e-scrita ISSN 2177-6288

V. 8 – 2017.2 – SILVA, Larissa Goulart

COMPILATION OF A BRAZILIAN ACADEMIC WRITTEN ENGLISH CORPUS

Larissa Goulart da Silva[1]

ABSTRACT: The aim of this paper is to describe the compilation of a corpus of Brazilian academic English (BrAWE). This corpus allows other researchers to use it to investigate linguistic aspects related to English for Academic Purposes (EAP) produced by Brazilian students in comparison to other corpora such as BAWE or MICUSP, for instance. The students represented in it were Brazilian students doing part of their undergraduate degree in British universities. The corpus is divided into several subcorpora representing 13 academic genres and fours field of expertise.
Keywords: corpus linguistics, academic genres, English for academic purposes

A CRIAÇÃO DE UM CORPUS DE INGLÊS ACADÊMICO ESCRITO POR BRASILEIROS

RESUMO: O objetivo desse trabalho é descrever o processo de compilação de um corpus de inglês acadêmico escrito por estudantes brasileiros (BrAWE). Esse corpus permitirá que outros pesquisadores investiguem aspectos linguísticos relacionados a inglês para fins acadêmicos (IFA) produzido por estudantes brasileiros em comparação com outros corpora como o BAWE ou o MICUSP, por exemplo. Os estudantes representados são brasileiros realizando parte da sua graduação em universidades britânicas. O corpus está dividido em diversos subcorpora representando 13 gêneros acadêmicos e quatro áreas de estudo.
Palavras-chave: linguística de corpus, gêneros acadêmicos, inglês para fins específicos.

1. INTRODUCTION

Studies concerning how Brazilian students write academic English have become increasingly necessary, specially taking into consideration the ongoing process of internationalization of Brazilian higher education (CARNOY, 2013). According to Sarmento et al (2016, p.81) there are different internationalization strategies, such as, the establishment of international institutional networks, research projects developed in cooperation with other universities, academic mobility programmes, attraction of foreign students and researchers, among others. Nevertheless, English is usually the language adopted by these foreign universities which create cooperation networks in Brazil, even in the case of universities in countries, where English is not the mother tongue. Therefore, in order for Brazilian students

1 Mestre em Ensino de Língua Inglesa pela Universidade de Warwick, professora do programa Idiomas sem Fronteiras - UFRGS.

and researchers to be able to share their research in an international academic dialogue they have to master English for Academic Purposes (hereafter, EAP). Hence, the need for studies that investigate the academic language aspects specific of Brazilian students, so as to translate these studies into teaching materials that will meet the needs of students (HYLAND, 2016).

Considering the above mentioned, the objective of this paper is to describe the compilation of a Brazilian Academic Written English (BrAWE) corpus that can be used by researchers and teachers of Brazilian EAP to investigate the specificities of the academic language written by Brazilian students. This corpus comprises academic texts written by Brazilian students and submitted as part of their evaluation for their undergraduate courses in the UK. In addition, it is worth highlighting that these texts were written with the purpose of demonstrating the acquisition of disciplinary knowledge and skills (GARDNER & NESI, 2013) rather than attesting linguistic proficiency. BrAWE is a comparable corpus to the British Academic Written English Corpus (BAWE). The objective is that future investigations can be developed comparing EAP written by Brazilian students and students who received excellent grades in British Universities.

This paper is divided into five sections; the next section provides a brief overview of corpus linguistics and its applications to linguistic studies. It also presents previous studies on Brazilian English written corpora. Section three describes the corpus compilation. Section four addresses the corpus design and how it is divided in its subcorpora. Finally, section five presents some possible applications of the corpus for teachers and researchers of EAP.

2. CORPUS LINGUISTICS AND CORPORA OF BRAZILIAN STUDENTS.

As this paper describes the compilation of an Academic English corpus, this section will provide a brief introduction to corpus linguistics. According to McEnery and Hardie (2011, p.02) corpus linguistics "is not a monolithic, consensually agreed set of methods and procedures for the exploration of language. While some generalisations can be made (...) it is very important to realise that corpus linguistics is a heterogeneous field". Therefore, the next paragraphs describe the view of corpus linguistics adopted in this paper for the compilation of BrAWE.

