Academic vocabulary: a corpus linguistic study on how Brazilian students write academic English

Author: Larissa Goulart Da Silva

University of Warwick

British Council ELT Master's Dissertation Awards: **Special Commendation**

Academic vocabulary: a corpus linguistics study on how Brazilian students write academic English.

Larissa Goulart DA SILVA
1565725

Dissertation submitted in partial fulfilment of the requirements for the degree of MA in English Language Teaching (with a Specialism in English for Specific Purposes)

Centre for Applied Linguistics
University of Warwick

September, 2016

Acknowledgements

A big thank is due to all the Brazilians studying in the UK who collaborated with my research by sending their texts. Without your help, this study would not have been possible.

I would also like to thank the Hornby Trust for the funding and the British Council for the support provided during this year in the UK.

I also want to express my gratitude to my lecturers at Warwick University, especially to my personal tutor Steve Mann and my supervisor Sue Wharton for their help with my dissertation.

In addition, there are not enough words to express how thankful I am to my research group in Brazil – "the simonetes" – thank you for discussing my dissertation ideas with me, for the skype meetings, and for helping me with the final revision of this dissertation.

Finally, thanks for the person who makes this group possible, Simone. Thank you for being my supervisor over the past six years, for encouraging me to apply for the Hornby scholarship, and for everything else.

Abstract

According to Vongpumivitch, Huang, and Chang (2008) and Shawn (1991) the use of academic vocabulary is one of the main difficulties encountered by English as a foreign language students when reading and writing academic English. Therefore, the aim of this study is to explore the use of academic vocabulary in a corpus of Brazilian students in order to propose pedagogical implications for the teaching of EAP in Brazil. The research questions are: a) what is the vocabulary profile of assignments written by Brazilian students?; b) how does it compare to the vocabulary profile of other academic corpora presented in the literature review?; c) what words in the AWL do Brazilian students use?; and d) how does the use of academic words differ between Brazilian students and students represented in the BAWE corpus?. In order to answer these questions a corpus of assignments written by Brazilian students was compiled and analysed using Range and Sketch Engine. The results show that the corpus of Brazilian students present the same coverage of AWL words as other academic corpora. However, the word forms selected are different, some of the reasons for this are the underuse of lexical bundles and derivations processes. These findings provide insights on topics that could be addressed in classes of EAP for Brazilian students.

Table of Contents

1. **Introduction** 10

1.1 Background of the study 10

1.2 Aims of the research 13

1.3 Organization of the study 14

2. **Literature Review** 15

2.1 Introduction 15

2.2 What is Academic Vocabulary? 15

2.3 The role of vocabulary in language learning 17

2.4 Why teach Academic Vocabulary? 20

2.5 Academic word lists: a timeline 22

2.6 AWL coverage in other corpora 27

3. **Methodology** 33

3.1 Introduction 33

3.2 Corpus Linguistics 33

3.3 The case for a specialized corpus 34

3.4 Data collection 36

3.5 Corpus compilation 38

3.5.1 Cleaning the data 38

3.5.2 Corpus organization 40

3.5.2.1 Contextualization 41

3.5.2.2 What is the purpose for building this specialized corpus? 44

3.5.2.3 How large should be the specialized corpus? 44

3.6 Data analysis 45

4. **Results and Findings** 47

4.1 Introduction 47

4.2 The AWL coverage 47
4.2.1 Life Sciences' lexical profile 49
4.2.2 Physical Sciences' lexical profile 51
4.2.3 Discussion 52
4.3 The use of academic vocabulary: BAWE compared to Brazilian Written English 52
4.3.1 The top 40 AWL in BAWE and Brazilian students 53
4.3.1.1 High frequency words in BAWE 56
4.3.1.2 High frequency word in the corpus of Brazilian students 60
4.3.2 Differences on the first sublist 64
4.3.2.1 Suffixes 65
4.3.2.2 Prefixes 66
5. **Conclusion** 70
5.1 Introduction 70
5.2 Summary of key findings 70
5.3 Research evaluation and limitations 71
5.4 Pedagogical implications 71
5.5 Further research suggestions 72
References 74
Appendices 81

List of Tables

Table 1 – Areas covered by the Science without Borders programme 11
Table 2 - Academic words x High frequency words x Portuguese 17
Table 3 – A timeline of academic word lists 23
Table 4 - Academic Word List Coverage 28
Table 5 - Numbers of texts and words by disciplinary group 43
Table 6 - AWL coverage in the corpus of Brazilian students 48
Table 7 - Most frequent academic words 53
Table 8 - Evidence: BAWE compared to Brazilian students 57
Table 9 - Response: BAWE compared to Brazilian students 58
Table 10 - Team: BAWE compared to Brazilian students 59
Table 11 - Modifiers of “energy: Brazilian students 61
Table 12 - Modifiers of “energy”: BAWE 61
Table 13 - Energy: BAWE compared to Brazilian students 62
Table 14 - Construction: BAWE compared to Brazilian students 63
Table 15 - Concentration: BAWE compared to Brazilian students 64
Table 16 - Word forms of the word “concept” 65
Table 17 - Suffixes in BAWE versus Brazilian students 66
Table 18 - Prefixes in BAWE x Brazilian students 67

List of Figures

Figure 1 - SwB webpage 36
Figure 2 - Example from text NWUT01I184 39
Figure 3 - Excerpt from LS text. 50
Figure 4 - Excerpt from PS corpus 51
Figure 5 - Evidence: BAWE 57
Figure 6 - Evidence: Brazilian students 57
Figure 7 - Response: BAWE 58
Figure 8 - Response: Brazilian students 58
Figure 9 - Team: BAWE 59
Figure 10 - Team: Brazilian students 59
Figure 11 - Energy: BAWE 60
Figure 12 - Energy: Brazilian students 60
Figure 13 - Construction: BAWE 62
Figure 14 - Construction: Brazilian students 62
Figure 15 - Concentration: BAWE 63
Figure 16 - Concentration: Brazilian students 63

List of Appendices

Appendix 1 - Email sent to Brazilian Students 81
Appendix 2 - Top 50 off-list words in Life Sciences 82
Appendix 3 - Top 50 off-list words in Physical Sciences 84
Appendix 4 - AWL words in BAWE and the corpus of Brazilian students 86
Appendix 5 - First sublist of the AWL 124
Appendix 6 – Ethical Form 139

List of Abbreviations

AH – Arts and Humanities

AKL – Academic Keyword List

AVL – Academic Vocabulary List

AWL – Academic Word List

BASE – British Academic Spoken English corpus

BAWE – British Academic Written English corpus

CAPES - Coordination for the Improvement of Higher Education Personnel

CNPq - National Council for Scientific and Technological Development

EAP – English for Academic Purposes

EFL – English as a Foreign Language

GSL – General Service List

LL – Log-likelihood

LS – Life Sciences

LwB – Languages without Borders

MICASE – Michigan Corpus of Academic Spoken English

MICUSP – Michigan Corpus of Upper-Level Student Papers

NAWL – New Academic Word List

NGSL – New General Service List

SwB – Science without Borders

PS – Physical Sciences

SS – Social Sciences

UFRGS – Federal University of Rio Grande do Sul

UK – United Kingdom

UWL – University Word List

1. Introduction

Having an extended vocabulary is essential for reading and writing in a foreign language. Some studies (Laufer, 1989; Hirsh & Nation, 1992; Hsueh-Cho & Nation, 2000; Nation, 2013) argue that it is necessary to know around 95% to 98% of the words in a given text to gain appropriate understanding of it. In academic settings, formal writing requires students to have a vast knowledge of academic words and to actively use it on written assignments in order to achieve academic success (Kaur and Hegelheimer, 2005; Coxhead, 2011). This study uses corpus linguistics tools to investigate how Brazilian students produce academic vocabulary when writing assignments during their exchange programmes at British universities. Academic vocabulary is different from frequent words - words used in everyday life - or discipline specific words - words related to a specific field of study. Academic words are lexical items that are recurrent in academic texts (Chung and Nation, 2003; Clark and Ishida, 2005) and, therefore, are not usually acquired in everyday interaction.

This chapter presents a background for this study; it explains my motivation to pursue this topic as well as the UK-Brazil academic exchanges in higher education, which provides the setting for the students who participated in this research (1.1). In the following section (1.2), I present the aims of this study and its research questions. Finally, in the last section (1.3) I describe the outline of this study.

1.1 Background of the study

My motivation to investigate written academic English derives from my experience as an English teacher in the Languages without Borders (LwB) programme, at the Federal University of Rio Grande do Sul (UFRGS), in which I taught academic English for Brazilian students. In my experience, I have noticed that Brazilian students, especially at lower levels of proficiency, tend to use high frequency words rather than academic ones assuming that the former will grant them better marks as they are less similar to Portuguese words. Regarding the LwB, this is a Federal government policy present in 63 different Brazilian public universities (Brazil, 2015). This programme offers free English classes to all university community (staff, students, and professors). LwB objectives are slightly different from language centres in two aspects. Firstly, LwB employs novice teachers as a way to encourage teacher development during initial teacher training. In 2015, for example, at LwB – UFRGS, nineteen out of twenty teachers were undergraduate students in

their final year (Sarmento and Kirsch, 2015). Therefore, another motivation for this study is to provide insights into academic vocabulary that can help novice teachers in the LwB programme to prepare their classes. The second aspect is that LwB was created to promote the internationalization of Brazilian universities (Brazil, 2014) by focusing on English for Academic Purposes (EAP) classes.

Languages without Borders is a policy that first emerged as a branch of the Science without Borders (SwB) programme, which is a scientific mobility programme created in 2011. Initially most students participating in SwB selected Portugal as their country of destination due to the fact that both countries share Portuguese as their first language. The, then, ministry of education Aloisio Mercadante, noticing that the linguistic gap hindered students to select universities in other foreign countries, decided to create the LwB to improve the linguistic proficiency of the academic community in Brazilian universities. As most of the students who collaborated with the current research were SwB former or current students, the following paragraphs provide a short explanation of this educational policy.

Science without Borders is a large-scale scholarship programme that aims at strengthening, expanding and consolidating the internationalization of Brazilian higher education by promoting students and researchers exchange between international universities of excellence and Brazilian universities. This programme provides scholarships for students and researchers, in the areas presented in the table below, to take part in an academic mobility.

Table 1 – Areas covered by the Science without Borders programme

Engineering and other technological areas	Minerals
Pure and Natural Sciences	Biotechnology
Health and Biomedical Sciences	Nanotechnology and New Materials
Information and Communication Technologies	Technology for prevention and mitigation of natural disasters
Aerospace	Biodiversity and bioprospecting
Pharmaceuticals	Marine Sciences
Sustainable Agricultural Production	Creative Industry
Oil, Gas and Coal	New technologies for constructive engineering
Renewable Energy	Formation of technical personnel

Source: SwB website

The SwB programme promotes different modalities of exchange according to student's level of education: undergraduate students participate in a one-year exchange, doctoral students can take part on a full PhD course for three years or participate as a visitor graduate student for a year or a semester, and postdoctoral students, usually, can develop a project during the course of one academic year. Since the establishment of the programme in 2011, the United Kingdom (UK) has received 10,740 Brazilian students in 87 British universities. These students came from 131 different Brazilian universities, representing different regions of Brazil. According to the data presented on the SwB official website, from these 10,740 students, 774 were visiting PhD students, 571 were full PhD students, 8,864 were visiting undergraduate students, and 531 were postdoctoral students.

