http://dx.doi.org/10.22168/2237-6321-21490

A quantitative analysis of collocations in Brazilian and British students' academic writing

Uma análise quantitativa de colocações em produções acadêmicas de estudantes brasileiros e britânicos

Marine Laísa MATTE (UFRGS)
marine.matte@ufrgs.br
Rozane Rodrigues REBECHI (UFRGS)
rozane.rebechi@ufrgs.br

Recebido em: 20 de jan. de 2019.
Aceito em: 15 de maio de 2019.

MATTE, Marine Laísa; REBECHI, Rozane Rodrigues. A quantitative analysis of collocations in Brazilian and British students' academic writing. **Entrepalavras**, Fortaleza, v. 9, n. 2, p. 195-213, maio-ago/2019.

Abstract: Language is formulaic in nature, which means that appropriate writing goes beyond knowing isolated words. The objective of this paper is to analyze the quantitative difference in the use of collocations of the Academic Collocation List (ACL) in two academic corpora, the British Academic Written English (BAWE) and the Brazilian Academic Written English (BrAWE). In order to conduct this analysis, we used corpus linguistics as our methodology. The results show that only few collocations of ACL came up as statistically significant when we investigated the BrAWE corpus in comparison with BAWE, indicating that Brazilians use academic collocations appropriately when compared to British. This research points to the importance of focusing on collocations in Academic English teaching contexts, since they cooperate to guarantee conventionality in language. Moreover, both British and Brazilians use academic collocations that are not necessarily present in ACL, suggesting a possible mismatch between what is prescribed and what is actually used in authentic language.

Keywords: Collocations. Academic English. Academic Collocation List. BAWE. BrAWE.

v. 9 (2)
195-213
maio-ago
2019

Resumo: A língua é formulaica por natureza, o que significa que a redação adequada vai além do conhecimento de palavras isoladas. O objetivo deste artigo é analisar a diferença quantitativa do uso de colocações da *Academic Collocation List* (ACL) em dois *corpora* acadêmicos, o *British Academic Written English* (BAWE) e o *Brazilian Academic Written English* (BrAWE). Para realizar essa análise, utilizamos a Linguística de *Corpus* como metodologia. Os resultados apontam que poucas colocações do ACL se mostraram estatisticamente significativas quando investigamos o *corpus* BrAWE em comparação com o BAWE, indicando que os brasileiros utilizam as colocações acadêmicas de forma adequada quando comparados aos britânicos. Esta pesquisa ressalta a importância do foco em colocações nos contextos de ensino de inglês acadêmico, uma vez que elas ajudam a garantir a convencionalidade no uso da língua. Além disso, observamos que britânicos e brasileiros usam colocações acadêmicas que não estão necessariamente presentes na ACL, sugerindo um possível descompasso entre o que é prescrito e o que é realmente usado na linguagem autêntica.

Palavras-chave: Colocações. Inglês acadêmico. *Academic Collocation List*. BAWE. BrAWE.

Introduction

Writing has been the focus of several studies in the field of English for Academic Purposes (EAP). Nevertheless, scholars have concentrated their investigations on the study of isolated words, either by creating word lists, such as the Academic Word List (AWL) (COXHEAD, 2000), the Academic Keywords List (AKL) (PAQUOT, 2010), and the Academic Vocabulary List (AVL) (GARDNER and DAVIES, 2014), or by analyzing those words in use (De COCK et al.1998; GRANGER 1998; LORENZ 1999; FOSTER 2001; NESSELHAUF 2005). In spite of agreeing that non-native speakers are acquainted with the formulaic characteristic of languages, those researchers state that non-natives at times underuse some linguistic constructions. Once scholars argue that language is formulaic in nature (DURRANT and SCHMITT, 2009), it is essential to give special attention to collocations, sequences of words that frequently co-occur (MCENERY and HARDIE, 2011).

Hence, lists that present isolated words do not help in the improvement of writing, especially if we take into account that fluency in a text is guaranteed mainly by the appropriate use of formulaic language (CHOI, 2016). Prodromou (2008) claims that mastering formulaic language is an important step towards the achievement of idiomatic production. Thus, rather than simply learning isolated academic words, it is worth knowing how to use them in context with a specific purpose.