Biber, Conrad and Reppen (1998, p.01) argue that corpus linguistics studies focus on language in use, in other words, "how speakers and writers exploit the resources of their language" rather than language structure, or "what is theoretically possible in a language". For McEnery and Hardie (2011, p.01) corpus linguistics refer to investigations that "deal with some

set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions". Conrad (2002, p.76) complements this definition stating that "a corpus is a large, principled collection of naturally occurring texts that is stored in electronic form (accessible on computer)". Therefore, given these definitions, it is possible to assume that a corpus contains a collection of natural occurring texts, compiled in order to represent a target domain of language use. In the case of BrAWE, the texts were originally students' assignments written for their UK universities and they were compiled together to represent academic English written by Brazilian students.

According to Biber, Conrad and Reppen (1998, p.04) in order to analyse the collection of texts that a corpus comprises, it is necessary to employ some kind of computer analysis. Nevertheless, these authors also claim that corpus linguistics analysis depends both on quantitative and qualitative techniques. Regarding this combination of both techniques, Conrad (2002, p.77) argues that "recognizing patterns of language use necessarily entails assessing whether a phenomenon is common or unusual - a quantitative assessment. At the same time, numbers alone give little insight about language". Hence, as a way to make this corpus available for both researchers and teachers to explore the linguistic patterns of Brazilian academic English the corpus described here is available on Sketch Engine[2], which is an online tool for corpus analysis that can be assessed through any computer.

In addition, regarding this division between empirical data and intuition, we can say that even though the research questions proposed in corpus linguistics studies are based on the researcher's intuition, the primary focus of corpus linguistics analysis is empirical data, or the actual use of language. Endorsing this claim, Flowerdew (2004, p.13) highlights the role of corpus linguistics in providing "attested examples of language patterns based on empirical data". Several researchers (CONRAD, 2002; CONNOR & UPTON, 2004; STUBBS, 2007) also point out that corpus linguistics studies can give evidence of recurring patterns in language, provide examples of lexico-grammatical aspects of language use, and determine what is typical and unusual in given circumstances. In sum, corpus linguistics studies rely on computerized tools to gain understanding of recurrent language patterns, which occur in texts produced for a communicative purpose outside the corpus.

Turning to studies on EAP written by Brazilian university students, it is worth highlighting two corpora recently compiled and their differences to the corpus presented here. The first one is *Br-ICLE*, according to Sardinha (2001) this corpus is composed of

[2] https://www.sketchengine.co.uk/

argumentative essays written by advanced learners taking the undergraduate course in English teaching or translating, these students represent different universities in Brazil. Even though this corpus comprises texts written by Brazilian students in higher education contexts it does not necessarily represents academic English as students were required to follow a predetermined theme for their written assignment. The other initiative is the *CorISF-English*, this corpus represents texts written by students in higher education contexts who are studying English at the Languages without Borders programme. This corpus represents written pieces of Brazilian students from different fields of expertise nevertheless these texts are not necessarily academic as genres such as letters were required of students. In addition, these two corpora, *CorISF* and *BR-ICLE* are different from the one presented here in several aspects, the texts in BrAWE were not written with the purpose of building a corpus, for example and the students represented here are from different areas of knowledge doing part of their undergraduate degree in an English speaking country, therefore they have already achieved a high level of proficiency attested by an IELTS test.

The aim of this section was to provide a brief overview of the understanding of corpus linguistics adopted in this paper and to establish the difference between BrAWE and other corpora of Brazilian students in higher education. The following section seeks to provide a recount of the corpus compilation process.

3. CORPUS COMPILATION

3.1 The case for a specialized corpus

The corpus described here is characterizes as a specialized corpus, in the sense that it comprises texts from a specific community, Brazilian students, and it represents a given type of texts, written assignments. Hunston (2002, p.14) argues that specialized corpora are:

> A corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language. Researchers often collect their own specialised corpora to reflect the kind of language they want to investigate. (HUNSTON, 2002, p.14)

Connor and Upton (2004, p.02) comparing the applications of general corpora and specialized corpora state that "while general corpora are important and provide a critical

foundation for the study of language structure and use, they are less conducive for analysing language use in specific academic and professional situations." Therefore, a specialized corpus gives a better understanding of the feature being studied than a generalized corpus. Furthermore, regarding corpus compilation these authors say that specialized corpora focus on texts of one specific genre or situation. However, in BrAWE more than one textual genre is represented, yet texts collected for the corpus represent the same language variety, Brazilian academic writing.

