.23
the registration process		
consultation +	24,679	7.17
the consultation process		
approval +	23,754	7.13
approval process		
assessment +	22,668	6.97
the assessment process		
evaluation +	22,225	6.97
the evaluation process		
electoral +	20,918	6.92
the electoral process		
complex +	23,599	6.90

Figure 19: modifiers of *process* in the general English corpus

Regarding the verbs that collocate with *process* as the object of the sentence, *use*, *improve*, *do*, *repeat*, *start* and, *understand* are overused, while *base* is the only underused collocate in this category. Again, with the exception of *use* and *improve*, the remaining SSD verbs have zero occurrences in BrAWE (cases of underused collocates), and zero occurrences in BAWE (cases of overused collocates).

Require and *consist* are the two SSD verbs that collocate with *process* in the 'subject of' category. Both of them are overused in BrAWE and *consist* has no occurrences in BAWE. The overuse of the collocations *process* + *require* might be explained by the fact that this verb has the third higher MI score in the general English corpus (Figure 20). Thus, once again, the overuse of this specific collocation is a consequence of transferring the collocational knowledge from the general language to academic English registers.



verbs with "process" as subject		
		13.66
involve +	25,898	8.58
process involves		
begin +	18,033	7.37
process begins		
require +	12,226	7.20
process requires		
take +	26,408	7.18
process takes		

Figure 20: Verbs that collocate with *process* as a subject in the general English corpus

4.2.6 Collocations of *analysis*

In this subsection, the collocates that accompany the node analysis will be described. This node is the fifth most frequent in BrAWE and 13th most frequent in BAWE. There are 23 collocates in the Brazilian corpus, and 35 in the British corpus. Table 19 contains the total amount of collocates with the node under discussion:

Table 19
Collocations of analysis in both corpora

	ANALYSIS			
		BrAWE	BAWE	LL
modifier	statistical	18	30	9.12*
	network	0	29	-12.1**
	detailed	0	28	-11.68**
	far[46]	0	26	-10.85**
	critical	0	20	-8.34**
	data	4	19	-0.03
	regression	6	13	1.76
	above	0	12	-5.01
	in-depth	0	11	-4.59
	thorough	0	11	-4.59
	cost-benefit	0	9	-3.76
	quantitative	13	7	20.44*
	first	6	0	20.04*
	other	5	0	16.70*
	qualitative	4	6	2.40
	deep	4	0	13.36*
	complete	4	0	13.36*
verb (object of)	perform	18	35	6.79*
	use	11	26	2.55
	carry	3	22	-0.85
	provide	7	20	0.82
	conduct	9	14	5.11
	require	5	9	2.2
	undertake	0	7	-2.92
	base	3	7	0.72
	show	4	6	2.4
	make	22	6	46.88*
	run	0	5	-2.09
	allow	3	5	1.52
	set	0	4	-1.67
	include	0	4	-1.67
	do	10	0	33.40*
verb (subject of)	show	9	29	0.55
	suggest	0	9	-3.76
	use	0	8	-3.34

[46] The lemma far is used as ‘further’ in the texts.

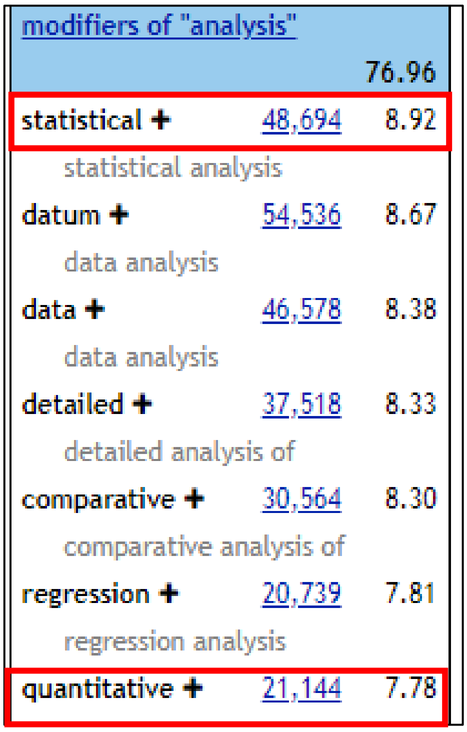
	reveal	2	7	0.06
	enable	0	4	-1.67
	indicate	0	4	-1.67
	follow	0	4	-1.67
	take	0	4	-1.67
	make	4	0	13.36*
Number of collocates		23	35	