Learning formulaic language and collocational sequences can be a great challenge for students. Following this idea, Bahns and Eldaw (1993, p. 108) explain that "collocations have been largely neglected in EFL instruction and that learners are therefore not aware of collocations

as a potential problem in language learning." Furthermore, when it comes to collocations, it is impossible not to mention the crucial role they play in a text. Biber and Conrad (1999) attest that ambiguity avoidance and clarity are guaranteed with the appropriate use of collocations. In other words, this means that fluidity in a text is given by collocational density, which in turn is a key characteristic of formulaic language. Finally yet importantly, it is impossible not to mention Firth's famous quote "you shall know a word by the company it keeps" (FIRTH, 1957, p. 179). This reinforces the argument that language is formulaic and sequences of words recur. The aim of this paper is to analyze the quantitative difference of Academic Collocation List's collocations in Brazilian Academic Written English (BrAWE) and British Academic Written English (BAWE) corpora.

In the following section, we present an overview of the concept of collocation the way it is used in this study and the ACL. The methodology is explained in section three, and the results found in this analysis are given in section four. In the last section, some final remarks are made.

Collocations

The following subsections present an overview of collocations and introduce the ACL.

Defining collocations: a quick glimpse at some notions

This section aims at demonstrating how different authors understand collocations and at presenting the definition of collocation as adopted in this analysis.

The definition of terms usually leads to incongruence depending on the authors. When it comes to defining the object of this study, it is not different, as "there seems to be no absolute definition of collocations" (ACKERMANN and CHEN, 2013, p. 244). Hill (2000) understands collocations as multi-word combinations. Shimohata et al. (1997, p. 476) conceptualize collocation as "a recurrent combination of words, ranging from word level to sentence level." Moreover, Shimohata et al. (1997) classify collocations in two types, one being "an uninterrupted collocation which consists of a sequence of words, the other is 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" (SHIMOHATA et al, 1997, p. 476).

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

The association of collocations with formulaic language is pointed out by Choi (2016), who states that formulaic sequences can be used as an overarching term for collocations. Wray (2000) explains that we retrieve collocations from our memories every time we use them. Therefore, the author's comprehension of collocations is that they are prefabricated elements of language. Sinclair (1991) affirms that when considered through corpus linguistics perspectives, collocations have to do with how likely words can co-occur.

Despite the plethora of definitions on collocations, they all have one thing in common, that is, they all refer to frequently combined words. In this paper, collocations are understood as a sequence of two words that co-occur more frequently together in a text than it would be expected by chance.

Academic Collocation List

Ackermann and Chen (2013) designed the ACL by using computational search of the most frequent combinations of words, statistical data of the Mutual Information (MI)[1] of the word combinations and revision of experts to determine whether the words in the preliminary list were appropriate. Based on the written curricular component of the Pearson International Corpus of Academic English (PICAE) the final version of the ACL is composed of 2,468 entries.[2]

The ACL is different from the Academic Formulas List (AFL) (SIMPSON-VLACH and ELLIS, 2010), since the latter is composed of 3, 4 and 5-gram sequences which are frequent in both written and spoken corpora, and ACL is composed of 2-gram sequences only. Together with the AWL, the two lists (ACL and AFL) play a complementary part in EAP teaching environments. The authors point out that "In addition to the Academic Word List and the Academic Formulas List, the ACL provides a further tool for EAP teachers to construct appropriate teaching materials and help students focus on frequent lexical items beyond individual words." (ACKERMANN and CHEN, 2013, p. 246).

A recent research based on the ACL is Frankenberg-Garcia t al.'s (2018), which aims at creating a writing tool to help learners of English enhance their academic writing performance when it comes to

[1] MI is a value that indicates how strong the link between two items is. The higher the MI score, the stronger the relation between the items.

[2] The complete list is available at https://www.eapfoundation.com/vocab/academic/acl/. [June, 2018]

the appropriate use of collocations. Along with ACL, the authors gather lemmas from the AVL and the AKL in order to build their collocation list.

In the next section, the methodology is presented.

Methodology

Following we explain how corpus linguistics is used in this study, the corpora analyzed – BAWE and BrAWE –, as well as the tools used to process them.