Hoey (2007, p.10) corroborates Connor and Upton (2004) argument that a specialized corpora can provide better understanding of how a linguistic feature is used in a specific context. This author, when discussing the analysis of priming, argues that "specialized corpora may be more revealing than general corpora, since general corpus may iron out the primings associated with particular genres or domains". Although this author is referring specifically to priming, his argument can be expanded to other aspects of language analysis that might be explored using BrAWE.

In addition, Connor and Upton (2004, p.02) present two advantages of using specialised corpora, the first one is that "it includes complete texts for a specific purpose instead of sample of texts" and the second one is that "specialized corpora are often small and collected by the analyst, these corpora often include more contextual information about the communicative situation than larger, general corpora". Hence, the aim of this paper is to provide the contextual information for future investigations exploring Brazilian EAP in BrAWE. Finally, specialized corpora can play an important role in understanding language of a more specific academic nature, as general corpora may not be suited for this role (FLOWERDEW, 2004, p.14).

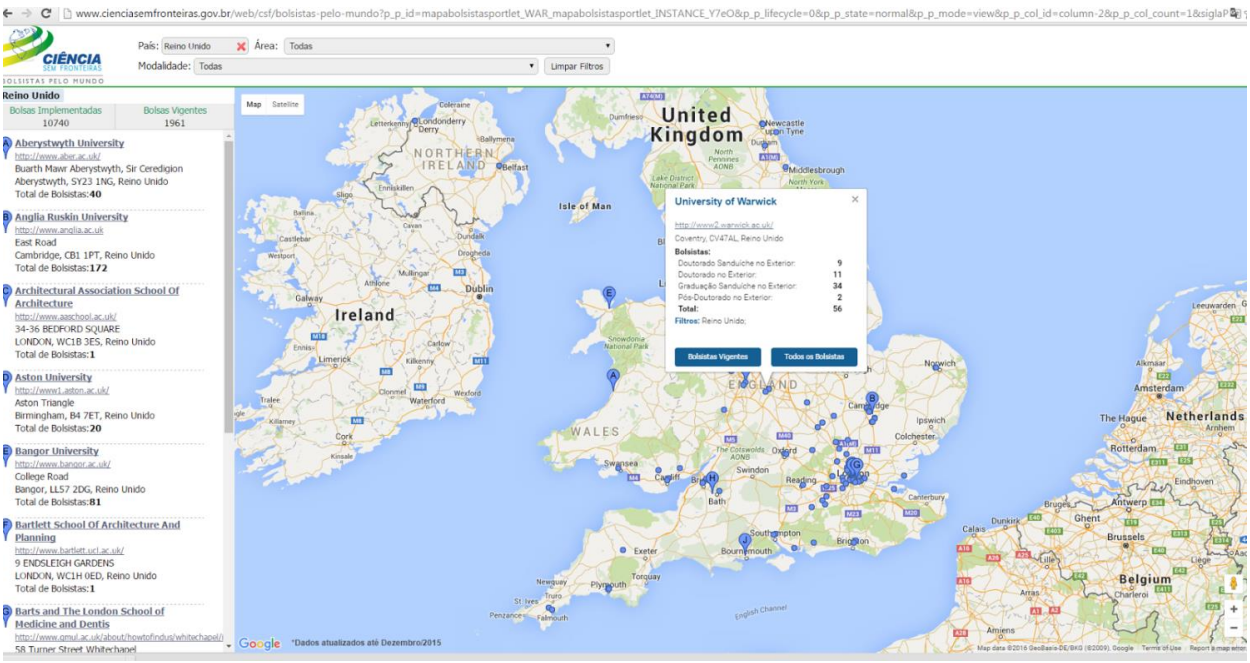
3.2 Data collection

In this section, I will discuss the steps of data collection and some problems encountered in this process. For the data collection, students were contacted through three different ways: the official webpage of the SwB Programme, Facebook, and advertisements in students' newsletters. The next paragraphs explain the procedure carried out to contact the students and gather their texts.