For those students whose fields of study are not encompassed by SwB covered areas, mainly arts, humanities and social sciences, other initiatives from the Brazilian government provide different opportunities to study abroad. CAPES (Coordination for the Improvement of Higher Education Personnel) and CNPq (National Council for Scientific and Technological Development) offer full PhD grants, while other foreign initiatives, such as the recently created Newton Fund, Santander Universities, and the Chevening Programme, allow Brazilian students to study in the UK. However, these are provided in smaller amounts and this is why these disciplines are underrepresented in the corpus compiled for this study.

Considering the substantial amount of scholarships provided in the last seven years, the number of Brazilian students and researchers participating in academic activities in the UK has largely increased and, as a consequence, so has the number of research partnerships and articles co-authored between researchers in both countries. Stern (2016) states that the number of co-authored papers between UK and Brazilian researchers has increased by 196% in the last seven years. This increase on publications highlights the need for studies on EAP written by Brazilian researchers. Moreover, this ongoing academic exchange between Brazil and English speaking countries, in general, has encouraged Brazilian universities to offer disciplines in English, which emphasizes the need for research in an even more specific EAP field: students' academic writing to fulfil coursework requirements.

Regarding previous investigations on academic English written by university students, according to Gardner and Nesi (2013) researchers of EAP in the UK have neglected students' coursework writing due to the difficult process of collecting a representative body of written work. These authors also state that "this type of writing is more varied than research writing, and manifests itself through a wider range of genres" (Gardner and Nesi, 2013:26). Therefore, even

though several researchers (Freitas, 2016; Sarmento, Scortegagna and Silva, 2014; Dayrell and Aluisio, 2008, etc) have investigated Brazilian EAP, they have either focused on research writing or students' interlanguage, hence the area of Brazilian students written EAP to fulfil coursework demands is still unexplored.

Finally, given the context provided, this research seeks to investigate one aspect of Brazilian students' academic writing: vocabulary. I hope that the outcome of this research can be useful for Brazilian teachers and students of EAP. Furthermore, this dissertation aims at providing some reflections into the field of EAP and university students writing. The objective of this section was to present my motivation for this study and the context in which the students who participated in this study are inserted. The next section (1.2) provides the aims and questions that guide this investigation.

1.2 Aims of the research

First, the broad objective of this study is to collaborate to the ongoing process of internationalization in Brazilian universities by investigating Brazilian EAP in assignments written as part of students' coursework in the UK. In addition, I hope that the corpus compiled for this study can be useful for other investigations on Brazilian EAP.

Gass and Selinker (2008:137) argue that "language learning is largely lexical learning" and Cobb and Horst (2015:185) refer to the "centrality of vocabulary in acquiring language generally". As previously mentioned, Kaur and Hegelheimer (2007) claim that, in academic contexts, using academic vocabulary is central to students' success. Considering that academic vocabulary is an aspect perceived as particularly troublesome for Brazilian students, the focus of this investigation is academic words and their use.

In conclusion, this investigation seeks to answer the following questions

a) What is the vocabulary profile of assignments written by Brazilian students?
b) How does it compare to the vocabulary profile of other academic corpora presented in the literature review?
c) What words in the Academic Word List (AWL) do Brazilian students use?
d) How does the use of academic words differ between Brazilian students and students represented in the British Academic Written English (BAWE) corpus?

1.3 Organization of the study

In order to answer the questions related to how Brazilians use academic vocabulary, I have read previous studies and discussed how their results are relevant for this investigation in the literature review (Chapter 02). I have also compiled a corpus of assignments written by Brazilian students in the UK (Chapter 03) and analysed the use of academic vocabulary in this corpus using Sketch Engine and Range (Chapter 04). Finally, I draw some conclusions and pedagogical implications based on the results of this study (Chapter 05).

2. Literature Review

2.1 Introduction

This chapter examines previous research related to the topic of vocabulary and academic vocabulary in language learning. The next section (2.2) discusses the definition of what academic vocabulary is. The following section, then, provides a brief overview on studies that investigated the existence of a vocabulary threshold for reading general English and academic texts (2.3). The subsequent section (2.4) presents studies that argue in favour of teaching academic vocabulary. This chapter continues to present a timeline of studies, which compiled academic word lists (2.5), and, finally, it presents previous studies that investigated the Academic Word List (AWL) coverage across academic corpora (2.6).

Before discussing previous studies conducted on the area of academic vocabulary, it is important to present some definitions used in these investigations. Firstly, some studies refer to word families, which "consist of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (Bauer and Nation, 1993:253). The words *analysed*, *analysers*, *analyses*, *analysing*, *analysis*, for example, are all members of the word family *analyse*. According to Nation (2001) research on vocabulary relies on word families because if students know one or two words in a specific word family they are more likely to understand the other words in that same family. Other studies presented here refer to lemmas, which are "words with a common stem, related by inflection only, and coming from the same part of speech" (Gardner and Davies, 2014).

In addition, most of these studies use the General Service List (GSL) (West, 1953) as a reference, in spite of the criticisms which say that it is dated (Richards, 1974; Carter, 2012), it lacks in range (Engels, 1968), or it is built based on subjective criteria (Brezina and Gablasova, 2015). The GSL is an inventory of 2,000 high frequency words considered useful for learners of English as a foreign language. Sometimes, it is referred to as the first 2,000 most frequent words in English.

2.2 What is Academic Vocabulary?

According to Baumann and Grave's (2010) there is a plethora of explanations associated with academic vocabulary. Li and Pemberton (1994:184) state that academic vocabulary is the label given "to this group of words that occur across a range of academic disciplines". Townsend

and Kiernan (2015:113) add to this definition saying that "academic vocabulary words are typically abstract, technical, nuanced, and/or densely packed with meaning". Some researchers (Hyland and Tse, 2007; Gardner and Davies, 2014), however, claim that a small amount of the words in the AWL, for example, overlap with the 2,000 most frequent words in English putting into question, then, the distinction between academic and high frequency words. Snow (2010:450), however, argues that "there is no exact boundary when defining academic language; it falls toward one end of a continuum (...), with informal, casual, conversational language at the other extreme."

For the purposes of this study, it is worth distinguishing between general academic vocabulary and discipline specific (academic) vocabulary. General academic vocabulary, also referred simply as academic vocabulary, is a group of words that are frequent in academic texts across different disciplines. On the other hand, discipline specific vocabulary or technical words/terms (Fisher and Frey, 2008; Harmon, Wood and Medina, 2009) are recurrent words in one academic domain, for example, *abdominal, laboratory* and *melanoma* are more recurrent specifically in the field of health sciences (Lei and Liu, 2016). Beck et al (2002) calls discipline specific words Tier 3 words and general academic vocabulary Tier 2 words due to its frequency on academic texts.

In this study the AWL will be used as a reference of academic vocabulary, (further clarification on this decision is provided in section 2.6). McCarthy, O'Keefe and Walsh (2009) present the following explanation on English vocabulary:

> "English often has two kinds of words for the same thing: words whose origin lies in northern Europe (the Nordic and Anglo-Saxon world) and words which came from further south (the Mediterranean world - French, Latin and Greek words). Often, the Greek or Latin word for something is considered more formal than the Anglo-Saxon word for the same thing. Example include *commence* versus *start*, *ascend* versus *go up*, and *depart* versus *leave*." (p.07)

Words with origins in Greek or Latin are considered more formal insomuch that around 80% of the academic vocabulary in English comes from Latin or Greek roots (Coxhead and Nation, 2001; Coxhead, 2000; Farid, 1985). This might explain why students whose first languages derive from Latin, such as Portuguese, sometimes find it difficult to perceive if words like the ones presented in the table below are academic words in English or if they are using a Portuguese word in English.

Table 2 - Academic words x High frequency words x Portuguese

Academic Word	**GSL**	**Portuguese**
commence	begin	começar
obtain	get	obter
indicate	point out/show	indicar

Nevertheless, speakers of Portuguese as a first language are not the only students to whom academic vocabulary might be an obstacle. According to Vongpumivitch, Huang and Chang (2009) and Li and Pemberton (1994) learners of English as a foreign language have pinpointed vocabulary as one of the most difficult aspects of reading and writing for academic purposes. The studies below provide further evidence on this claim in EAP reading (Park, 2012; Clark and Ishida, 2005) and writing (Shaw, 1991; Townsend et al, 2012).

2.3 The role of vocabulary in language learning

This section will offer an overview into studies on the role of vocabulary in language learning. Several studies (Gass and Selinker, 2008; Cobb and Horst, 2015; Townsend and Kiernan, 2015) argue that having ample lexical knowledge grants better reading and writing skills for students of English as a Foreign Language (EFL). Research on the influence of vocabulary in school-aged children's reading comprehension (Garcia, 1991; Snow, 2010; Stahl and Fairbanks, 1986) confirms the impact of general academic vocabulary in students' achievement. Vocabulary acquisition, however, should be only one among a range of goals addressed in the foreign language classroom (Nation, 2013). Nevertheless, according to Nation (2013) it is possible to approach the skills of listening, speaking, reading and writing and focus on spelling, pronunciation and grammar through the viewpoint of vocabulary. In the EAP context, Townsend et al (2012:498) remind us that although some studies refer to "academic vocabulary as an independent construct, academic vocabulary, by definition, is academic because of its role within academic language". These authors continue by arguing that although research on academic vocabulary does not guarantee a full picture of students' ability when using academic English, it can be an entry point to the overall academic English proficiency (Townsend et al, 2012:503). Finally, given both these researchers

claims, it is the teachers' role to balance the study of vocabulary alongside other aspects of the English language in their classroom practice.

Regarding the correlation between vocabulary size and reading comprehension measures, several studies (Hirsh & Nation, 1992; Hsueh-Cho & Nation, 2000; Nation, 2013) attempt to determine a vocabulary threshold that would allow learners to develop an appropriate understanding of texts. In order to explore this threshold, Hirsh and Nation (1992) investigated the vocabulary size needed to read unsimplified texts for pleasure. These authors analysed the coverage of the GSL in three novels, and their findings suggest that students require a vocabulary of around 5,000 words families in order to read for pleasure. However, the researchers considered that pleasure reading meant encountering one unknown word per 3.3 lines, or knowing 97 to 98% of the words in a text and did not take into account students' actual understanding of the text. Therefore, although this investigation established that the GSL covered only 85 to 90% of the novels, and that students' would have to learn words beyond the GSL in order to understand these narratives, it did not take into account students' lexical knowledge based on reading comprehension tests.

Hsueh-Cho and Nation (2000) explored the relationship between the density of unknown vocabulary and the degree of text comprehension by testing sixty students' comprehension after reading four versions of the same text with different lexical densities. The first group of students read the complete text, while the second, third, and fourth group of students read the text with, respectively, 5%, 10%, and 15%, of the words changed into nonsense ones, in other words, the fourth group only knew 85% of the words in the text. The results of the reading comprehension tasks suggested that some students on the second group, those who knew 95% of the words, were not able to gain adequate comprehension of the text (Hsueh-Cho and Nation, 2000:415). Therefore, this study confirms the claim that students need to be familiar with 95% or more of the words in a text in order to understand it.