Corpus Linguistics

This study uses corpus linguistics (CL) as its methodology. "CL comprehends compilation and exploitation of corpora [...]. As such, it focuses on language exploitation through empirical evidence, extracted by a computer" (SARDINHA, 2000, p. 325. Our translation).[3] Corpora are compiled with the purpose of characterizing a specific portion of authentic language. For instance, if we intend to investigate how language is used in articles of a certain field, the corpus should contain articles of that field. If the purpose is to conduct a research on how discourse markers are used in job interviews, recordings of these interviews could be transcribed in order to compose a corpus that represents this genre.

An important distinction to be made is the difference between possibility and probability. While in Chomsky's view virtually any language construction is possible, for CL language is considered a probabilistic system. This concept follows an empirical approach and assumes a descriptive perspective when language is faced. Biber, Conrad and Reppen (1998, p. 1) claim that CL focuses on "how speakers and writers exploit the resources of their language", which is different than "what is theoretically possible in a language".

BAWE and BrAWE corpora

The British Academic Written English Corpus (BAWE) was compiled by Alsop and Nesi (2009) as part of the project called 'An Investigation of Genres of Assessed Writing in British Higher Education'.

[3] "A Lingüística de Corpus ocupa-se da coleta e exploração de corpora [...]. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador."

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

This academic corpus is composed of written texts produced by undergraduates and masters' students and contains texts of four big areas: Life Sciences (LS), Physical Sciences (PS), Social Sciences (SS), and Arts and Humanities (AH). These texts were categorized according to 13 academic genres: Case Study, Critique, Design Specification, Empathy Writing, Essay, Exercise, Explanation, Literature Survey, Methodology Recount, Narrative Recount, Problem Questions, Proposal, and Research Report. The students represented in BAWE received outstanding grades (merit and distinction) for their assignments.

The other academic corpus used for this investigation is the Brazilian Academic Written English (BrAWE) (SILVA, 2017), whose characteristics are quite similar to those of BAWE. BrAWE is considerably smaller – 768,323 tokens as opposed to the 3,312,196 tokens in BAWE – and it contains assignments of Brazilian students studying at British universities. These assignments received only passing grades. Considering that LS, SS and PS are the most representative areas in BrAWE, a subcorpus of BAWE was created in order to make both corpora comparable. However, we refer to this corpus simply as BAWE to avoid misunderstandings. Table 1 summarizes the characteristics of the study corpora:

Table 1 – BAWE and BrAWE corpora[4]

	BAWE	**BrAWE**
Number of assignments	2,761	380
Words[4]	2,768,588	657,859
Tokens	3,312,196	768,323

Source: the authors.

Methodological procedures

As it was stated previously, the objective of this paper is to analyze the quantitative difference of ACL's collocations in the BrAWE and BAWE corpora. Nevertheless, due to time and space constraints we would not be able to analyze all of the collocations identified. Besides, because the ACL is organized according to the nodes' (search words') alphabetical order rather than on the frequency of collocations, such as the AWL (COXHEAD, 2000) and the AVL (GARDNER and DAVIES, 2014), we chose the 10 most frequent collocation nodes in BAWE, used here

[4] There is a quantitative difference between words and tokens because punctuation marks are counted as tokens by Sketch Engine.

as the reference corpus, that is, the corpus the study corpus (BrAWE) is compared with. Therefore, the 10 most frequent collocation nodes presented in ACL were selected in BAWE in order to be analyzed in a contrastive way with BrAWE. Figure 1 illustrates the 'nodes' – referred as 'headwords' in ACL – and the collocations that contain those specific 'nodes':

Headword	Collocations
ability	cognitive ability
abstract	abstract concept
abuse	sexual abuse
academy	(in) academic circles, academic achievement, academic career, academic community, academic debate, academic discipline, academic discourse, academic institution, academic journal, academic life, academic performance, academic research, academic skills, academic study, academic success, academic work, academic world, academic writing, academic year
accept	accept responsibility, acceptable behaviour, socially acceptable, widespread acceptance, (be) commonly accepted, (be) generally accepted, (be) universally accepted, (be) widely accepted
access	allow access (to), deny access (to), direct access, easy access, electronic access, equal access, free access, gain access (to), give access (to), have access (to), internet access, limited access, online access, open access, provide access (to), public access, ready access, unlimited access, easily accessible, readily accessible
account	brief account, comprehensive account, historical account
accurate	great accuracy, accurate assessment, accurate description, accurate information, accurate measurement, accurate picture, accurate record
achieve	achieve (a) goal, achieve (an) objective, achieve (an) outcome, academic achievement
acquire	acquire knowledge, newly acquired

Figure 1 – Extract of ACL's nodes in alphabetical order
Source: the authors

In order to identify the most frequent nodes of ACL's collocations in BAWE, the nodes[5] were used as a whitelist[6] in Sketch Engine[7] and arranged according to their frequency. As a result, we identified a total of 57 nodes in BAWE (see appendix A). As mentioned above, for the purposes of this paper only the 10 most frequent collocation nodes were selected, as shown in Table 2.