The first attempt to contact students was through the official SwB webpage, which contains a tab entitled Scholars throughout the World (Bolsistas pelo Mundo). In this tab, it is possible to see a map of the world with blue dots representing universities, which received

Brazilian students. The image below exemplifies how the UK is depicted on this map with the dots representing universities with Brazilian students.

Image 01 - SwB webpage



Once you click on a university, the page shows the number of Brazilian students enrolled, for example, in the image above the blue dot representing Warwick University was selected. After that, the user has two options: to view all scholars who have attended that university (Todos os Bolsistas) or just the current ones (Bolsistas Vigentes). At the early stages of data collection, "all scholars" was selected, however, soon the researcher has encountered some obstacles with this course of action on the webpage that made me opt to contact only current scholars.

The main difficulty using the SwB webpage was the restricted number of daily emails allowed. In order to protect students' privacy, this page does not provide students' emails, instead it mediates the correspondence between one person and the scholars. In other words, it would allow one person to send emails through the page, however one could not actually have access to students' emails. Therefore, for this research, there was a limit set by the page to contact only five students daily, and this limit is arbitrarily settled in order to avoid phishing and spam. In order to circumvent this problem, the author used four different email accounts (two institutional and two personal ones) to contact students, thus it was possible to send twenty emails a day instead of only five. Nevertheless, this was still a significant small amount considering that the goal was to contact all scholars. Hence, a decision was made to focus on

current scholars because this group was more likely to reply to this contact email. The researcher has also opted to send emails only to undergraduate students for three reasons: there are more undergraduate students than any other modality of the Programme, these students tend to write more assignments than PhD or visiting students and, finally, one of the objectives of the compilation was to make BrAWE similar to BAWE, which only contains texts of undergraduates' and masters' students.

Notwithstanding, it was part of the goals of the compilation to achieve as many students as possible, therefore students were contacted through Facebook in groups of Brazilians in the UK. This informal setting encouraged more students to reply. In addition, in the final stages of data collection I was also able to get in contact with former scholars with the help of SwB Network (Rede CSF) and the Educational secretariat of the Brazilian Embassy who have advertised this research on their weekly newsletters and put the researcher in contact with former scholars. This way texts of participants who had been SwB scholars from 2014 until 2016 were integrated in the corpus. Finally, the Brazilian Association of Students and Researchers in the UK had helped in the data collection through its mailing list. With the help of this organization more assignments from students participating in other mobility programmes and from other fields of study (arts and humanities and social sciences) were added to the corpus.

Regarding the email sent to students, it consisted of a text in English and Portuguese explaining the background and the objectives of the study, and it also described how students could participate in the study. In this passage, it was explained to students that their texts would be processed anonymously and that they could delete from their texts any sensitive issue or research results that they would not like to share.

4. CORPUS DESIGN

The aim of this section is to present in detail the steps of corpus compilation. I will explain the process of cleaning the data and the decisions made in the process of organizing the corpus. The final version of the corpus contains 380 texts of 186 students from 59 universities in the UK, comprising 768,323 tokens in Sketch.

4.1 Cleaning the data

Although all the texts that compose the corpus were originally in portable documents, participants sent these files in different formats (pdf, docs and odt). However, Sketch Engine, the software used to store this corpus only reads TXT files. Hence, it was necessary to clean the data and convert the texts to TXT.

For the purposes of this study each text was cleaned following the ensuing procedures, firstly PDF files were converted into Word format and then misspellings of words and mis-ordering of sentences that resulted from the conversion were corrected. This procedure was important because after the conversion, the lines from the same paragraph are split and sometimes words are divided into two parts and this would affect the final word count. After the correction of the problems resulting from the conversion, extra-linguistic features, such as, tables, images, charts and formulas were excluded from the TXT file. Nevertheless, whenever Unicode characters were used in a sentence, as it is shown in the example below which uses the character ±, they were left in the corpus for two reasons: the first one is that it would not be feasible to exclude all occurrences of Unicode characters in all the texts due to time constraints, and the second one is that these characters, when embedded in the sentence, perform a grammatical function, therefore they are relevant for the understanding of the sentences.