Although these two studies focus on non-academic vocabulary, they corroborate previous researchers' claims that it is necessary to have an extensive vocabulary in order to gain understanding of texts in a foreign language. Li and Pemberton (1994:183), for example, state that if students do not recognise high frequency words they might not be able to carry out top-down reading strategies as they will be employing "too much of the brain's capacity in trying to process this incoming information".

Considering a vocabulary threshold in academic texts, Laufer (1989:317) investigated the correlation between lexical knowledge and the quality of the reading comprehension in academic texts. This researcher asked 100 first year university students to read an academic text and answer

comprehension questions. Students were also asked to highlight any unknown words in the text. Her findings sustain the argument that in order to understand a text 95% of the words or more have to be familiar to the reader. This finding corroborates Coxhead and Nation's (2001) claim that learners of EAP, after acquiring the first 2,000 high frequency words in English, should focus on learning academic vocabulary to gain more text coverage with their vocabulary. According to them, "knowing the 2,000 high frequency words and the words in the academic word list will give close to 90% coverage of the running words in most academic texts. When this is supplemented by proper nouns and technical vocabulary, learners will approach the critical 95% coverage threshold needed for reading" (Coxhead and Nation, 2001:260).

Clark and Ishida (2005:228) investigated the differences between placed and promoted students in an advanced EAP in-sessional course. These authors discovered that, even though promoted students had taken EAP classes and attended lessons in their university courses for one semester, they did not improve in academic vocabulary tests. Furthermore, this finding corroborated teachers' perception that placed students had stronger linguistic foundations than promoted students. In addition, these authors state that "without sufficient knowledge of academic vocabulary, they (students) cannot deal with reading materials efficiently for various types of academic tasks given to them". This study confirms that academic vocabulary can be an indicator of students' overall academic English proficiency (Townsend et al, 2012).

Park (2012) conducted questionnaires and interviews with foreign students at the same university in which Clark and Ishida's (2005) study was carried out. These interviews showed that students attributed their low reading comprehension ability to their limited vocabulary size (Park, 2012). Furthermore, instructors interviewed for this research also pointed out that students displayed a "difficulty in understanding general academic vocabulary" (Park, 2012:10). Therefore, the instructors interviewed sustain Flowerdew's (1993) claim that content teachers are less likely to introduce general academic vocabulary to students.

Given the results of the studies presented above, it is clear that EAP students need to have a good command of not only the 2,000 most frequent words in English, but also of academic and technical vocabulary. Nevertheless, understanding vocabulary is not the same as producing appropriate vocabulary for academic texts, as students might be able to understand academic words but might not be able to use them in their own texts. The studies below present how students deal with academic vocabulary production.

Shaw (1991) interviewed 22 foreign students who were writing their theses or dissertations in English. These students declared that "vocabulary and finding the right word for the context"

(p.195) were their most significant problem when writing EAP. Much of the same response was given by the 128 foreign students surveyed by Leki and Carson (1994).

Regarding the impact of vocabulary knowledge in academic achievements, Townsend et al (2012:502) investigated the correlation between academic success and general academic vocabulary knowledge in a group of 339 students. These students took a general academic vocabulary test and the results were compared with their academic achievements. Students who scored higher in the vocabulary test also achieved higher marks in their content tests. Furthermore, Townsend et al (2012) analysis took into account students' linguistic background (English only or EFL). English-only students outperformed students whose first language was not English, supporting Li and Pemberton's (1994) claim that academic vocabulary is more problematic for non-native students.

Engber (1995), in his study of ESL students writing in an academic context in the USA, found that students who incurred in high amounts of lexical errors presented lower lexical density and lower lexical variation received worse grades. Thus, corroborating Santos's (1988) earlier findings that content lecturers judge lexical errors as the most serious of students' language errors and are more likely to penalize students' final grade based on it than other language mistakes.

The objective of this section was to present previous studies that investigated the vocabulary threshold for general and academic purposes and studies on students' difficulties with reading and writing for academic purposes. The next section presents arguments in favour and against teaching general academic vocabulary.

2.4 Why teach Academic Vocabulary?

Coxhead and Nation (2001) claim that general academic vocabulary should be emphasized in course plans and teaching materials because: a) academic vocabulary is common to a wide range of academic texts and generally not so common in other texts; b) academic vocabulary accounts for a significant number of words in academic texts; c) academic vocabulary is generally not as well-known as technical vocabulary, and d) English teachers can usefully help the learners with academic vocabulary, while this might not be true for technical vocabulary, in which the teacher probably do not have the necessary background to understand the words. Hyland and Tse (2007) and Durrant (2014), however, question the usefulness of a general academic word list. Hyland and Tse (2007) criticize especially argument "b" which says that academic vocabulary accounts for a large portion of words in an academic text. According to these researchers, it would be more helpful if students learned discipline specific vocabulary rather than general academic vocabulary.

They claim that a discipline specific vocabulary would cover larger portions of the texts and, therefore, would promote better reading comprehension and, as a consequence, active use when writing academic texts. Nevertheless, Trimble (1985) claims exactly the opposite. According to Trimble (1985), discipline specific vocabulary is not enough of a problem for students of EFL, due to its low frequency in academic texts. Corroborating Trimble's (1985) argument, previous investigations (Campion and Elly, 1971; Praninskas, 1972) uphold the existence of a compendium of academic words that perform a supportive role in academic texts. Besides, previous studies conducted on the coverage of academic words in discipline specific corpora (see section 2.6), such as Chung and Nation's (2003) research on anatomy and applied linguistics textbooks and Chen and Ge's (2007) study on a corpus of medical research articles confirm the supportive role of academic vocabulary in academic texts. Furthermore, considering the applications of vocabulary compendiums into teaching, most EAP classrooms are not discipline specific (Durrant, 2016), therefore a general academic word list might be more useful as a resource for English teachers. Finally, as Coxhead (2011:357) points out, "a number of subject-specific vocabulary lists have been developed recently to address the needs of particular learners (e.g., Wang, Liang, & Ge, 2008; Coxhead & Hirsh, 2007; Ward, 2009; Chung, 2009)" among other lists compiled after the publication of Coxhead's article, such as Lei and Liu's (2016) medical academic word list and Yang's (2015) nursing academic word list. These lists can be used to supplement the vocabulary of EAP students with regard to their specific area of study/research.

Hyland and Tse (2007) also question the suitability of a general academic word list to all fields of expertise based on the fact that even if students were to acquire a general academic vocabulary, words might assume different meanings according to their fields of study. They exemplify this claim with the word "attribute", which is recurrent in linguistics as the verb "to accredit" and in economics as the noun "feature" (Hyland and Tse, 2007:245). Nevertheless, according to Nagy and Townsend (2012:96) this feature is part of the nature of academic vocabulary, its "dictionary entries tend to include many definitions". In addition, Townsend et al (2012) point out that meaning differentiation also occurs in high frequency words. Therefore, teachers can make these distinctions salient to students either by using specific vocabulary lists in a supportive role or even with students searching for occurrences of the academic words in texts from their areas of study.

Finally, several researchers (Flowerdew, 1993; Corson, 1997; Snow, Lawrence, and White, 2009) argue that content teachers might target discipline specific vocabulary for students whereas they are less likely to do the same with general academic words. While other researchers (Cowan, 1974; Hutchinson and Waters, 1987; Farrell, 1990) also sustain that EAP teachers might not be

able to target this discipline specific vocabulary and that EAP learners would not encounter difficulties with this vocabulary. These claims supports the view presented on argument “d” which mentions that English teachers can help learners with a general academic vocabulary and might not be as helpful with technical vocabulary.

2.5 Academic word lists: a timeline

The objective of this section is to discuss the evolution of academic word lists throughout the years, taking into consideration the methodology used to create these lists and their applications. The table below presents nine general academic word lists compiled from 1971 (Campion and Elley) to 2014 (Gardner and Davies; and Browne, Culligan and Phillips).

Table 3 – A timeline of academic word lists

Author/ Year	Name	Outcome	Procedure	Applications	Access
Campion and Elley, 1971	Academic Vocabulary List	Two lists, one containing the complete academic vocabulary with 3,200 words and the other containing the top 500 words in the academic vocabulary.	This list was compiled based on a frequency word count of university textbooks and high school students rating of these words difficulty if encountered in a text. The final list also excludes the 5,000 most frequent words in English according to Thorndike and Lorge (1944).	This list used to be the reference for the academic section of the Vocabulary Level Test (Nation, 1983).	
Praninskas, 1972	American University Word List	This list contains 507 word families organized in alphabetical order, frequency of each word form or frequency of word families.	It was compiled based on the texts of ten textbooks used in first year undergraduate classes in the courses of chemistry, biology, psychology, physics, English, literature, sociology and history at the American University of Beirut. The manual corpus used to create this list contained 30,844 running words. The final vocabulary lists excludes proper nouns, dates, foreign words, abbreviations, discipline specific words and words that appeared in the GSL.	Textbooks (Yorkey, 1981; Farid, 1985; Valcourt and Wells, 1999).	
Lynn, 1973	Word list	This list contains 10,000 types organized according to frequency of occurrence.	In order to compile this list, Lynn (1973) checked 52 textbooks and 4 cycle-styled handouts used at the University of Nanyang looking for the words which students annotated or wrote the translation in their academic texts.		
Ghadessy, 1979			Ghadessy (1979) also looked into student's annotations on textbooks, examining a manual corpus of student's textbooks with more than 478 running words.		

Author/ Year	Name	Outcome	Procedure	Applications	Access
Xue and Nation, 1984	A University Word List (UWL)	This list contains 836 word families	Xue and Nation (1984) combined the items on the four previous word lists (Campion and Elly, 1971; Praninskas, 1972; Lynn, 1973 and Ghadessy 1979), excluding the words in the GSL.	This list used to be the reference for *Range* and the *VocabProfile* tool in LexTutor.	Available at https://www.learnthat.org/word_lists/view/7797
Coxhead, 2000	A new Academic Word List (AWL)	A list containing 570 word families divided into 10 frequency-ranked sublists. Therefore, the first sublist accounts for one third of the words in an academic corpus, while the second sublist accounts for half of this amount. Each sublist, with the exception of the last one, contains 60 words families.	This list was compiled based on an academic corpus of 3.5 million words. The texts in this corpus included the areas of arts, commerce, law, and science. The criteria to include the words in the AWL were: frequency - the words had to occur more than 100 times; range - the words had to occur at least 10 times in each subcorpora; and specialised occurrence - the words could not be part of the GSL.	- The Vocabulary Level Test (Schmitt, Schmitt and Clapham, 2001). - LexTutor - Vocabpofile (Cobb, n.d.). - Textbooks (Savage and Mackey, 2010; Huntley 2006; Zimmerman et al, 2012; Schmitt and Schmitt, 2011; Mifflin, 2006). - Oxford Student's Dictionary. - Longman Exams Dictionary.	Available at http://www.victoria.ac.nz/lals/resources/academicwordlist/
Paquot, 2010	Academic Keyword List (AKL)	The AKL contains 930 potential academic words divided into nouns, verbs, adjectives, adverbs, and others.	This list was based on two corpora of professional writing - Micro-Concord corpus collection B (Scott and Johns 1993) and the Baby BNC Academic Corpus - and two corpus of student writing - the Louvain Corpus of Native Speaker Essays and the British Academic Written English Pilot Corpus. The criteria to select the words in this list were: keyness - the words had to appear significantly more in the academic corpus than in a comparison corpus of fiction; frequency, range, and distribution.	Louvain EAP Dictionary (Granger and Paquot, 2010).	Available at https://www.uclouvain.be/en-372126.html