*Overused. **Underused

Based on the table above, it should be noted that *network, detailed, far, critical, above, in-depth, thorough, cost-benefit, undertake, run, set, include, suggest, use, reveal, enable, indicate, follow,* and *take* are exclusively used in BAWE. Nevertheless, *first, other, deep, complete, do, and make* are only used by Brazilians.

There are 14 collocates whose differences in frequency are statistically significant, out of which 10 are overused and four are underused by Brazilians. The overused collocates are: *statistical*, *quantitative*, *first*, *other*, *deep,* and *complete* in the 'modifier' category; *perform, make,* and *do* in the 'verb (object of)' category; and *make* in the 'verb (subject of)' category. When it comes to the underused collocates, *network*, *detailed, far* and *critical* are SSD, all in the 'modifier' category.

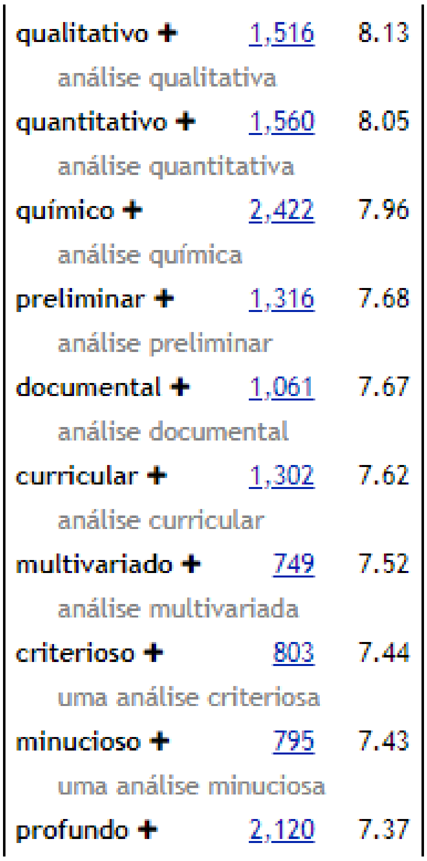
When analyzing the Word Sketch in the general English corpus for the node *analysis*, *statistical* and *quantitative* collocate as the modifiers as well.



modifiers of "analysis"		76.96
statistical +	48,694	8.92
statistical analysis		
datum +	54,536	8.67
data analysis		
data +	46,578	8.38
data analysis		
detailed +	37,518	8.33
detailed analysis of		
comparative +	30,564	8.30
comparative analysis of		
regression +	20,739	7.81
regression analysis		
quantitative +	21,144	7.78

Figure 21: Modifiers of *analysis* in the general English corpus

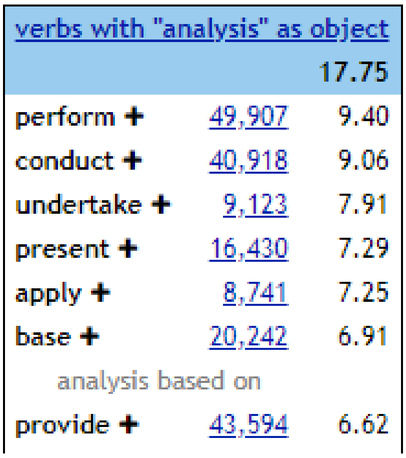
Eight of the overused collocates are absent from BAWE: *first*, *other*, *deep*, *complete*, *do*, and *make.* The overuse of *quantitative* and *deep* by Brazilians might also be explained due to L1 transfer. As shown in Figure 22, ‘quantitativa’ (cognate for *quantitative*) and ‘profunda’ (equivalent for *deep*), are among the Portuguese words that collocates with the node ‘análise’.