Image 02 - Example from text NWUT01I184

The temperature set-point of the reaction was 30.0 ± 0.05ºC.

In addition, textual features like titles, headings, acknowledgements, summary, block quotes, genuine lists, references, and appendix were excluded from the TXT file. However, as the goal is that the corpus might become useful for different research purposes in the future, all these textual features were kept in the PDF file version of the corpus. Moreover, BAWE has a PDF format that allows for investigations concerning genre structure, therefore to keep a PDF format in BrAWE might be helpful for investigations concerning how Brazilians structure academic genres in English.

Finally, after cleaning the texts and saving them in TXT and PDF, the text files were labelled. This naming of the files helped to organize them in different genres and later insert this information as metadata in Sketch Engine. I have also verified if the texts in the corpus were representative of most of the universities which received Brazilian students in the UK.

The end product of the corpus contains texts from 59 British universities, among the 87 that have hosted Brazilian students. Therefore, it represents a significant amount of universities. The next step was to divide the text into the same four disciplinary groups as BAWE - Life Sciences, Social Sciences, Arts and Humanities and Physical Sciences and categorize them according to the classification of genre families discussed in Gardner and Nesi (2013).

4.2 Corpus organization

According to Conrad (2002, p. 77) "corpus design is crucial to reliable and generalizable results (...) it is important to note that the size of the corpus, the types of texts included, the number of texts, the sampling procedure, and the size of each sample are all important considerations." A number of researchers (MCENERY, XIAO AND TONO, 2006; TIMMIS, 2015) also discuss the relevance of three concepts when compiling a corpus: representativeness, balance and sampling. McEnery, Xiao and Tono (2006, p.16) define balance as "the range of texts categories included in the corpus". These authors also discuss corpus sampling, they argue that "samples are scaled-down versions of a larger population" (p.19). According to them some aspects to be considered in sampling are the division of text-chunks and the proportion of samples in each text category (MCENERY, XIAO AND TONO, 2006, p.20). The last aspect, representativeness, is associated to the idea that the texts in a corpus should represent a wide range of text types (or genres) produced by a range of different users of the language variety being studied (MCENERY, XIAO AND TONO, 2006).

McEnery, Xiao and Tono (2006) also complement these definitions saying that the appropriate balance, sampling and representativeness of a corpus are associated with its intended uses, the research questions, and the group of language users it aims to represent. McEnery and Hardie (2011, p.10) claim that "balance, representativeness and comparability are ideals which corpus builders strive for but rarely, if ever, attain."

Nevertheless, a corpus linguist researcher should not dismiss these aspects when compiling a corpus as McEnery and Hardie (2011, p.10) declare "although balance and representativeness remain largely heuristic notions (...) this does not mean to say that the concepts are of no value". Considering balance in the BAWE corpus Alsop and Nesi (2009, p.73) state that

> simple random sampling would have been the most statistically valid way of achieving representation, had it been possible to identify the full range of assignments produced within each of the participating universities, and to acquire a proper sample from this resource pool. Unfortunately, we had no real

> means of assessing the volume or nature of assignments that would be at our disposal. (ALSOP; NESI, 2009, p. 73)

The same issue of sampling applies for the compilation of the corpus presented here since there was no information regarding the total number of Brazilian students in each university, their areas of study, and the genres they were asked to write. Furthermore, I also depended on the number of participants who would agree to participate in this corpus compilation. Thus, the final product of this corpus compilation is associated to what McEnery and Hardie (2011, p.11) define as an opportunistic corpus "they represent nothing more nor less than the data that it was possible to gather for a specific task". Regarding this type of corpus Sinclair (2005, p.81) also states that

> compilers should make the best corpus they can in the circumstances, and their proper stance is to be detailed and honest about the contents. From their description of the corpus, the research community can judge how far to trust their results, and future users of the same corpus can estimate its reliability for their purposes. (SINCLAIR, 2005, p.81)

Therefore in order to be detailed and honest about the content of the corpus the next sections aim at contextualizing the corpus and answering the questions relevant for this corpus presented in Flowerdew's (2004, p.25) set of general guidelines for building a specialized corpus.