Author/ Year	Name	Outcome	Procedure	Applications	Access
Gardner and Davies, 2014	A new Academic Vocabulary List (AVL)	This is a list of the 3,000 top lemmas occurring in all academic domains of the Corpus of Contemporary American English. A version of this list organized in word families for research and teaching purposes is also available.	Gardner and Davies' (2014) AVL derives from 120 million words subcorpus of academic English in COCA. The words were included in this list based on: ratio - words had to occur at least 50% more in the academic subcorpus than in other non-academic subcorpora; range - words had to occur in seven of the nine academic disciplines with more than the expected frequency; dispersion - the word had to occur with a dispersion above 0.80; discipline measure - the word could not occur more than three times the expected frequency in any of the disciplines.		Available at http://www.academicvocabulary.info/
Browne, Culligan and Phillips, 2014	New Academic Word List (NAWL)	The NAWL contains 963 lemmas derived from a 288 million-word corpus. This list is available in three different forms: lemmas, headwords only and based on frequency indices.	The corpus used to compile this list incorporated the academic subsection of the Cambridge English corpus, best-selling textbooks, the MICASE corpus and the BASE corpus. Words were selected based on frequency, dispersion and appropriateness. In addition, words in the New General Service List (NGSL, Browne, Culligan and Phillips, 2014) were excluded.	- LexTutor - Complete Vocab Profile (Cobb, n.d.). - New Academic Word List Test (Bennett and Stoeckel, 2015).	Available at http://www.newacademicwordlist.org/

The first point to consider regarding this table is that almost all of the word lists presented here, with the exception of Lynn (1973), Ghadessy (1979) Paquot (2010), and Gardner and Davies (2014), exclude high frequency words whether by eliminating GSL words or words in other general lists, such as the New-GSL or the most frequent words in the *Teacher's book of words* by Thorndike and Lorge (1944). However, Paquot (2007) questions the exclusion of GSL words, the findings in her investigation indicate that two thirds of the nouns in her AKL are also in the GSL, while only one third of these nouns occur in the AWL. In addition, Gardner and Davies (2014) argue that it is not useful to distinguish between academic and high-frequency words, as academic words make an additional contribution to general vocabulary. Therefore, these authors chose not to associate the AVL with any general English word list.

Furthermore, in Lynn's (1973) and Ghadessy's (197) word lists students' annotations and translations written in their textbooks were taken into account, therefore these lists do not omit GSL words as these words might be relevant to those students. The methodology adopted to build these two lists is somewhat questionable since students' annotations do not necessarily mean that they did not understand the words. Another aspect to consider is that Lynn's (1973) word list was intended specifically for the context of students at Nanyang University in Singapore. Nevertheless, it was used as a reference, together with Campion and Elley's (1971) Academic Vocabulary List, Praninskas' (1972) American university word list and Ghadessy's (1979) word list, to form the University Word List (1984). Therefore, although the UWL was widely used in vocabulary research before Coxhead's AWL, it also combined the methodological weakness of all the previous lists. Coxhead's (2000) AWL was the first general academic vocabulary compendium to rely on computer tools in order to determine frequency, range and distribution on its compilation.

A second point to be noted is the use of non-professional writing in the reference corpus, Paquot (2010) and Browne, Culligan and Phillips (2014) were the only authors to include students' writing in the corpus that originated their word lists. Nesi (2008:03) argues that "novice writers do not begin by writing for publication, and their early attempts at academic writing are likely to be assessed texts produced in the context of a course of study" therefore, pointing out to the need to consider the texts of novice writers in the production of academic word lists. Browne, Culligan and Phillips (2014) were also the first researchers to include academic spoken English in their reference corpus.

Finally, we have to consider the dichotomy between word families and lemmas in more modern vocabulary lists. Although word families have been criticised for not taking into account parts of speech and, also, because derived and inflected words are not always related in meaning to the headword in a family (Gardner and Davies, 2014), they are extensively used in the

production of word lists, as we can perceive by the table above. Even the AVL, originally composed of lemmas, has a version in word families. Browne, Culligan and Phillips (2014) are the only authors not to make available a word family or word form version of their academic word list, nevertheless they state on their website that this list is still under development, hence there is a possibility that in the future a version with word families, or word forms will be made available. It is also possible to consider that recent word lists are also available as word families due to the easiness to explore word families in a corpus.

Despite the fact that the AWL has been criticised for being dated (Gardner and Davies, 2014) especially because it was compiled on top of the GSL (West, 1953), for using word families (Nagy and Townsend, 2012) and for representing general academic English (Hyland and Tse, 2007) it is still widely used as a reference for writers of teaching materials, EAP researchers, and it is also a source for The Vocabulary Level Test (Schmitt, Schmitt and Clapham, 2001). Furthermore, several researchers have explored the AWL coverage in different academic corpora, as it is presented in the next section, while more recent word lists, such as the AVL, have only recently started to be independently evaluated (Durrant, 2016). Hence, for the reasons listed above I use the AWL as a reference in this study.

The aim of this section was to discuss the development of word lists and justify the use of the AWL in this study. In the next section, I present previous studies that investigated the AWL coverage in academic corpora.

2.6 AWL coverage in other corpora

This section discusses the results of previous studies, which have explored the AWL coverage in academic corpora, written or spoken. The table below presents the procedures used on these studies, the basis for the corpus compilation, the subcorpora divisions, the number of words and the AWL coverage

Table 4 - Academic Word List Coverage

Study	Procedure	Corpus Compilation	Corpus	Subcorpus	Number of Running words	AWL coverage
Coxhead, 2000	Coxhead analysed the AWL coverage in three different corpora: one academic corpus divided into four subcorpora, another academic corpus and one corpus of fiction texts.	Coxhead compiled a corpus of research articles, university textbooks, textbook chapters, sections of the Wellington Corpus of Written English (Bauer, 1993), Brown Corpus (Francis & Kucera, 1982), the Lancaster-Oslo/Bergen Corpus (Johansson, 1978), the MicroConcord academic corpus (Scott and Johns, 1993) and laboratory manuals.	1st academic corpus	Science	875,000	9.10%
				Arts	875,000	9.30%
				Commerce	875,000	12%
				Law	875,000	9.40%
				Total	*3,513,330*	*10%*
		The second subcorpus was compiled in order to evaluate the AWL.	2nd academic corpus		678,000	8.50%
		The subcorpus of fiction contained 50 texts from Project Gutenberg's collection of texts.	Corpus of fiction		7,763,733	1.40%
Chung and Nation, 2003	The aim of this study was to compare the frequency of technical vocabulary in applied linguistics and anatomy.	One textbook of each area was used for the analysis.	Discipline specific academic corpus	Anatomy	11,356	8.60%
				Applied Linguistics	5,137	17.40%
Cobb and Horst, 2004	These authors wanted to test the AWL reliability and coverage in different academic corpora and across different disciplines.	Text segments from the *Learned* section of the Brown corpus (Francis & Kucera 1979).	Academic corpus	Zoology	2,026	7.31%
				Linguistics	2,031	12.60%
				Sociology	2,084	13.44%
				History	2,036	14.49%
				Social Psychology	2,059	14.38%
				Development	2,023	12.26%
				Medicine (anatomy)	2,024	6.72%
				Total	*14,283*	*11.60%*
		Randomly selected Reader's Digest non-fiction articles.	Non-academic Corpus			5.56%

Study	Procedure	Corpus Compilation	Corpus	Subcorpus	Number of Running words	AWL coverage
Hyland and Tse, 2007	This study tests the AWL coverage in professional and students writing.	The corpus compiled for this study consisted of research articles, textbook chapters, academic book reviews, scientific letters, dissertations, theses and final-year undergraduate project theses in three different areas: sciences, engineering and social sciences.	Academic corpus	Engineering	551,891	11.10%
				Social Sciences	1,822,660	11.00%
				Sciences	838,926	9.30%
				Total	3,213,477	10.06%
Chen and Ge, 2007	The aim of this study was to investigate the coverage of the AWL in health science studies.	This corpus consisted of 50 research articles on medical studies.	Discipline specific academic corpus		190,425	10.07%
Konstantakis, 2007	In this study Konstantakis (2007) evaluates the AWL and the GSL coverage in an academic corpus of business English and then proposes a Business Word.	This study used Nelson (2000) corpus of Published Material which encompasses 33 Business English textbooks.	Discipline specific academic corpus		600,000	4.66%
Ward, 2009	In this study Ward (2009) compiles a word list for engineering students and compares the coverage of the AWL and the BEL in an engineering corpus.	The corpus used for this study contained random pages of 25 textbooks in engineering faculties.	Discipline specific academic corpus		250,000	11.30%
Martinez, Beck and Panza, 2009	Martinez, Beck and Panza, 2009 investigate which words in the AWL are more relevant to agriculture researchers.	This corpus consisted of research articles in the agricultural sciences.	Discipline specific academic corpus		826,416	9.06%

Study	Procedure	Corpus Compilation	Corpus	Subcorpus	Number of Running words	AWL coverage
Vongpumivitch, Huang and Chang, 2009	This study investigates the most frequent AWL words in a corpus of applied linguistics research articles.	200 research articles from five applied linguistics journals have been collected in order to compile the corpus for this investigation.	Discipline specific academic corpus		1,500,000	11.17%
Li and Qian, 2010	The aim of this study was to profile the presence of AWL items in a financial corpus.	This study used the Hong Kong Financial Services Corpus which has 25 different text genres.	Discipline specific academic corpus		6,300,000	10.46%
Dang and Webb, 2014	Dang and Webb, 2014 investigated the academic profile of academic spoken English.	These researchers used the BASE corpus for the analysis. This corpus consists of 160 lectures and 39 seminars.	Academic corpus	Arts and Humanities		3.82%
				Life and Medical Sciences		4.27%
				Physical Sciences		4.28%
				Social Sciences		5.21%
				Total	1,658,403	4.41%

As can be seen in the table above, some of the studies represent the coverage in general academic corpora and some investigated the coverage in discipline specific corpora. Furthermore, only one study takes into account spoken academic corpus (Dang and Webb, 2014). However, few studies have investigated the AWL coverage across general academic corpora, the first one being Coxhead (2000), who tested the AWL coverage in two academic corpora. It is possible to notice that in the second academic corpus the AWL coverage is lower than the first one, and this can be attributed to two aspects: the first one is that the first academic corpus was the basis for the AWL compilation, thus it might have influenced the high coverage. The second one is that the second academic corpus was skewed as it overrepresented the science subcorpora and this discipline presented the lower AWL coverage in the analysis of the first corpus. Coxhead and Nation (2001) claim that the AWL covers 10% of academic texts, however they did not test it in an independent academic corpus. Cobb and Horst (2004) were the first ones to test the AWL coverage in different disciplines. Their study might be used as a reference in order to verify the coverage of the AWL across different study areas. Cobb and Horst (2004) research supports the argument that the AWL covers around 10% in a general academic corpus. Nevertheless, it is important to consider that their corpus was composed of parts of the *Learned* section of the Brown Corpus which was also used in Coxhead's (2000) academic corpus. Therefore, until Hyland and Tse's (2007) analysis, the AWL coverage had not been tested in an independent academic corpus. These researchers verified the AWL coverage in a corpus of professional and learner writing built specifically for this purpose. Their corpus shows an even higher AWL coverage than Coxhead (2000), confirming, then, the supportive role of the AWL in academic texts of different disciplines. Recently, Dang and Webb (2014) have also attempted to verify the AWL coverage in an academic corpus, however their corpus consisted of spoken English. This study highlights the differences between written and spoken academic English as the AWL coverage is significantly lower (4.41%) in the spoken corpus.