qualitativo +	1,516	8.13
análise qualitativa		
quantitativo +	1,560	8.05
análise quantitativa		
químico +	2,422	7.96
análise química		
preliminar +	1,316	7.68
análise preliminar		
documental +	1,061	7.67
análise documental		
curricular +	1,302	7.62
análise curricular		
multivariado +	749	7.52
análise multivariada		
criterioso +	803	7.44
uma análise criteriosa		
minucioso +	795	7.43
uma análise minuciosa		
profundo +	2,120	7.37

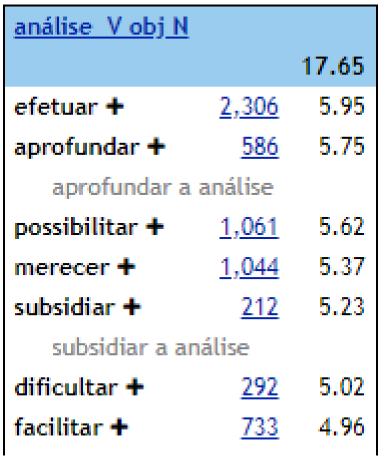
Figure 22: modifiers of *analysis* in the Portuguese corpus

Perform is the verb with the highest MI score in the general English corpus, as shown in Figure 23. Besides, ‘efetuar’, which is a possible equivalent for *perform* in Portuguese, is the verb that most strongly collocates (MI = 5.95) with ‘análise’ (the cognate for *analysis*) in the Portuguese corpus (Figure 24).



verbs with "analysis" as object		
		17.75
perform +	49,907	9.40
conduct +	40,918	9.06
undertake +	9,123	7.91
present +	16,430	7.29
apply +	8,741	7.25
base +	20,242	6.91
analysis based on		
provide +	43,594	6.62

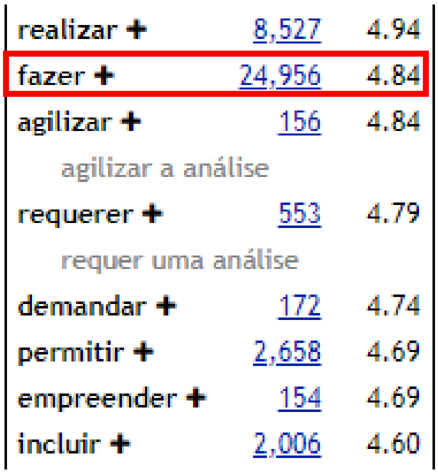
Figure 23: perform + analysis in the general English corpus



análise V obj N		
		17.65
efetuar +	2,306	5.95
aprofundar +	586	5.75
aprofundar a análise		
possibilitar +	1,061	5.62
merecer +	1,044	5.37
subsidiar +	212	5.23
subsidiar a análise		
dificultar +	292	5.02
facilitar +	733	4.96

Figure 24: 'efetuar' in the Portuguese corpus

Furthermore, it should also be highlighted that the verb 'fazer' (a possible equivalent for *do*) as an object collocates with *analysis* in the Portuguese corpus as well:



realizar +	8,527	4.94
fazer +	24,956	4.84
agilizar +	156	4.84
agilizar a análise		
requerer +	553	4.79
requer uma análise		
demandar +	172	4.74
permitir +	2,658	4.69
empreender +	154	4.69
incluir +	2,006	4.60

Figure 25: 'fazer' collocating with *analysis* in the Portuguese corpus

The next section aims at both answering the research questions, and discuss some patterns found throughout this corpus-based study.

4.3 Answering the research questions and discussing some findings

This section aims to discuss the main findings of this investigation. In order to do that, the research questions will be answered individually.

1- *Is there a statistically significant difference in the frequency of the nodes in BAWE and BrAWE?*

Data reveals that out of the 125 nodes analyzed, 89 show statistically significant difference, out of which 49 are underused in BrAWE and the remaining 40 are overused (Table 6). Additionally, the top five nodes in both corpora are somehow different. In BAWE, *system*, *result*, *value*, *figure* and *process* are the most frequent nodes. In BrAWE, however, the five most frequent nodes are *system*, *result*, *process*, *value* and *analysis*. As stated in section 4.1, *figure*, which is the fourth most frequent node in BAWE, occupies only the 24th position in BrAWE. Conversely, *analysis*, among the top five nodes in BrAWE, is the 13th most frequent node in BAWE. Moreover, *data* and *example* have an intriguing behavior, as their frequencies in the corpora are strikingly different. *Data* is in the 10th in BAWE and only 98th in BrAWE, while *example* is in the 64th most frequent node in BAWE, and only the 12th most frequent in BrAWE.