[5] All the ACL's nodes were gathered in a txt file. Then, the list was uploaded as a whitelist in Sketch Engine and 57 came up as frequently used in BAWE, our reference corpus. For the purposes of this study, only the 10 most frequent nodes and their collocates were analyzed in both BrAWE and BAWE.

[6] The whitelist tool is used when the researcher is willing to analyze only the words in the list. For this study, the nodes of ACL were gathered so that they could be used as the whitelist in Sketch Engine.

[7] Sketch Engine is a tool used to explore how language works. Thus it is useful for lexicographers, translators, researchers in CL and language learners interested in studying the behavior of language through the analysis of texts stored in databases. https://www.sketchengine.eu/

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

Table 2 - Collocations associated to ACL's top 10 nodes

Position	Node	Frequency	Collocations in the ACL
1	used	9,188	commonly used, extensively used, frequently used, widely used
2	time	8,995	brief time, prime time
3	different	7,677	entirely different, fundamentally different, markedly different, qualitatively different, radically different, slightly different, substantially different, totally different, widely different
4	use	7,505	continued use, use criteria, use effectively, use resources, use sparingly, use statistics, use (a) format, use (a) method, use (a) methodology, use (a) procedure, use (a) source, use (a) strategy, use (a) technique, use (a) theory, use (an) approach, use (an) analysis, use (a) definition, use (the) concept, use (the) data, widespread use
5	people	6,569	indigenous people
6	system	6,397	binary system, capitalist system, complex system, comprehensive system, dynamic system, economic system, educational system, integrated system, legal system, solar system, transport system
7	order	6,299	established order, high order, natural order
8	new	6,015	entirely new, new initiative, new insight, new perspective
9	important	5,439	clearly important, equally important, increasingly important
10	example	5,299	classic example, obvious example, prime example, provide (an) example, specific example, striking example, typical example

Source: the authors.

Appendix B shows the 10 most frequent nodes, in bold, with their respective collocates according to ACL. After gathering a total of 64 collocations, we analyzed them separately in both corpora (see appendix C) through "Search - simple query" in Sketch Engine, tool which retrieves the search word(s) in their context of use (KWIC – keyword in context), as shown in Figure 2, which presents the five

concordance lines of the collocation *extensively used* in BrAWE:

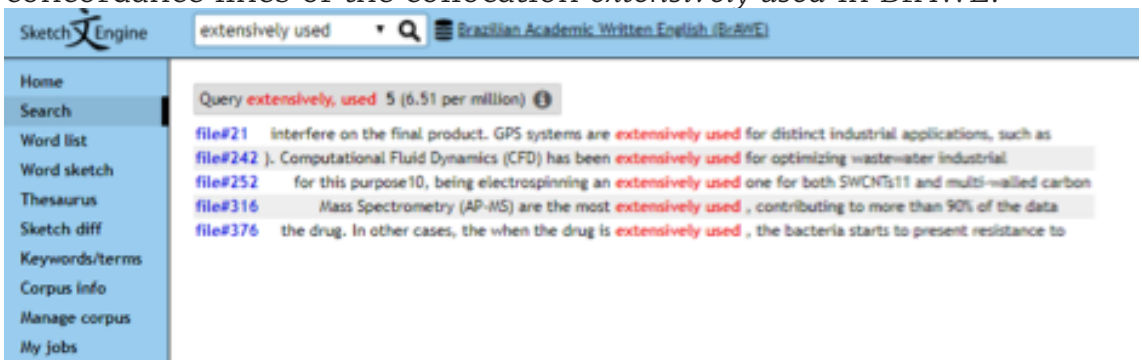

Figure 2 – Concordance lines of the collocation *extensively used* in Sketch Engine
Source: Sketch Engine.