It is important to highlight that the AWL does not cover all disciplines to the same extent; in Coxhead's (2000) study the AWL coverage in the discipline of commerce was 3% higher than in sciences, while in Cobb and Horst's (2004) investigation of the *Learned* corpus, the AWL covered only 6.72% of the medicine corpus and 14.49% in the History corpus. Finally, in Hyland and Tse's the "sciences" subcorpus is also the corpus with the lower coverage (9.30%) while engineering's coverage was 3% higher.

Most of the studies (Chung and Nation, 2003; Chen and Ge, 2007; Konstantakis, 2007; Ward, 2009; Martinez, Beck and Panza, 2009; Vongpumivitch, Huang and Chang, 2009 and Li and Qin, 2010), reported in the table above, investigated the AWL coverage in discipline specific

corpora in order to test its validity in a specific field of study or to build on top of the AWL a technical word list. All these studies are relevant as they support the view that the AWL is a useful word list across disciplines. They also confirm that the AWL usually covers a mean of 10% of an academic corpus, except for Konstantakis (2007) and Dang and Webb (2014) corpus of spoken English.

The objective of this section was to present previous studies, which analysed the AWL coverage in academic corpora as this is one of the objectives of this study. In the next section, I will discuss the corpus compilation and the methodology used in order to analyse the AWL coverage in the Brazilian Academic English corpus.

3. Methodology

3.1 Introduction

As previously mentioned, this study uses corpus linguistic tools to investigate the academic vocabulary used by Brazilian students in written assignments. In this section, I will present corpus linguistics and its applications as a methodology (3.2), explain the relevance of compiling a specialized corpus in order to answer the research questions (3.3), and discuss the ethical issues involved in the data collection (3.4). In addition, I will describe the process of compilation (3.5). Finally, the last section explains the steps taken in the analysis of the corpus (3.6).

3.2 Corpus Linguistics

In this study corpus linguistics is adopted as a methodology, regarding this issue McEnery and Hardie (2011:02) claim that 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". Subsequently, these authors present the difference between corpus-based studies and corpus-driven studies. According to them, corpus-based studies view corpus linguistics as a methodology used to explore a theory or a hypothesis, while corpus-driven studies consider "the corpus itself as a source for hypothesis about language" (McEnery and Hardie, 2011:03).Thus, this dissertation is a corpus-based study as it sets out to explore how Brazilian students use academic vocabulary, using corpus linguistic tools based on a pre-existing list of academic words, the AWL, rather than using the corpus itself as a source to create a list of academic vocabulary used by Brazilian students.

Furthermore, Biber, Conrad and Reppen (1998:01) 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 (2011:01) 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 (2002:76) 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 the study presented here, 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 (1998:04) 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 (2002:77) 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, in the case of this study the quantitative part refers to the percentage of AWL coverage in texts, while the qualitative part refers to the analysis, through concordance lines, of how students use the academic words in their texts

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 (2004:13) highlights the role of corpus linguistics in providing "attested examples of language patterns based on empirical data". Several researchers (Conrad, 2002; Connor and 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. For the purposes of this study, I have used Range and Sketch Engine - two electronic resources - to investigate the use of academic vocabulary in texts written as coursework by Brazilian students.

The aim of this section was to provide a brief overview of the characteristics of corpus linguistics studies. The next section discusses the relevance of a specialized corpus to answer the research questions in this study.

3.3 The case for a specialized corpus

The aim of this section is to expose the motives to compile a specialized corpus for the purposes of this study and to introduce what the literature on the field says about the compilation of a specialized corpora. The corpus compiled for this study is 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 (2002:14) argues that specialized corpora "aim to be representative of a given type of text and it is used to investigate a particular type of language". Nevertheless, in this study, more than one type of text is represented, yet texts collected for the corpus are examples of the same language variety, that is, Brazilian academic writing.

Connor and Upton, (2004:02) 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.

Hoey (2007:10) 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, such as academic vocabulary.

In addition, Connor and Upton (2004:02) 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 as "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". Further attention is given to the issues of sampling and genre in section 3.5.2.1.

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, 2004:14). Considering the arguments presented above, it is possible to assume that in order to answer the research questions introduced in section 1.2 a specialized corpus of Brazilian students' academic writing needed to be compiled.

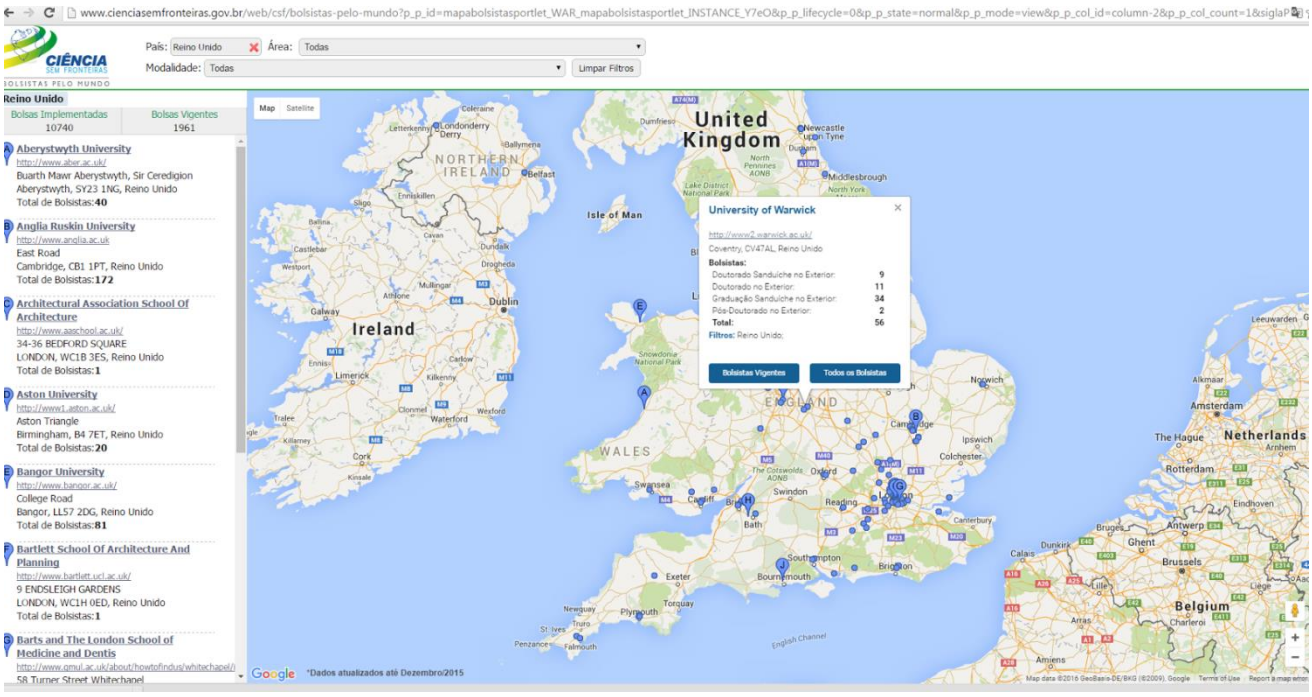
The aim of this section was to explain the reasons why I have compiled this specialized corpus of Brazilian academic writing. The next section presents the ethical issues and the data collection process.

3.4 Data collection

Since this study involves people, it is reasonable to take into consideration issues concerning ethics and students participation. Therefore, 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.

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

Figure 1 - SwB webpage



Once you click on a university, the page shows the number of Brazilian students enrolled, for example, in the figure above I have clicked on the blue dot representing Warwick University. 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 my data collection,

I selected “all scholars”, however, soon I encountered some obstacles with this course of action on the webpage that made me opt to contact only current scholars.

The main difficulty that I encountered 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 me to send emails through the page, however I could not actually have access to students’ emails. Therefore, I was limited to contact only five students daily, and this limit is arbitrarily settled by the webpage in order to avoid phishing and spam. In order to circumvent this problem, I used four different email accounts (two institutional and two personal ones) to contact students, thus I could send twenty emails a day instead of only five. Nevertheless, this was still a significant small amount considering that my goal was to contact all scholars. Hence, I have decided to focus on current scholars because this group was more likely to reply to my email. I 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, I wanted my corpus to be as similar as possible to BAWE, which only contains texts of undergraduates’ and masters’ students.

Notwithstanding, I still wanted as many students as possible to participate in my research, therefore I have contacted students through Facebook groups of Brazilians in the UK. I expected that in this informal setting more students would reply. In addition, in the final stages of my 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 my research on their weekly newsletters and put me in contact with former scholars. This way I have received texts of participants who had been SwB scholars from 2014 until 2016. Finally, I also went to the Brazilian Association of Students and Researchers in the UK annual conference to talk to these students who are here and then I received more assignments from students participating in other mobility programmes and from other fields of study (arts and humanities, and social sciences).

Regarding the email sent to students, it consisted of a text in English explaining the background and the objectives of the study, and it also described how students could participate in the study (Appendix 01). In addition, I wrote a more informal text in Portuguese explaining who I was and summarizing the research goals, and this text in Portuguese was used as an introduction to the e-mails sent. In this passage, I 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.

Taking into account the consent form, as I was not going to meet the participants in person and I wanted to make the process of participation as easy as possible for students, instead of asking participants to print, sign and scan a consent form, whenever a student agreed to participate in the research, they were asked to send their texts through their personal e-mail rather than signing a formal consent form.

Finally, I would like to point out some lessons learned through my data collection. The first one is to always contact first the stakeholders; it took me some time to approach the Brazilian Embassy and if I had talked to them sooner my data collection would have finished remarkably earlier. The second thing that I learned is to make clear for the participants that the research does not involve any kind of judgement towards their proficiency; some participants were embarrassed to send their texts as they believed I would judge their writing. The third lesson learned is to always use my institutional email if possible; some students only replied to my second attempt to contact them because the second email was through my university account and that was more trustworthy, according to them.

The aim of this section was to present the measures taken in order to contact students and collect their written assignments. The next section explains the corpus compilation in detail.

3.5 Corpus compilation

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 670,314 tokens in Range (and 768,323 tokens in Sketch Engine as this tool counts punctuations marks as tokens), the next section provides further information about the corpus compiled for this study.

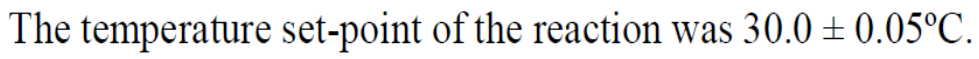
3.5.1 Cleaning the data

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, Range and Sketch Engine, the softwares used in this study, only read TXT files. Hence, in order to proceed with the analysis using these softwares 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, first 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 account and the AWL coverage. 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 ±, I have decided to leave these occurrences 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.