This noteworthy difference in the position occupied by *data* and *example* in both corpora may be explained by two reasons. When it comes to *data*, students in BAWE vary the words that collocate with the node. On the other hand, the variety of collocations that have *data* as the node word in BrAWE is much more restricted. This pattern of NNS repeating favored items and NS using a wider range of collocations is found in Durrant and Schmitt (2009), Howarth (1998) and Simpson-Vlach, Ellis and Maynard (2008). This difference may also suggest a difference in the content of the texts, as far as the higher presence of the word *data* can be related to contents more evidence based. Regarding the node *example*, it is among the top 15 most frequent nodes in BrAWE and only the 64th position in BAWE due to a L1 transfer (Howarth, 1998), meaning that Brazilian students generally use collocations with this specific node in Portuguese ('por exemplo'), and transfer this collocational knowledge into academic registers.

2- *Is there a statistically significant difference in the frequency of the collocates of these nodes in BAWE and BrAWE? If so, does this difference indicate overuse or underuse? Is it possible to identify the motivations for such differences?*

A positive answer can be provided. In quantitative terms, there is a statistically significant difference in the frequencies of collocations in the two academic corpora at stake. From the six nodes that were analyzed, - *system*, *result*, *value*, *figure* and *process*, 164 collocates emerged in BrAWE and 262 in BAWE, out of which 94 are SSD, i.e. 67 are overused, and 27 are underused by Brazilians in comparison to the BAWE. When observing the nodes individually, it should be noted that the number of overused collocates is higher than the number of underused collocates in the six nodes analyzed.

Table 20.
Number of overused and underused collocates in the six nodes

Node	Number of overused collocates	Number of underused collocates
system	7	6
result	23	1
value	14	11
figure	1	0
process	12	5
analysis	10	4
TOTAL	67	27

Based exclusively on the collocates that emerge from the six nodes discussed in this investigation, Brazilian students apparently transfer collocations they already know and use in general contexts into academic written registers. For instance, the collocations *create* + *system*, *result* + *show*, *result* + *suggest*, *result* + *demonstrate*, *process* + *require*, and *perform* + *analysis* are all overused in BrAWE. These collocates (*create*, *show*, *suggest*, *demonstrate*, *require* and *perform*) have a high MI in the general English corpus. Hence, students from BrAWE seem to master the collocations at stake in situations where general English is required, and transfer this collocational knowledge into academic registers. Regarding the pair *create* + *system* and *design* + *system*, students could have chosen to use them interchangeably whenever they are referring to the process of starting something new in order to vary the text and enrich it in terms of variety of lexical items. This argument is sustained by Howarth (1998), who points out that L2 students' texts present a lower density of combinations, which seems to be the case when it comes to the collocations of *system*.

A similar case can be observed with the node *result*. Thus, instead of overusing the modifiers *good* and *positive*, Brazilian students could have chosen *accurate* and *interesting* more often, as they are the collocates with occurrences in BrAWE and BAWE and did not came up as SSD in the comparison of the frequencies in both corpora. Once again, this outcome corroborates Howarth (1998) who explains that selecting conventional phraseologies is a challenge even for advanced learners of English. That is to say that even advanced learners tend to use fewer combinations of words in a way that these specific combinations are usually overused by learners.

Following this same line of thought, regarding the node *value*, with the exception of *final*, *absolute*, *negative* and *good*, the other six collocates underused in BrAWE – *market*, *extreme*, *measured*, *nutritional*, *intrinsic*, and *net* – have a more specialized meaning, even though they are used in at least two areas of expertise in BAWE. Considering that these collocates have zero occurrences in BrAWE and, thus, are underused in this corpus, the argument in Simpson-Vlach, Ellis and Maynard (2008) is validated, since the authors claim that NS use a wider range of collocations, thus varying the combinations of words and enriching the texts.