In order to determine significant statistical difference between the occurrences, we used Log Likelihood (LL) calculator[8] (RAYSON, 2003). According to this test, if the outcome is 3,84 (negative or positive) or more, there is a 95% chance that the difference between the two corpora used in the investigation is not random. The positive (+) outcomes indicate an overuse of the given collocation in the first corpus, in this case BrAWE, in relation to the second one, in this case BAWE. A negative (–) outcome reveals an underuse of the collocation in the BrAWE corpus in comparison to BAWE. Figure 3 contains an example of how the calculator tool shows the results.

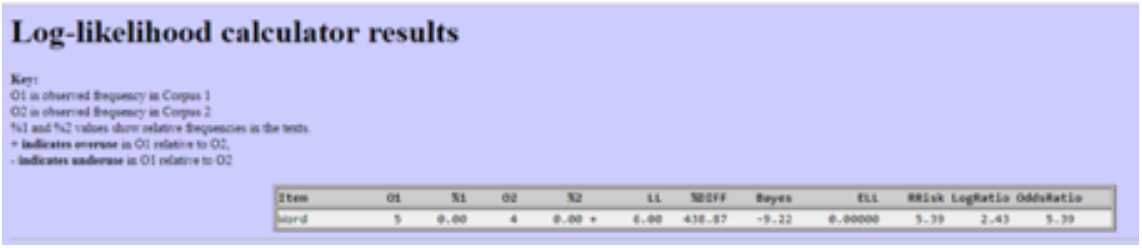

Figure 3 – Log-likelihood calculator – results of the collocation *extensively used*
Source: Log Likelihood. Available at: http://ucrel.lancs.ac.uk/llwizard.html.

By observing the table in Figure 3, it is possible to see that under "O1" we have the frequency in corpus 1 (BrAWE) – 5 –, and under "O2", the frequency in corpus 2 (BAWE) – 4. The number 6.00 under LL indicates the outcome of the statistical test. In this example, the result is positive (+) and higher than 3.84, meaning that the collocation is overused in BrAWE in comparison to BAWE. In this study, only LL

[8] Available at: http://ucrel.lancs.ac.uk/llwizard.html. Access: 9 Jan. 2019.

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

outcomes were taken into account. Other parameters are explained at the website.

Having outlined the methodological procedures, we will discuss some findings in the next section.

Findings

Considering the object of this investigation – collocations –, CL tools fit our needs and goals as they allow us to search for specific collocations in the corpora analyzed. For example, by typing the combination *entirely different* in both BAWE and BrAWE, the software comes up with the frequencies.

As can be observed in appendix C, from the 64 collocations of ACL only seven came up as statistically significant (the LL outcomes were higher than 3,84, positive or negative) in BrAWE and BAWE corpora. These occurrences, highlighted in gray in appendix C, represent only around 10.9% of the total data. In other words, 64 is the total amount of collocations analyzed, and the statistically significant outcomes (7) – gathered in Table 7 – account only for 10.9% of the collocations investigated.

Table 3 - statistically significant collocations

COLLOCATION	BrAWE	BAWE	LL
extensively used	5	4	6,00
widely used	18	38	5,64
binary system	2	0	6,68
legal system	0	31	-12,83
transport system	5	1	11,71
equally important	0	10	-4,17
increasingly important	0	15	-6,26

Source: the authors.

Interestingly, the statistically significant collocations are composed by only three nodes – *used*, *system* and *important*. Considering a comparison between the two academic corpora used here, *extensively used*, *widely used*, *binary system* and *transport system* are overused in BrAWE, whereas *legal system*, *equally important* and *increasingly important* are underused by Brazilians in their academic writing compared to the British students represented in BAWE.

Regarding the part of speech (POS) of the collocations, adverbs

(**adv**), verbs in the past participle (**vpp**), nouns (**n**) and adjectives (**adj**) are represented in the statistically significant collocations. The ones with the **vpp** as a node are preceded by an adverb (*extensively used*, *widely used*). Three collocations with a **n** as a node – *system* – are preceded by an adjective (*binary system, legal system* and *transport system*). The collocations with an **adj** as a node are preceded by an adverb, the only possible POS to come before an adjective (*equally important* and *increasingly important*).