Figure 2 - Example from text NWUT01I184



In addition, textual features like titles, headings, acknowledgements, summary, block quotes, genuine lists, references, and appendix were excluded from the TXT file. However, as I would like the corpus to be useful for other research purposes in the future, all these textual features were kept in the PDF file version of the corpus.

Finally, after cleaning the texts and saving them in TXT and PDF, the text files were labelled. This naming of the files helped me 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 final 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 and some of the universities that are not represented only received PhD students, therefore they could not contribute to the corpus. 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), further explanation on these subcategories is provided in sections 3.5.2.1.

In this section, I have discussed the process of data cleaning and provided a brief overview into the division of the corpus. In the next section, I will present how the corpus is organized in further detail as well as discuss the issues of sampling, balance and representativeness.

3.5.2 Corpus organization

According to Conrad (2002:77) "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 (2006:16) 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, 2006:20). 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 (2011:10) 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 (2011:10) 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 (2009:73) 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."

The same issue of sampling applies for the corpus compiled for this study since I did not know the 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, at first I did not know if I would receive 2, 10, or 100 texts. Thus, the final product of this corpus compilation is associated to what McEnery and Hardie (2011:11) 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 (2005:81) 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."

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 (2004:25) set of general guidelines for building a specialized corpus.

3.5.2.1 Contextualization

As previously stated, 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 the corpus presented here are significantly smaller than the other two, however I have decided to include the texts collected from these areas in the corpus so that these partitions might be complemented with more texts in the future.

Additionally, the texts were divided into genre families in order to get more general information about the corpus. I also believed that the text genres represented in the corpus might influence the results, even though the use of academic vocabulary across genres is not taken into account in this study. 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 of classifying the texts into genres was carried out by only one person, the researcher, and with limited time, I have tried to go through the texts twice comparing

the text features with the genre descriptions provided by these authors. Furthermore, whenever I was in 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 compiled 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 I required to include the texts in the corpus was that the assignment had received at least a pass, yet I did not ask for any proof for the grade assigned to the assignments.

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 corresponds 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. It is important to highlight that although the table shows the numbers for SS and AH, these fields of study were not taken into account in the qualitative part of this study.

Table 5 - 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 can be seen 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.

3.5.2.2 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, the corpus compiled for this study has two main purposes: the first one is to provide empirical data towards the use of academic vocabulary by Brazilian students, and the second one is to serve as a resource in future studies of Brazilian Academic writing.

3.5.2.3 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, 2004:26). 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 can be considered as a medium-size corpus.

Now, 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), 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. Therefore, the corpus seems to be of appropriate size to investigate the linguistic feature, which is the focus of this study, that is, academic vocabulary.

In this section, I have explained the corpus compilation, data cleaning, and how the corpus is organized. The aim of the next section is to explain how I proceeded with the analysis.

3.6 Data analysis

In order to answer the first and second research questions, I have analysed the lexical profile of Brazilian students' assignments using Range (Nation and Heatley, 2002). This software allows the researcher to compare the vocabulary of up to 32 text files. Nation and Heatley (2002) explain that "for each word in the texts, it (Range) provides a range or distribution figure, a headword frequency figure, a family frequency figure, and a frequency figure for each of the texts the word occurs in". However, the function used in this study is the comparison of the words in a corpus to a pre-selected word list, in other words this procedure provides the coverage of a word list in a set of texts inserted in the software. In the study conducted in this dissertation, the software has given me the percentage of academic words used in the texts written by Brazilian students. Therefore, I have used the AWL and the GSL as baselists as they are already available in the software.

Even though Range is useful to determine the vocabulary profile of large quantities of text, it only provides percentages, and to answers questions three and four I needed qualitative information regarding the use of AWL words in the corpus. For this purpose, I used Sketch Engine to check the frequency and the concordance lines of AWL words in the assignments. In order to do that I have used the whitelist tool in Sketch Engine, when a whitelist is added to this website the outcome of a word list shows only the frequency and the occurrences of the words in this whitelist. Hence, I have added the AWL as a whitelist. Furthermore, to compare how academic words are used in contexts in BAWE and in the corpus of Brazilian students writing I have selected the 40 highest frequency academic words in BAWE and in the Brazilian EAP corpus and analysed

the concordance lines for the top 3 highes frequency words in both corpora. In addition, I have explored the word forms selected by Brazilian students.

Finally, I would like to highlight that even though I am using BAWE as a reference corpus in this study the aim of this research is not to propose a comparison between native speakers and non-native speakers, since BAWE represents a group of students in British universities who have achieved high marks, regardless of their nationality. Therefore, the comparison in question here is how the use of vocabulary in merit and distinction assignments differs from Brazilian students assignments.

The aim of this section was to explain how the analysis of vocabulary was conducted the next section will present the results of the data analysis and discussion answering each research question.

4. Results and Findings

4.1 Introduction

The aim of this chapter is to present the results of the corpus analysis. The next section presents the percentage of the AWL coverage in Brazilian students' assignments. The following section depicts the frequency and the use of the AWL words in both corpora, BAWE and the corpus of Brazilian students' writing, with reference to the concordance lines of the most frequent words. The next section shows the different AWL word forms selected by both groups of students. Data analysis and discussion are displayed together in this section to help answer the research questions.

4.2 The AWL coverage

This section aims at answering the first two research questions, presented in section 1.2

a) What is the vocabulary profile of assignments written by Brazilian students?

b) How does it compare to the vocabulary profile of other academic corpora presented in the literature review?

It is worth remembering that according to Coxhead and Nation (2001) the lexical profile of an academic text is composed of roughly 80% of high frequency words (GSL), 10% of academic words (AWL), 5% of technical words, and 5% of low-frequency words. In addition, previous studies, presented in section 2.6, corroborate, to some extent, this claim. In order to determine the lexical profile of the corpus of Brazilian students I have created table 06 below, which depicts the percentage of academic words written by Brazilian students. The results were generated using the software Range. Therefore, this table presents the coverage of the AWL, the GSL and off-list words in the corpus of Brazilian students divided per genre and area of study. In the first row the number between brackets represents the number of tokens in each subcorpora.

Table 6 - AWL coverage in the corpus of Brazilian students

	AH (7,887)			SS (67,907)			LS (224,979)			PS (369,541)			Total (670,314)		
	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)	GSL (%)	AWL (%)	OFF (%)
Case Study				78.60	9.70	11.70	74.10	8.90	17.00	79.00	10.40	10.50	77.60	9.90	12.50
Critique				79.70	11.00	9.40	77.00	11.00	12.00	77.20	11.10	11.80	77.60	11.00	11.40
Design										78.70	10.40	10.80	78.70	10.40	10.80
Empathy Writing															
Essay	81.70	5.20	13.10	80.60	9.40	10.00	69.50	9.60	20.90	77.30	10.50	11.30	73.90	9.60	16.50
Exercise				86.40	5.20	8.40	70.80	8.40	20.80	78.70	9.90	11.40	77.60	9.50	12.90
Explanation							73.00	8.00	19.10	78.60	8.90	12.50	77.90	8.90	13.20
Literature Survey							71.00	9.30	19.80	62.40	7.50	30.10	69.30	8.90	21.80
Methodology Recount							69.30	8.20	22.50	79.20	9.40	11.30	75.60	8.90	15.50
Narrative Recount							74.50	11.80	13.80	80.90	10.40	8.60	78.40	10.90	10.60
Problem Question							78.60	12.40	9.10	78.50	7.80	13.70	80.50	9.70	9.80
Proposal							76.30	11.30	12.40	77.80	12.40	9.80	77.40	12.20	10.30
Research Report							69.10	9.10	21.80	75.40	10.70	13.90	73.10	10.20	16.70
Total	81.70	5.20	13.10	80.30	9.80	9.90	71.30	9.40	19.30	77.90	10.20	12.00	**76.00**	**9.80**	**14.20**

On previous investigations, the AWL covered around 9 to 11% of academic written English corpora, with some small variations that could be attributed to discipline specificities as depicted in Cobb and Horst (2004) and Konstantakis (2007). It is possible to perceive that in the corpus of Brazilian students, the AWL covers around the same percentage of words in all subcorpora except for Arts and Humanities, however this result should be taken as something to be explored in the future, since the corpus of AH contains only 7,887 words, thus not being representative enough to allow for generalizations.

Taking into account the distinct use of academic vocabulary in different disciplines, previous studies, like the one by Cobb and Horst (2004), have reported a disparity on the AWL coverage between the corpus of Medicine (6,72%) and Sociology (13,44%). Thus, these results suggest that the AWL covers different percentages of a corpus depending on its discipline. However, it is difficult to compare the coverage of each of the subcorpora compiled for this study with previous investigations due to the fact that discipline division is not always clear in these past studies. Hyland and Tse (2007) and Coxhead (2000), for instance, report the AWL coverage on two different "science" subcorpora, nevertheless "science" is a term that encompasses many fields of study and these authors do not clarify which disciplines are included in their "science" subcorpora. Hence, the next paragraphs are an attempt to contrast the AWL coverage in the Brazilian subcorpora of PS and LS with the most similar corpora found in previous studies.

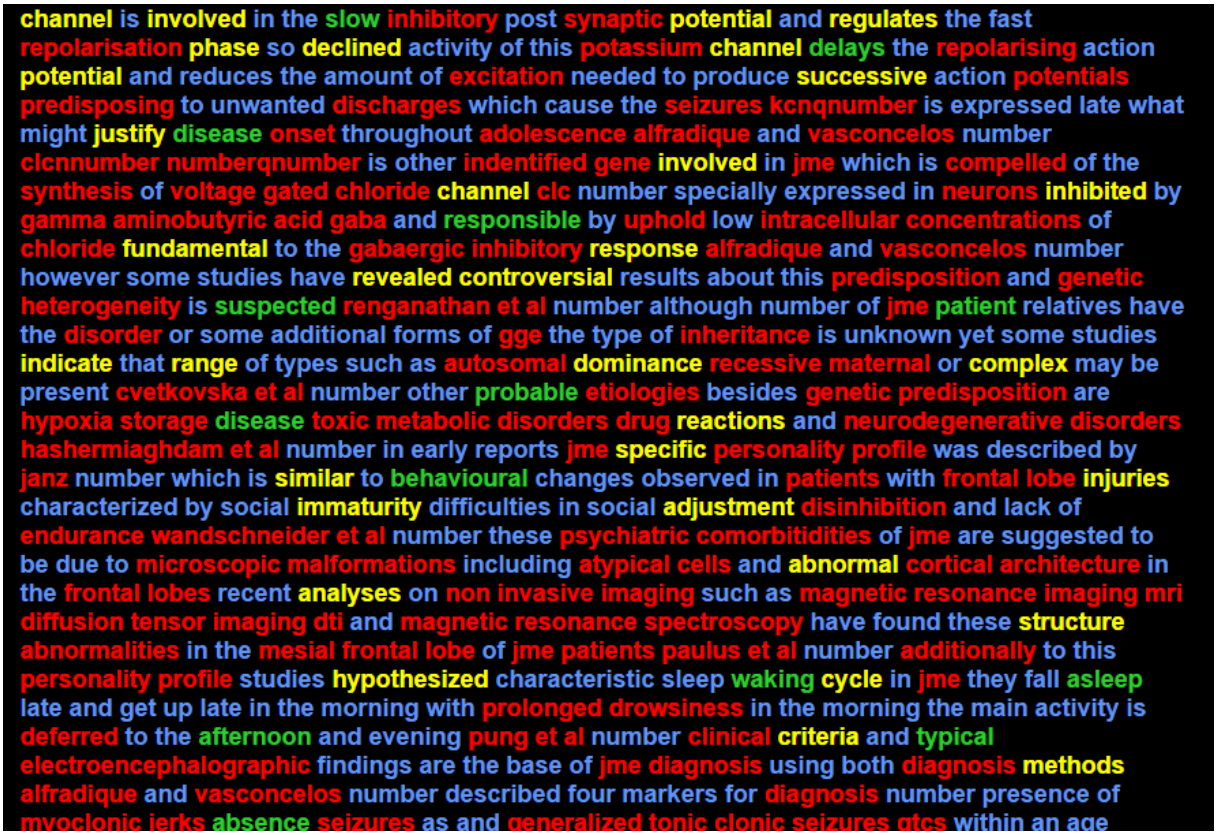
It is possible to notice that overall the use of academic words in these two subcorpora, LS and PS, falls on the expected average. In addition, similar to previous studies, the Life Sciences (Chen and Ge, 2007; Cobb and Horst, 2004; Chung and Nation, 2003) subcorpus has a slightly lower AWL coverage than the Physical Sciences (Ward, 2009; Hyland and Tse, 2007), with the AWL covering 9.40% and 10.20% respectively. A salient difference between the lexical profile of LS and PS is the amount of off-list words, especially considering the 10% occurrence of off-list words predicted by Nation and Coxhead (2001). Therefore, the next sections explore the lexical profiles of these subcorpora separately.