Moreover, another phenomenon common to learners of English is the underuse of collocations with a high MI score in corpus that follow native norms (Durrant & Schmitt, 2009). For instance, the modifier that collocates with the node *process* and has the highest MI in BAWE is *development* (MI= 9.56). On top of that, this collocation is underused by the Brazilians, corroborating the pattern found by Durrant and Schmitt (2009).

Furthermore, along with transference from general language to academic writing contexts, the findings of this study indicate a strong influence of L1, in this case, the Portuguese language. According to Gitsaksi (1999), Laufer and Waldmann (2011) and Selistre (2010), the latter focusing on Brazilians writing academic texts as well, L1 influencing L2 production is a recurring issue faced by learners of a second language. For instance, with the node *system*, Brazilians overuse the collocation *create + system* possibly because they are acquainted with it in Portuguese, i.e. ‘criar’ has a high MI score in the Portuguese corpus, meaning that this verb and ‘sistema’ have a high-strength relationship. Vestiges of Portuguese are also observed in the overuse of the collocate *demonstrate* in the ‘verb (subject of)’ category with the node *result*, since ‘demonstrar’ is the verb that second best collocates with ‘resultado’ in the Portuguese corpus (Figure 15).

In the next chapter, the final remarks are provided along with suggestions of collocations for ColloCaid's database.

Chapter 5: Final remarks

In this chapter, a brief summary of this investigation and its main limitations will be provided. Also, contributions and pedagogical implications will be discussed. Finally, further research on the field of academic writing will be suggested, specifically regarding how Brazilian students produce academic collocations in English.

This corpus-based study aimed at discussing the use of collocations by Brazilians studying in British universities. Thus, based on the 125 nouns listed in Frankenberg-Garcia et al. (2018), a comparative analysis of collocations in two corpora – the Brazilian Academic Written English Corpus (BrAWE; Goulart, 2017) and the British Academic Written English (BAWE; Alsop & Nesi, 2009) was conducted. After providing the reader with a literature review of what is understood by academic language, collocations, and previous studies on the subject, the analysis was based on the following definition of collocation created for the purposes of this research:

> a combination of two words that are associated due to statistical probabilities of occurring together

This definition is determined by the differences in the frequencies measured with the Log-likelihood calculator. Hence, if the result of the LL test is 6.63 or higher, there is a 99% chance of accuracy in the results, meaning that they are not random (p<0.01). The co-occurrence of two words is necessarily formed by a node - one of the 125 nouns (Frankenberg-Garcia et al., 2018) and one of the three categories below, as indicated in Chapter 3:

- ✓ Modifier: adjectives that come before the node
- ✓ Verb (object of): used when the node is the object of the verb

- ✓ Verb (subject of): used when the node is the subject of the verb

The outcomes of this corpus-based study show some statistically significant differences not only in the frequencies of the nodes in both corpora, but also in the individual collocates.

On the assumption that difficulties in producing collocations are related to lack of language exposure (Orenha-Ottaiano, 2015), it was possible to conclude that because there is a high incidence of transferring collocations from the general language into academic writing, it is necessary to teach Brazilian students less frequent collocates in general academic language. Furthermore, the findings reveal L1 transference (Gitsaksi, 1999; Laufer & Waldmann, 2011; Selistre, 2010), in this case, from the Portuguese language. Data revealed that the density of collocations in BrAWE is smaller. Thus, by increasing the students' repertoire in academic collocations, their writing will consequently improve, sounding formulaic and fluent. Considering a context of EAP teaching, collocates that did not came up as being SSD in the comparison between BrAWE and BAWE do not need to be emphasized. Nevertheless, collocations such as *design + system, measured + value, good + value, decision-making + process, detailed + analysis,* and *further + analysis* are worth teaching for Brazilian students.