Both examples of collocations in which an adjective is the node have no occurrence in BrAWE. Possible explanations are that (i) Brazilian students do not qualify their papers in terms of importance, as if they were not supposed to measure how important what they are writing is or (ii) they simply do not establish comparison of importance between what they are analyzing, describing or researching. Still regarding these two collocations (*equally important* and *increasingly important*), we observed that they are mostly used in essays. Of the 10 occurrences of *equally important* in BAWE, 6 of them are found in essays, which could indicate a recurrent type of collocation used in this text type.

Another finding to be highlighted is the fact that of all 64 collocations in ACL, 13 have no occurrences in BrAWE and BAWE, which is the case of *brief time*, *prime time*, *markedly different*, *use criteria*, *use statistics*, *use (a) format*, *use (a) procedure*, *use (a) strategy*, *comprehensive system*, *established order*, *new initiative*, *clearly important*, and *striking example*. Thus if these 13 collocations and the 7 statistically significant ones are disregarded, the remaining 44 collocations cannot be considered overused or underused in BrAWE, which might indicate that collocations are appropriately used by Brazilian students, when compared to British, at least if we consider only the ones analyzed in this paper.

As can be observed in the data, the statistical significant outcomes are not so numerous Hence, ACL might not be the most suitable list when it comes to collocations in academic English.

Final remarks

The aim of this study was to conduct a quantitative analysis on the use of collocations in BAWE and BrAWE corpora in order to verify if they are used proportionally by native and non-native speakers. Therefore, the ACL's 10 most frequent collocational nodes in BAWE were selected and later compared to BrAWE. A statistical test was run

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

in order to determine the significant outcomes in both corpora. Out of 64 collocations of ACL, only seven came up as statistically significant, four overused by Brazilian students and the remaining three underused in the Brazilian corpus.

This investigation brought together theoretical assumptions of collocations in a study that relied basically on corpus linguistics. In general terms, what could be concluded is that ACL's collocations seem not to be a writing issue in Brazilian assignments. However, more collocations could be investigated to confirm this idea.

As pointed out throughout this paper, language is formulaic in nature and collocations are one of the linguistic elements that are part of the umbrella term 'formulaic language'. Hence, collocations should be a pedagogical concern in any teaching context. As suggestions for follow-up studies, more ACL nodes could be analyzed to verify whether the tendency of having such low occurrence of statistical significant collocations remains. Moreover, conducting a study in which ACL and AFL are compared could be of great value, since both lists deal with formulaic language and ACL's outcomes were not statistically significant in this paper.

References

ACKERMANN, Kirsten; CHEN, Yu-Hua. Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. **Journal of English for Academic Purposes**, v. 12, n. 4, p. 235-247, 2013.

ALSOP, Sian; NESI, Hilary. Issues in the development of the British Academic Written English (BAWE) corpus. **Corpora**, v. 4, n. 1, p. 71-83, 2009.

BAHNS, Jens; ELDAW, Moira. Should we teach EFL students collocations?. **System**, v. 21, n. 1, p. 101-114, 1993.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. **Corpus linguistics: Investigating language structure and us***e*. Cambridge: Cambridge University Press, 1998.

BIBER, Douglas; CONRAD, Susan. Lexical bundles in conversation and academic prose. **Language and Computers**, v. 26, p. 181-190, 1999.

CHOI, S. Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, *21*(3), P. 403-426, 2016.

COXHEAD, Averil. A new academic word list. **TESOL quarterly**, v. 34, n. 2, p. 213-238, 2000.

DE COCK, Sylvie; GRANGER, Sylviane; LEECH, Geoffrey; MCENERY, Tony. An automated approach to the phrasicon on EFL learners. **Learner English**

on Computer, Sylviane Granger (ed.), P. 67–79. London: Addison Wesley Longman, 1998.

DURRANT, Philip; SCHMITT, Norbert. To what extent do native and non-native writers make use of collocations?. **IRAL–International Review of Applied Linguistics in Language Teaching**, v. 47, n. 2, p. 157–177, 2009.

FIRTH, John R. A synopsis of linguistic theory, 1930–1955. **Studies in linguistic analysis**, 1957.

FOSTER, Pauline. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In: **Researching pedagogic tasks**. Routledge,. p. 85–104, 2001.

FRANKENBERG-GARCIA, Ana et al. Developing a writing assistant to help EAP writers with collocations in real time. **ReCALL**, 2018. Draft submitted by the author. Available at: http://epubs.surrey.ac.uk/846264/.