4.3 Contextualization

The texts in the corpus were divided into four main areas - Social Sciences (SS), Arts and Humanities (AH), Physical Sciences (PS), and Life Sciences (LS). The advantages of using these groupings are that they are the same as other corpora such as BAWE, BASE, MICASE, and MICUSP (ALSOP AND NESI, 2009). Nevertheless, contrary to BAWE which contained more texts in AH and SS than in PS and LS (ALSOP AND NESI, 2009), the AH and the SS partitions of BrAWE are significantly smaller than the other two.

Additionally, the texts were divided into genre families in order to get more general information about the corpus. Therefore, I have read the texts received and compared them to the genre families described in Gardner and Nesi (2013). These authors describe 13 genre families - Case Study, Critique, Design Specification, Explanation, Exercise, Essay, Empathy Writing, Literature Survey, Methodology Recount, Narrative Recount, Problem Question, Proposal and Research Report. Even though this process was carried out by only one person, the researcher, and with limited time, each text was read twice and compared to the textual

features with the genre descriptions provided by these authors. Furthermore, whenever there was a doubt regarding the text typology I opted to include the text in the prevailing objective of the assignment. In other words, if the main objective of a text was to perform a critique, but one section was dedicated to a case study, the whole text was still classified as a critique.

Finally, it is worth noting the special case of the "final report", SwB students are asked to write a final report of their activities during their one-year exchange, these reports varied from 5 to 20 pages. I have chosen to include this piece of writing in the corpus because they are relevant for future SwB scholars who will study in the UK. Even though these final reports are not taken into account to determine student's grades, they are assessed by the SwB programme. Hence, these texts were included in the corpus. Regarding their genre classification, as they have similar characteristics to the genre family Research Report they were included in this category.

Furthermore, texts in the BAWE corpus contain a series of other contextual information, such as, students' level of education, their grades, previous study background, gender, among other information. Nevertheless, students who contributed to BAWE had a financial motivation to proceed with their submission, while the students who contributed to the corpus presented here were solely motivated by their sympathy to the project. Thus, I did not want to ask many questions and risk them losing interest in participating on the research as it is reported in Alsop and Nesi (2009). Hence, the only prerequisite required to include the texts in the corpus was that the assignment had received at least a pass, yet there was no formal requirement for the students to show proof of their grades for each assignment.

Nevertheless, as a Brazilian in the UK who have met some of the students who participated in this study I can provide a brief context regarding students level of education. One of the criteria for students to apply for the SwB is to have finished at least 20% of their course credits in Brazil, during the process of application and acceptance they continue their regular course in Brazil, therefore when most students arrive in the UK, they are already in their final years of undergraduate course in Brazil, which are correspondent to the third year or masters' in the UK. Thus, the texts in the corpus presented here would be comparable to levels 3 and 4 in the BAWE corpus. The table below presents the number of tokens and assignments per genre and field of study.

Table 01 - Numbers of texts and words by disciplinary group

	AH		SS		LS		PS		Total	
	Texts	Tokens	Texts	Token	Texts	Tokens	Texts	Tokens	Texts	Tokens
Case Study			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							18	36,093	18	36,093
Empathy Writing										
Essay	4	7,887	13	20,906	46	82,975	31	48,401	94	160,169
Exercise			1	1,594	7	6,829	28	35,236	36	43,659
Explanation			7	11,371	11	14,976	29	54,266	47	80,613
Literature Survey					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	3,369	3	4,306	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**

As we can see PS is the area with more texts in the corpus, this is not surprising if we take into consideration the fact that SwB scholarships are awarded mainly for this area of study. In addition, Essays is the genre with more texts, followed by Methodology Recounts and Explanations and there is no texts of Empathy Writing, which was expected since in the BAWE corpus this is the genre with the lowest frequency.

The aim of this section was to contextualize the corpus, providing some background information about the students who collaborated with it and an overview into its organization. The next sections aim at answering Flowerdew's (2004) questions relevant for the corpus compiled for this study.

4.4 What is the purpose for building this specialized corpus?

According to Flowerdew (2004) and McEnery, Xiao and Tono (2006) the purpose of a corpus determines the other characteristics (e.g. size, balance, sampling and representativeness) of it. In this respect, BrAWE has one main goal to serve as a resource for studies regarding Brazilian Academic writing and comparative studies between BrAWE and BAWE.