With the intention of expanding ColloCaid's database, the outcomes of this study can be useful for more suggestions of collocations specifically addressed to help Brazilian users. ColloCaid was designed to serve as text editor that suggests collocations as the user types the text. Based on ColloCaid's system, whenever the user types one of the nodes of the collocations that are already inserted in the database, a variety of collocates shows up. Below, some collocates are suggested for each node analyzed. Only the underused collocates in BrAWE are suggested to be included in the mentioned tool[47]. For instance, with the node system, the collocates *new*, *reward*, *current*, *computer*, *communication*, and *design* would be suggested. Figures 26 and 27 portray how the tool would come up with the collocates when the user types the node *system*. As can be seen in Figure 26, if the user types the node *system*, it automatically gets underlined and suggestions of collocates appear after clicking on the node (Figure 27). In order to use that specific collocate, it is necessary to click on it to be incorporated into the text.

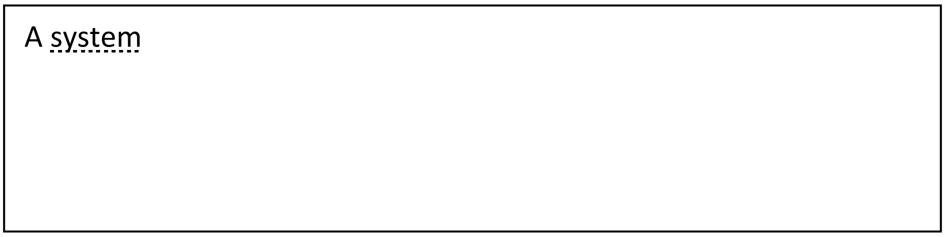


Figure 26. Node *system* in ColloCaid

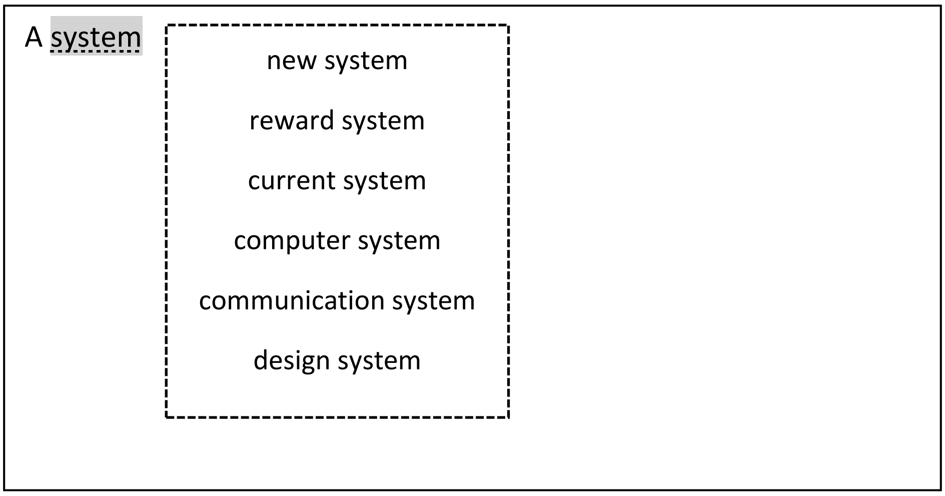


Figure 27. How the tool would suggest collocates

[47] There are no cases of underused collocates with the node *figure*.

After clicking on the collocation, the tool suggests a context that illustrates how the collocation can be used, as shown in Figure 28, in which the collocation new + system is inserted in a broader context:

> It is argued that employing a **new system** may involve a hefty financial cost and resources in terms of staff training.

Figure 28: Use of the collocation in context

Further suggestions of collocations composed of the nodes analyzed are given below:

Result

- ✓ verb (object of): *yield*

Value

- ✓ modifier: *final, market, extreme, measured, nutritional, intrinsic, negative, net, good*
- ✓ verb (object of): *take, provide*

Process

- ✓ modifier: *development, decision-making, political, new*
- ✓ verb (object of): *base*

Analysis

- ✓ modifier: *network, detailed, far, critical*

Concerning the limitations of this study, only the five most frequent nodes in BAWE and in BrAWE were chosen to be qualitatively analyzed. Thus, it would be desirable to undertake a qualitative research for all of the 125 nodes in order to best illustrate the preferred collocates.