4.2.1 Life Sciences' lexical profile

Previous studies that have investigated the AWL coverage in Life Sciences related corpora showed a somewhat lower coverage of academic words, Chung and Nation (2003), 8.60%; Cobb and Horst (2004), 6.72% and Chen and Ge (2007), 10.07%. While in Chung and Nation (2003) this is due to the amount of off-list words, which is 71.1%, in Cobb and Horst (2007) the AWL and the GSL combined cover 81.57% which falls in the coverage proposed by Coxhead and Nation

(2001). However, in the corpus of Brazilian students almost 20% of the words in LS are neither in the GSL nor in the AWL. Hence, I have decided to explore the off-list words in this subcorpus, the figure below represents the lexical profile of one text within this subcorpus. The words in red are off-list words, the words in yellow are from the AWL, and the blue and green words appear in the GSL.

Figure 3 - Excerpt from LS text.

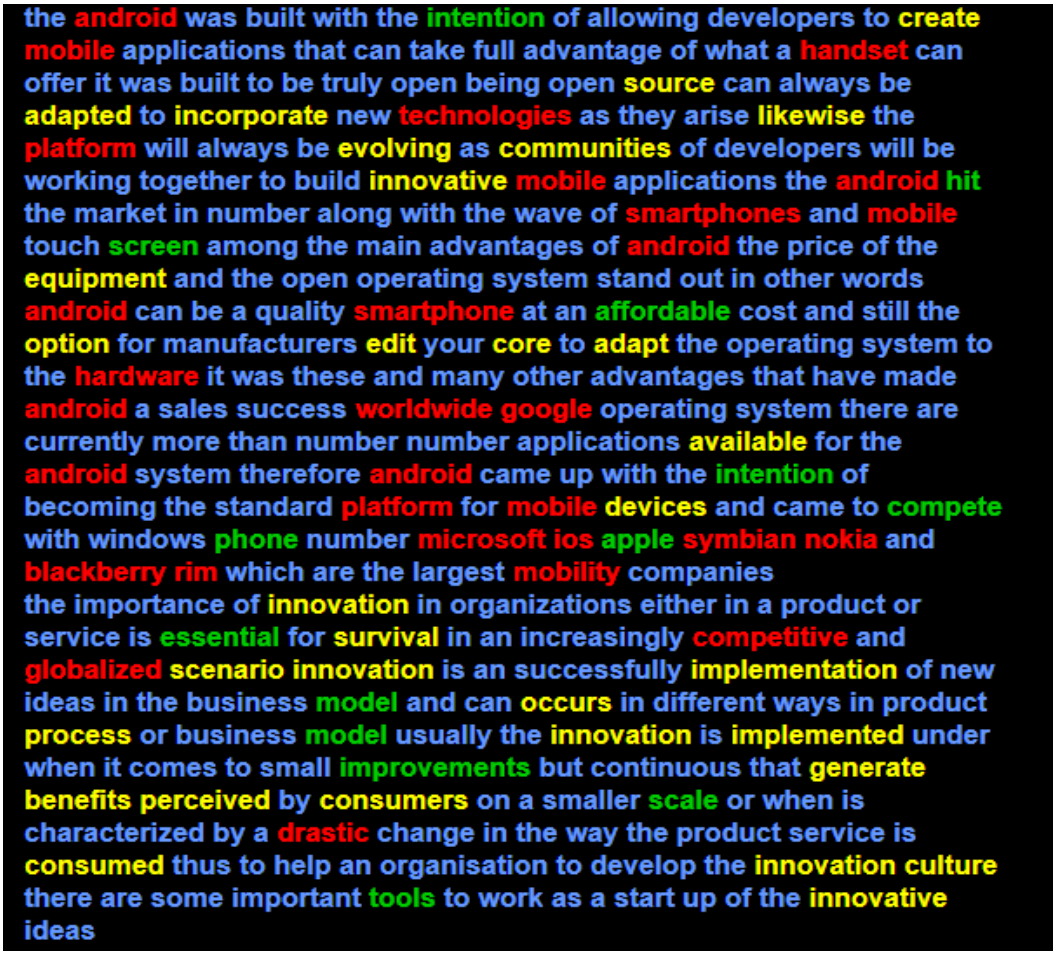


This excerpt indicates that the high percentage of off-list words is due to the use of discipline specific words (chloride, potassium, patients, etc). There are also some nonsense words, like, “kcqnumber”, or “clcnumber” these might be due to the software’s limitations. Still, this is a small excerpt from a subcorpus of 224,979 tokens. Therefore, I have looked at the list of the 50 most frequent off-list words in LS corpus (appendix 02). Based on this list, it is possible to assume that the off-list words presented in the LS corpus are mainly related to health issues. Furthermore, among the top 10 most frequent off-list words, except for the occurrences of “et al”, all the words are related specifically to health disciplines. Finally, these findings suggest that technical vocabulary plays an important role in academic writing for disciplines in the LS.

4.2.2 Physical Sciences' lexical profile

Turning to the AWL coverage in the subcorpus of PS, it is possible to compare it with the subcorpus of Engineering in Hyland and Tse, 2007 (11.10%) and Ward, 2009 (11.30%), in this case the AWL coverage in Brazilian students' assignments, 10.20%, is slightly below the figures found on previous corpus studies. Nevertheless, the use of off-list words in this corpus is lower than in LS, only 12%. Figure 4 below presents the lexical profile of one text in the PS corpus.

Figure 4 - Excerpt from PS corpus.



Analysing the 50 most frequent off-list words in this subcorpus we can observe that they are mainly related to discipline specific terms (appendix 03). With the exception of "et al" again and "graph", the top 10 most frequent words are related to the disciplines in PS. Nevertheless, when comparing the frequency of off-list words in PS and LS, it seems that knowing these words

in LS has a larger impact in writing than in PS, even though the corpus of PS is larger, the frequency of the top off-list words in LS is two-thirds the frequency in PS.

4.2.3 Discussion

Finally, with reference to the first research question, the vocabulary profile of Brazilian students writing for academic purposes is composed of 76% of high frequency words (GSL), 9.8% of Academic words (AWL) and 14.2% of off-list words. Given the results of the AWL coverage in the overall corpus, it is possible to assume that Brazilian students' written assignments do not differ in the use of the academic words from other academic corpora presented in the literature review. Notwithstanding, one important aspect to be taken into consideration is the considerable use of off-list words when compared to previous studies, especially in the subcorpus of LS, since, as previously mentioned, the AWL and the GSL usually cover around 85% to 90% of an academic corpus. This extended usage of discipline specific words in both corpora corroborates Hyland and Tse's (2007) and Durrant's (2014) argument that discipline specific terms are as important as academic words in academic English. Furthermore, exploring the list of off-list words used in the LS and PS subcorpora, we can see that only five words occur in both lists "et", "al", "Brazil", "x" and "cells", being "et al" related to the constraints of academic texts and "Brazil" a result of the students' background. This use of off-list words, along with the different lexical profiles of both subcorpora, suggests that the vocabulary used on these fields of study are appreciably different.

The aim of this section was to present the AWL coverage in the corpus of Brazilian students writing. Although the overall results suggest that the use of AWL by these students is similar to the use of academic words in previous studies, it is still necessary to analyse which academic words these students select the next section is dedicated to it.

4.3 The use of academic vocabulary: BAWE compared to Brazilian Written English

The aim of this section is to answer the last two research questions, presented in section 1.2

c) What words in the Academic Word List (AWL) do Brazilian students use?

d) How does the use of academic words differ between Brazilian students and students represented in the British Academic Written English (BAWE) corpus?

In order to do that I have generated a word list based on the words in the AWL using Sketch Engine whitelist tool. This list depicts the frequency of AWL words in BAWE and the corpus of Brazilian students. However, for the investigation presented in this section I have taken into account only the subcorpora of PS and LS from both corpora, as already mentioned, the other two subcorpora, AH and SS, are substantially smaller to allow for in depth analysis. Furthermore, the BAWE subcopora are composed of texts from students in the levels 3 or 4, as they match the same level of the texts written by Brazilian students.

The AWL has 570 word families, which are translated into 3,120 word forms including different spellings of the same word, such as “labor” and “labour” or “analyse” and “analyze”. Taking into account the complete list, texts in the BAWE subcorpus have used 999 word forms, and Brazilian students have used 998 word forms (appendix 04). Due to space constraints, I have decided to present the first 40 most frequent AWL words in both corpora and analyse the concordance lines of the top 3 frequency words depicted in table 07 below for the corpus of Brazilian students and BAWE. The decision to analyse the overall list instead of focusing only on the first sublist is justified by the fact that several high frequency academic words in both corpora are not in the first sublist (e.g. found, energy, project, design) and these would not be represented on the results. Nevertheless, I have also explored the words in the first sublist, but with a different focus, in this case I have given special attention to the word forms being used, since this aspect is not salient when analysing the complete AWL. The next sections focus on these two analyses.

4.3.1 The top 40 AWL in BAWE and Brazilian students

The table below depicts the frequency of 54 words, these are the 40 most frequent words in BAWE and in the corpus compiled for this study. Column titles, underlined and italicized words will be explained below.