4.5 How large should be the specialized corpus?

Flowerdew (2004) states that the size of a specialized corpus is associated with the phenomenon being studied. According to this author, if the feature being investigated is

frequent in common language, than the corpus can be smaller, while if the analyst is investigating a low frequency feature, a larger corpus is necessary (FLOWERDEW, 2004, p.26). Gavioli (2002) discusses the necessity of having a large enough corpus in order for the results to be generalizable. Given these claims, it is possible to consider that a large corpus would be appropriate for the study of Brazilian EAP, nevertheless, pragmatic aspects have to be taken into consideration when estimating the adequate corpus size. As previously stated, the corpus presented here has as many texts as it was possible to obtain considering the time constraints and the representativeness of the language variety being investigated. Considering the corpus size in tokens, Sardinha (2000) claims that a corpus of around 500,000 tokens can be classified as a medium-size corpus. Thus, the corpus presented here would be classified as a medium-size corpus.

Turning to, considering the corpus and the linguistic feature being studied, McEnery, Xiao and Tono (2006) argue in favour of a closure/saturation measure in order to verify if a specialized corpus is of adequate size. According to this method, the corpus should be divided into different portions of the same size and with the addition of each portion, the number of lexical items should increase in the same proportion. However, Geng (2015) approaches the closure/saturation measure through a different perspective: this researcher has gradually decreased the size of her corpus and verified if the occurrences of the feature she was studying remained the same with each reduction. I have decided to do the same as Geng (2015) using academic vocabulary as a mean to observe the closure/saturation measure of BrAWE. Therefore, I have tested the AWL coverage in the corpus reducing randomly 5% of the corpus 18 times until only 10% of the corpus was left, the AWL coverage remained roughly the same (around 9%) in all of these reductions. Hence, the corpus seems to be of appropriate size to explore Brazilian EAP.

4.6 What genre is to be investigated?

The aim of this corpus is to act as a tool to investigate academic English written by Brazilian students as part of their assignments for university, thus different academic genres are part of this investigation. However, the assignments collected for this study were divided according to Gardner and Nesi (2013) genre family classification.

Gardner and Nesi (2013) argue that student writing differs from research writing in the sense that "assignments generally aim to demonstrate the acquisition of required skills and accepted knowledge" rather than convince the reader of research results and applications as it

is the case in professional research writing. Therefore, these authors have analysed other studies suggesting genre classifications of students writing and based on this literature review and on the similarities and differences encountered among the 2858 texts collected for BAWE, they proposed 13 genre families that represent the sum of assignments written by students.

In addition, regarding genre structure, a decision was made to include the full texts in the corpus rather than text segments. Although Flowerdew (2004, p.15) states that corpus based on "text-segments are useable for investigations of individual lexical or grammatical items", which is the case of the study presented here, this type of corpus "is unsuitable for top down-down genre based analysis, as for instance, the analysis of lexico-grammatical items in different sections of the text" which might be done in the future.

4.7 Is the specialized corpus representative of the genre?

Flowerdew (2004, p.26) claims that "specialized corpora are considered representative of the genre under investigation if they contain numerous texts from a variety of authors so that no one authorial style would dominate and typical lexical or grammatical patterns would be revealed." Therefore, even though most of the texts received were included in the corpus there was a set of criteria that determined the exclusion of some texts. The first one was related to the objectives of the corpus, professional texts, such as articles submitted to journals, or final dissertations, were excluded as they did not represent student writing for coursework. In addition, only three texts from the same participant in the same genre were included, this was an attempt to avoid overrepresentation of only one student writing style. I have chosen the number three arbitrarily in order not to discard many texts, but also not to allow one "authorial style to dominate the corpus" (FLOWERDEW, 2004, p.26). However, this criteria was not applied to the genre of Exercises as the number of words in each assignment in this case is usually significantly lower when compared to other genres.

In this section, the corpus compilation, the data cleaning and how the corpus is organized was explained. The aim of the next section is to present how BrAWE can be used in future investigations regarding linguistic aspects of Brazilian EAP.