In order to expand the understanding of how Brazilians use collocations in academic English written texts, it would be valuable to analyze collocations in one specific register across different disciplinary groups. Thus, instead of investigating the use of collocations in all genre families (Gardner & Nesi, 2013) that constitute BAWE and BrAWE, a narrower focus exclusively on essays, for instance, would allow for a better description of what is being produced by students when writing this specific register. To enable this kind of research, more texts should be added to BAWE so that statistical analyses are possible.

This investigation on academic English collocations in Brazilian L2 writing sheds light on the discussion about the internationalization of higher education, as English is the language for production and dissemination of knowledge (Baumvol, 2018). Even with scholars fighting against English imperialism (Pennycook, 1994), it is known that publications that are not written in English reach considerable fewer readers. As a consequence, the audience gets very limited, and scholars from marginalized countries become even more marginalized by not having a chance to participate in important scientific discussions. If the intention is to change this scenario, it is time to master academic writing skills in English and place Brazil among the important producers of knowledge internationally.

Formulaic sequences provide fluency and conventionality to the language. Considering that learning to write entails knowing how to use collocations properly, it is mandatory that these elements gain space in English teaching environments. (AlHassan & Wood, 2015; Li & Schmitt, 2009; Martinez & Schmitt, 2012). Based on the comparison of the two corpora at stake – BAWE and BrAWE – it is noted that academic collocations are not fully mastered by Brazilian students who write academic texts yet. For Sinclair (1991), learners operate on the open choice principle more than on the idiom principle. As a consequence, they may produce collocations that do not sound natural, i.e. not fluent. This lack of collocational competence was observed in the amount of outcomes that came up with statistically significant differences in the comparison between the data in the corpora.

Having this in mind, it is our role, as both EAP teachers and researchers, to contribute to the area and to help Brazilian learners of English to improve their academic writing skills. As pointed out by Hyland and Hamp-Lyons (2002, p. 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 ever more strongly informed foundations for pedagogic materials.". If the ideal scenario is to have more teaching of English in English and through English (Gardner, 2012), focusing on collocations is a good starting point. Some suggestions are given by Nesselhauf (2005, p. 253), for whom teaching collocations should begin with making students aware of this phenomenon. More than that, "It is essential that learners recognize that there are combinations that are neither freely combinable nor largely opaque and fixed (such as idioms) but that are nevertheless arbitrary to some degree and therefore have to be learnt."

Hopefully, this study will spur further research in the promising field of collocations.

References

Abreu-e-Lima, D. M.; Moraes Filho, W. B.; Barbosa, W. J. C.; Blum, A.S. (2016). O Programa Inglês sem Fronteiras e a política de incentivo à internacionalização do ensino superior brasileiro. In: Sarmento, S.; Abreu-e-Lima, D. M.; Moraes Filho, W. B. *Do Inglês sem Fronteiras ao Idiomas sem Fronteiras: A construção de uma política linguística para a internacionalização*. Editora UFMG, Belo Horizonte.

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.

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of studies in international education*, *11*(3-4), 290-305.

Ammon, U. (2006). Language planning for international scientific communication: An overview of questions and potential solutions. *Current issues in language planning*, *7*(1), 1-30.

Baisa, V. & Suchomel, V. (2014) SkELL: Web interface for English language learning. In Horák, A. & Rychlý, P. (ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing*. Karlova Studánka, Czech Republic, 5–7 December, 63–70.

Baumvol, L. K. (2018). *Language practices for knowledge production and dissemination: the case of Brazil*. PhD. Thesis. Universidade Federal do Rio Grande do Sul.

Berber Sardinha, T. (2004). *Lingüística de corpus*. São Paulo: Manole

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.

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

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1), 2-20.

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

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

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

Brasil. (2014). *Portaria 973*. Brasilia: MEC. Retrieved from <http://isf.mec.gov.br/ingles/images/pdf/novembro/Portaria_973_Idiomas_sem_Fronteiras.pdf.>.

Brasil. (2015). *Universidades NucLi*. Brasilia: MEC. Retrieved from: <http://isf.mec.gov.br/ingles/images/2015/janeiro/Universidades_NucLi_2015_novo.pdf>.