GRANGER, Sylviane. Prefabricated patterns in advanced EFL writing: Collocations and formulae. **Phraseology: Theory, analysis, and applications**, v. 145, p. 160, 1998.

HILL, Jimmie. Revising priorities: From grammatical failure to collocational success. In: M. Lewis (Ed.), **Teaching collocation: Further developments in the lexical approach** (pp. 47–69). Hove: Language Teaching Publications, p. 47–69. 2000.

LORENZ, Gunter R. **Adjective intensification: learners versus native speakers: a corpus study of argumentative writing**. Rodopi, 1999.

MCENERY, Tony; HARDIE, Andrew. **Corpus linguistics: Method, theory and practice**. Cambridge University Press, 2011.

NESSELHAUF, Nadja. **Collocations in a learner corpus**. John Benjamins Publishing, 2005.

PAQUOT, Magali. **Academic vocabulary in learner writing: From extraction to analysis**. Bloomsbury Publishing, 2010.

PRODROMOU, Luke. **English as a lingua franca: A corpus-based analysis**. A&C Black, 2008.

RAYSON, Paul. **Matrix: A statistical method and software tool for linguistic analysis through corpus comparison**. Tese de Doutorado. Lancaster University, 2003.

SARDINHA, Tony Berber. Lingüística de corpus: histórico e problemática. **Delta**, v. 16, n. 2, p. 323–367, 2000.

SHIMOHATA, Sayori; SUGIO, Toshiyuki; NAGATA, Junji. Retrieving collocations by co-occurrences and word order constraints. In: **Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, 1997. p. 476–481.

SILVA, L. G. Compilation of a Brazilian Written English Corpus. **E-scrita**,v. 8, n. 2, p. 32–47, 2017

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

SIMPSON-VLACH, Rita; ELLIS, Nick C. An academic formulas list: New methods in phraseology research. **Applied linguistics**, v. 31, n. 4, p. 487-512, 2010.

SINCLAIR, John. **Corpus, concordance, collocation.** Oxford University Press, 1991.

WRAY, Alison. Formulaic sequences in second language teaching: Principle and practice. **Applied linguistics**, v. 21, n. 4, p. 463-489, 2000.

Appendices

Appendix A – 57 most frequent nodes in BAWE

	Node	Frequency		Node	Frequency		Node	Frequency
1	used	9,188	**20**	process	4,408	**39**	life	3,558
2	time	8,995	**21**	World	4,296	**40**	result	3,49
3	different	7,677	**22**	Value	4,294	**41**	analysis	3,429
4	use	7,505	**23**	State	4,168	**42**	point	3,417
5	people	6,569	**24**	Form	4,090	**43**	language	3,386
6	system	6,397	**25**	Results	4,080	**44**	study	3,343
7	order	6,299	**26**	Model	4,048	**45**	further	3,294
8	new	6,015	**27**	development	3,915	**46**	rate	3,288
9	important	5,439	**28**	information	3,889	**47**	evidence	3,163
10	example	5,299	**29**	Need	3,884	**48**	means	3,116
11	way	5,296	**30**	Level	3,863	**49**	based	3,103
12	work	5,025	**31**	Theory	3,841	**50**	role	3,091
13	found	4,772	**32**	control	3,755	**51**	values	3,083
14	number	4,711	**33**	Group	3,639	**52**	set	3,065
15	power	4,544	**34**	research	3,63	**53**	individual	3,049
16	case	4,526	**35**	increase	3,617	**54**	effect	3,000
17	data	4,523	**36**	Part	3,613	**55**	nature	2,945
18	high	4,473	**37**	change	3,593	**56**	higher	2,861
19	market	4,416	**38**	Society	3,577	**57**	method	2,850