5. CORPUS APPLICATIONS

The aim of this paper was to introduce BrAWE corpus and its applications in the field of research and teaching of EAP in Brazil. So far, this corpus has been used to explore the use

of vocabulary by Brazilian students in contrast to students represented in BAWE. From this investigation, some aspects regarding Brazilian EAP could be observed, the first one was the overuse of the passive voice by Brazilian students when compared to BAWE. In addition, these students tend to use fewer multiword-units than BAWE students and there is also evidence that Brazilian students use significantly less process of morphological derivations in their texts. Nevertheless, there is a plethora of studies that are still possible to be developed especially considering the use of academic features, such as, technical vocabulary, concordances, lexical bundles, across different disciplines or across different genres.

Moreover, BrAWE gives EAP teachers a significant amount of authentic materials that can be used in class. EAP teachers have the opportunity to explore concordance and keywords of their students' fields of study, or even to analyse the occurrence of common mistakes in a corpus composed of assignments written in the same discipline as their students.

REFERENCES

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

BIBER, D; CONRAD, S; REPPEN, R. *Corpus linguistics: investigating language structure and use.* Cambridge: Cambridge University Press, 1998.

CARNOY, M. *University expansion in a changing global economy: the triumph of the BRICS?* Stanford, California: Stanford University Press, 1998.

CONNOR, U; UPTON, T. Introduction. In: _____ (Eds.), *Discourse in the professions*: *Perspectives from corpus linguistics*. Amsterdam: John Benjamins Publishing Company, 1998. P. 1 – 8.

CONRAD, S. Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics,* v. 22, p. 75-95, 2002.

FLOWERDEW, L. The argument for using English specialized corpora to understand academic and professional language. In CONNOR, U; UPTON, T. (Eds.), *Discourse in the professions*: *Perspectives from corpus linguistics*. Amsterdam: John Benjamins Publishing Company, 2004, p. 11 – 36.

GARDNER, S; NESI, H. A classification of genre families in university student writing. *Applied Linguistics,* v. 34, n. 1, p. 25-52, 2013.

GAVIOLI, L. Some thoughts on the problem of representing ESP through small corpora. In: KETTEMAN, B; MARKO, G. (Eds), *Language and Computers: Studies in Practical Linguistics*. Amsterdam: Rodopi, 2002, p. 293–303.

GENG, Y. *Appraisal in discussion sections of doctoral theses in the discipline of ELT/Applied Linguistics at Warwick University: A corpus-based analysis*. 2015. 336f. (PhD Theses) - The University of Warwick.

HOEY, M. Lexical priming and literary creativity. In: HOEY, M; MAHLBERG, M; STUBBS, M; TEUBERT, W (Eds.), *Text, discourse and corpora: Theory and analysis.* London: Continuum. 2007, pp. 7-30.

HUNSTON, S. *Corpora in applied linguistics*. Cambridge: Cambridge University Press. 2002.

HYLAND, K. General and specific EAP. In: HYLAND, K; SHAW, P. *The Routledge Handbook of English for Academic Purposes*, London, Routledge, 2016.

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

MCENERY, T; XIAO, R; TONO, Y. *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge. 2006.

SARDINHA, A. Linguística de corpus: histórico e problemática. *D.E.L.T.A*., v.16, n.2, p.323 – 367, 2002.

SARMENTO, S; DUTRA, D; BARBOSA, M; MORAES FILHO, W. IsF e Internacionalização: da teoria à prática. In: SARMENTO, S.; ABREU-E-LIMA, D.; MORAES FILHO, W. (Org). *Do Inglês sem Fronteiras ao Idiomas sem Fronteiras: A construção de uma política linguística para a internacionalização.* Belo Horizonte: UFMG, p. 77 – 104. 2016.

SINCLAIR, J. Corpus and text - Basic principles. In: WYNNE, M (Ed.), *Developing linguistic corpora: A guide to good practice.* Oxford: Oxbow Books. 2005, p. 1-16.

STUBBS, M. On texts, corpora and models of language. In HOEY, M; MAHLBERG, M; STUBBS, M; TEUBERT, W. (Eds.), *Text, discourse, and corpora*: *Theory and analysis.* London: Continuum. 2007, p. 127-162.

TIMMIS, I. *Corpus Linguistics for ELT: Research and Practice*. Abindgon: Routledge Taylor and Francis Group. 2015.

Enviado em: 27-11-16
Aceito em: 13-12-16