Charles, M. (2013). English for Academic Purposes. Paltridge, B., & Starfield, S. (Eds.). (2013). *The handbook of English for specific purposes* (Vol. 592). West-Sussex: Wiley-blackwell, 137-153

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), 403-426.

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

Coffin, C., Curry, M. J., Goodman, S., Hewings, A., Lillis, T., & Swann, J. (2003). *Teaching academic writing: A toolkit for higher education*. Routledge.

Conrad, S. (2002). 4. Corpus linguistic approaches for discourse analysis. *Annual review of applied linguistics*, *22*, 75.

De Cock, S.; Granger, S.; Leech, G; & McEnery, T. (1998) An automated approach to the phrasicon of EFL learners. In Granger, S. *Learner English on computer*, Routledge, 89-101.

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. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Flowerdew, J. (2019). The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. *Language Teaching*, *52*(2), 249-260.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second Language Learning, Teaching and Testing* (pp. 75–93). Harlow, UK: Longman.

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.

Gardner, S. (2012). Global English and bilingual education. In M. Martin-Jones, A. Blackledge, & A. Creese (Eds.), *The Routledge Handbook of Multilingualism* (pp. 247-263). London: Routledge.

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.

Gitsaki, C. (1999). Teaching English collocations to ESL students. *NUCB journal of language culture and communication*, *1*(3), 27-34.

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. (Eds.). (2002). International corpus of learner English. Louvain: Presses Universitaires de Louvain. http://www.uclouvain.be/en-cecl-icle.html.

Granger, S., C. Sanders & U. Connor. n.d. LOCNESS: Louvain Corpus of Native English Essays. https://www.uclouvain.be/en-cecl-locness.html [

Granger, S. & Paquot, M. (2015) Electronic lexicography goes local: Designs and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1): 118–141.

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

Hamp-Lyons, L. (2001) English for Academic Purposes. In: Nunan, D., & Carter, R. (Eds.). *The Cambridge guide to teaching English to speakers of other languages*. Ernst Klett Sprachen, 126-130.

Heuboeck, A., Holmes, J., & Nesi, H. (2010). *The BAWE corpus manual* (version III).

Hill, J. (2000). *Collocational Competence.* ETP English Teaching Professional.

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

Hughes, R. (1996). *English in speech and writing: Investigating language and literature*. Routledge.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

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

Hyland, K. (2016a). *Academic publishing: Issues and challenges in the construction of knowledge*. Oxford: Oxford University Press.

Hyland, K. (2016b). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, 31, 58–69.

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.

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

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

Mayor, M. (2013) *Longman collocations dictionary and thesaurus*. Harlow: Pearson Education.

McEnery*, T.;* Wilson*, A. (*1996*). Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T.; Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

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

Orenha-Ottaiano, A. (2015). Collocations workbook: um material de apoio pedagógico on-line baseado em corpus para o ensino de colocações em inglês. *Revista de Estudos da Linguagem*, *23*(3), 833-881.

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

Pennycook, A. (1994). *The cultural politics of English as an international language*. London: Longman.

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

Scarcella, R. (2003). *Academic English: A conceptual framework*. Technical Report. The University of California.

Selistre, I. C. T. (2010). Colocações, transferência linguística e elaboração de dicionários bilíngues escolares (inglês/português–português/inglês. *Acta Scientiarum. Language and Culture*, *32*(2), 271-278.

Shimohata, S., Sugio, T., & Nagata, J. (1997). Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the eighth conference on*

European chapter of the Association for Computational Linguistics (pp. 476-481). Association for Computational Linguistics.

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 University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied linguistics, 21*(4), 463-489.

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 A

https://drive.google.com/drive/folders/10YeTZtuhCaYvD9WtPLFDADG-0NJ-1y07?usp=sharing

This folder contains five Google Sheets organized according to the ranking of the most frequent to the least frequent node in BAWE. In order to facilitate visualization, the nodes were gathered in groups of 25. Thus, the names of the spreadsheets are *1-25*, *26-50*, *51-75*, *76-100* and *101-125*.