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

Appendix B - Top 10 ACL's nodes (in bold) with the respective collocates

	Component I	POS I	Component II	POS II
1	commonly	Adv	**Used**	vpp
2	extensively	Adv	**Used**	vpp
3	frequently	Adv	**Used**	vpp
4	widely	Adv	**Used**	vpp
5	brief	Adj	**Time**	n
6	prime	Adj	**Time**	n
7	entirely	Adv	**Different**	adj
8	fundamentally	Adv	**Different**	adj
9	markedly	Adv	**Different**	adj
10	qualitatively	Adv	**Different**	adj
11	radically	Adv	**different**	adj
12	slightly	Adv	**different**	adj
13	substantially	Adv	**different**	adj
14	totally	Adv	**different**	adj
15	widely	Adv	**different**	adj
16	continued	Adj	**use**	n
17	use	V	criteria	n
18	use	V	effectively	adv
19	use	V	resources	n
20	use	V	sparingly	adv
21	use	V	statistics	n
22	use (a)	V	format	n
23	use (a)	V	method	n
24	use (a)	V	methodology	n
25	use (a)	V	procedure	n
26	use (a)	V	source	n
27	use (a)	V	strategy	n
28	use (a)	V	technique	n
29	use (a)	V	theory	n
30	use (an)	V	approach	n
31	use (an)	V	analysis	n
32	use (a)	V	definition	n
33	use (the)	V	concept	n
34	use (the)	V	data	n
35	widespread	Adj	**use**	n

36	indigenous	Adj	**people**	n
37	binary	Adj	**system**	n
38	capitalist	Adj	**system**	n
39	complex	Adj	**system**	n
40	comprehensive	Adj	**system**	n
41	dynamic	Adj	**system**	n
42	economic	Adj	**system**	n
43	educational	Adj	**system**	n
44	integrated	Adj	**system**	n
45	legal	Adj	**system**	n
46	solar	Adj	**system**	n
47	transport	Adj	**system**	n
48	established	Adj	**order**	n
49	high	Adj	**order**	n
50	natural	Adj	**order**	n
51	entirely	Adv	**new**	adj
52	new	Adj	initiative	n
53	new	Adj	insight	n
54	new	Adj	perspective	n
55	clearly	Adv	**important**	adj
56	equally	Adv	**important**	adj
57	increasingly	Adv	**important**	adj
58	classic	Adj	**example**	n
59	obvious	Adj	**example**	n
60	prime	Adj	**example**	n
61	provide (an)	V	**example**	n
62	specific	Adj	**example**	n
63	striking	Adj	**example**	n
64	typical	Adj	**example**	n

ISSN 2237-6321

v. 9 (2)
195-213
maio-ago
2019

Appendix C - 64 collocations analyzed

COLLOCATION	BrAWE	BAWE	LL
commonly used	16	58	0,36
extensively used	5	4	6,00
frequently used	1	11	-1,04
widely used	18	38	5,64
brief time	0	0	-
prime time	0	0	-
entirely different	0	6	-2,50
fundamentally different	0	4	-1,67
markedly different	0	0	-
qualitatively different	0	1	-0,42
radically different	0	4	-1,67
slightly different	9	34	0,12
substantially different	0	2	-0,83
totally different	2	7	0,06
widely different	0	2	-0,83
continued use	1	3	0,09
use criteria	0	0	-
use effectively	0	5	-2,09
use resources	0	1	-0,42
use sparingly	0	1	-0,42
use statistics	0	0	-
use (a) format	0	0	-
use (a) method	1	4	0,00
use (a) methodology	1	0	3,34
use (a) procedure	0	0	-
use (a) source	0	1	-0,42
use (a) strategy	0	0	-
use (a) technique	1	4	0,00
use (a) theory	0	1	-0,42
use (an) approach	0	1	-0,42
use (an) analysis	0	1	-0,42
use (a) definition	0	2	-0,83
use (the) concept	2	3	1,20
use (the) data	5	17	0,21
widespread use	1	7	-0,23

indigenous people	1	7	-0,23
binary system	2	0	6,68
capitalist system	0	3	-1,25
complex system	5	12	1,11
comprehensive system	0	0	-
dynamic system	2	3	1,20
economic system	1	11	-1,04
educational system	0	6	-2,50
integrated system	0	2	-0,83
legal system	0	31	-12,83
solar system	1	1	0,98
transport system	5	1	11,71
established order	0	0	-
high order	2	5	0,39
natural order	1	3	0,09
entirely new	0	5	-2,09
new initiative	0	0	-
new insight	3	3	2,95
new perspective	2	11	-0,11
clearly important	0	0	-
equally important	0	10	-4,17
increasingly important	0	15	-6,26
classic example	3	5	1,52
obvious example	0	2	-0,83
prime example	1	6	-0,10
provide (an) example	1	9	-0,59
specific example	0	3	-1,25
striking example	0	0	-
typical example	0	6	-2,50

ISSN 2237-6321