Table 7 - Most frequent academic words

BAWE			Brazilian students			
Position	Word	Frequency	Position	Word	Frequency	LL
1	data	1829	3	data	707	-0.44
2	process	1529	1	process	792	-52.18

3	found	1235	5	found	538	-8.16
4	project	1054	4	project	628	-80.27
5	design	1047	8	design	449	-5.49
6	method	967	9	method	340	1.08
7	analysis	931	6	analysis	523	-52.18
8	*required*	*880*	*29*	*required*	*208*	*39.39*
9	*research*	*857*	*14*	*research*	*266*	*7.56*
10	function	853	12	function	286	2.77
11	*structure*	*816*	*28*	*structure*	*213*	*23.84*
12	energy	776	2	energy	709	-280.53
13	area	748	7	area	493	-89.64
14	*significant*	*734*	*30*	*significant*	*206*	*14.31*
15	*evidence*	*636*	*127*	*evidence*	*101*	*79.42*
16	factors	631	10	factors	328	-22.05
17	available	617	37	available	197	4.01
18	*ratio*	*609*	*67*	*ratio*	*151*	*22.52*
19	*period*	*607*	*80*	*period*	*132*	*36.1*
20	environment	602	16	environment	254	-2.41
21	range	600	20	range	244	-1.11
22	potential	599	36	potential	198	2.45
23	methods	592	21	methods	238	-0.8
24	similar	591	41	similar	190	3.53
25	*team*	*567*	*113*	*team*	*109*	*47.28*
26	*issues*	*563*	*56*	*issues*	*163*	*8.91*
27	areas	550	17	areas	251	-6.43
28	*individual*	*503*	*71*	*individual*	*144*	*8.61*
29	*normal*	*498*	*78*	*normal*	*133*	*12.94*

30	section	492	68	section	150	5.14
31	output	484	25	output	229	-8.1
32	*role*	*464*	*90*	*role*	*127*	*10.53*
33	stress	463	15	stress	260	-25.9
34	specific	462	11	specific	293	-47
35	*strategy*	*451*	*109*	*strategy*	*112*	*16.56*
36	*response*	*444*	*58*	*response*	*161*	*74.26*
37	approach	439	33	approach	201	-5.31
38	processes	438	24	processes	230	-16.37
39	physical	435	53	physical	165	-0.01
40	reaction	426	42	reaction	190	-3.83
70	concentration	340	13	concentration	281	-91.3
73	environmental	331	18	environmental	249	-65.3
67	site	354	19	site	248	-53.99
54	obtained	396	22	obtained	235	-29.57
83	equation	310	23	equation	231	-59.17
74	region	328	26	region	225	-46.14
60	maximum	377	27	maximum	220	-25.87
50	final	401	31	final	206	-12.93
58	complex	382	32	complex	206	-16.89
66	negative	359	34	negative	200	-19.21
64	technology	365	35	technology	199	-17.24
62	positive	374	38	positive	192	-12
79	achieved	316	39	achieve	192	-26.44
228	construction	146	40	construction	191	-128.01

The first column of the table represents a word position in the BAWE word list, the third column depicts its frequency in BAWE, the following column presents a word position in the

corpus of Brazilian students' word list, the sixth column shows its frequency in this corpus, and finally the last column presents the log-likelihood results.

Regarding the log-likelihood (LL) statistical test, Rayson (2002) has developed this test specifically for corpus linguistics analysts. According to this author, LL is appropriate to establish the statistical significance of the different frequencies of a word (or expression) between two corpora of different sizes. Rayson (2002:58) lists a series of advantages for using the LL test in preference to other statistical tests, among them is the fact that "LL has been shown to be better 'in general' than the chi-squared test" and that "the Mann-Whitney test is suitable only for mid to high frequency words and for comparing corpora of the same size" (p.58). Therefore, I have opted to use this test in the study presented here. Taking into account the reading of this test's results, if the LL outcome is 6.43 or more, it means that there is a 99% chance that the difference between the two corpora is not random. Hence, this test gives statistical information in order to determine if a word is significantly more frequent in a corpus in comparison to its use in another corpus. Sometimes the result of this test is negative (-), like in most results shown in table 07 above, this means that the first corpus, in this case BAWE, includes proportionally lower frequency of the word analysed than the second corpus, in this case the corpus of Brazilian academic English.

As can be seen in the table above, the amount of negative results suggest that writers in the BAWE corpus have used a wider range of academic vocabulary, while Brazilian students presented lower lexical variation. Furthermore, although the AWL does not account for parts of speech, it is possible to assume, based on table 07, that the most frequent academic words are either nouns or verbs, followed by adjectives.

4.3.1.1 High frequency words in BAWE

As can be seen in table 07 above, 20 words are used significantly more in BAWE than in the corpus of Brazilian students, 14 of these are statistically significant (LL 6.43 or more). These words are marked in italics. Taking into consideration the three most statistically underused words by Brazilian students – "evidence", "response" and "team" – and looking at their concordance lines, along with their different uses in the LS and PS subcorpora, we can consider some hypotheses for the reasons why they are underused. In the figures below, I present 10 randomly selected concordance lines for each word. It is worth mentioning that although I have taken into consideration all the concordance lines for the analysis, it would not be feasible, due to space constraints, to present all the concordance lines of each word here.

Concordance lines – evidence

Figure 5 - Evidence: BAWE

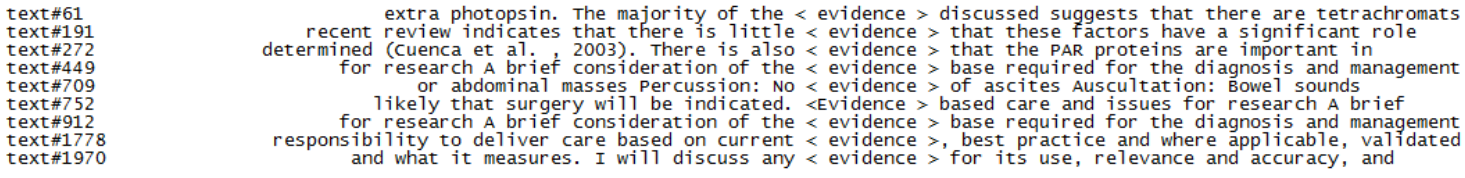
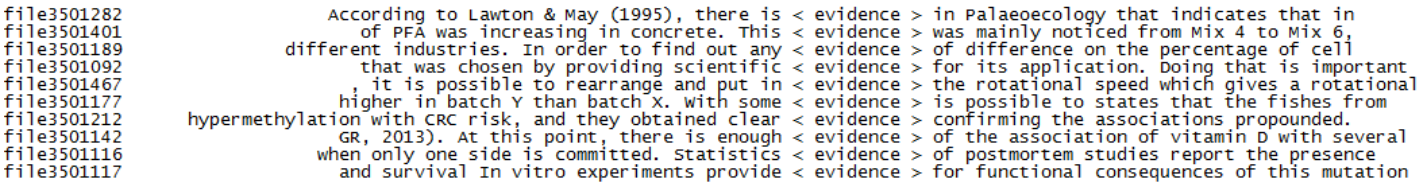


Figure 6 - Evidence: Brazilian students



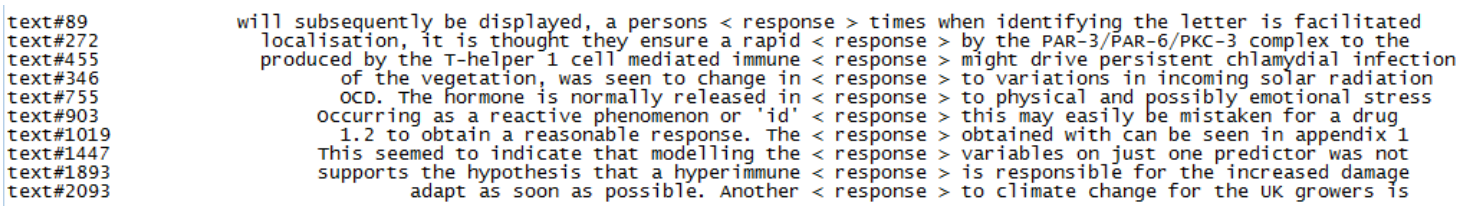
The first aspect to notice regarding the word “evidence” is that it is mainly used in the subcorpus of LS in both corpora. As can be seen from the table below the difference between LS in BAWE and the corpus of Brazilian students is greater than in PS. This might be due to the use of the expression “evidence base*” which appears 126 times in BAWE and only two times in Brazilian students writing (LL 64.89). Hence, Brazilian students’ underuse of the word “evidence” can be related to their limited use of this multi-word unit in LS

Table 8 - Evidence: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	Evidence	625	Evidence	71	122.16
PS	Evidence	135	Evidence	30	7.5

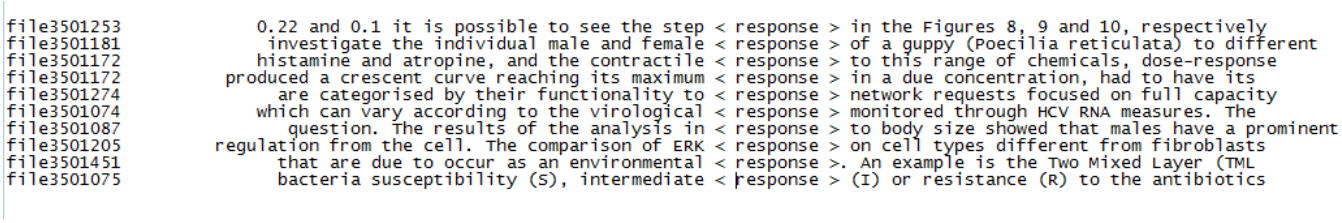
Concordance lines - response

Figure 7 - Response: BAWE



text#89	will subsequently be displayed, a persons < response > times when identifying the letter is facilitated
text#272	localisation, it is thought they ensure a rapid < response > by the PAR-3/PAR-6/PKC-3 complex to the
text#455	produced by the T-helper 1 cell mediated immune < response > might drive persistent chlamydial infection
text#346	of the vegetation, was seen to change in < response > to variations in incoming solar radiation
text#755	OCD. The hormone is normally released in < response > to physical and possibly emotional stress
text#903	Occurring as a reactive phenomenon or 'id' < response > this may easily be mistaken for a drug
text#1019	1.2 to obtain a reasonable response. The < response > obtained with can be seen in appendix 1
text#1447	This seemed to indicate that modelling the < response > variables on just one predictor was not
text#1893	supports the hypothesis that a hyperimmune < response > is responsible for the increased damage
text#2093	adapt as soon as possible. Another < response > to climate change for the UK growers is

Figure 8 - Response: Brazilian students



file3501253	0.22 and 0.1 it is possible to see the step < response > in the Figures 8, 9 and 10, respectively
file3501181	investigate the individual male and female < response > of a guppy (Poecilia reticulata) to different
file3501172	histamine and atropine, and the contractile < response > to this range of chemicals, dose-response
file3501172	produced a crescent curve reaching its maximum < response > in a due concentration, had to have its
file3501274	are categorised by their functionality to < response > network requests focused on full capacity
file3501074	which can vary according to the virological < response > monitored through HCV RNA measures. The
file3501087	question. The results of the analysis in < response > to body size showed that males have a prominent
file3501205	regulation from the cell. The comparison of ERK < response > on cell types different from fibroblasts
file3501451	that are due to occur as an environmental < response >. An example is the Two Mixed Layer (TML
file3501075	bacteria susceptibility (S), intermediate < response > (I) or resistance (R) to the antibiotics

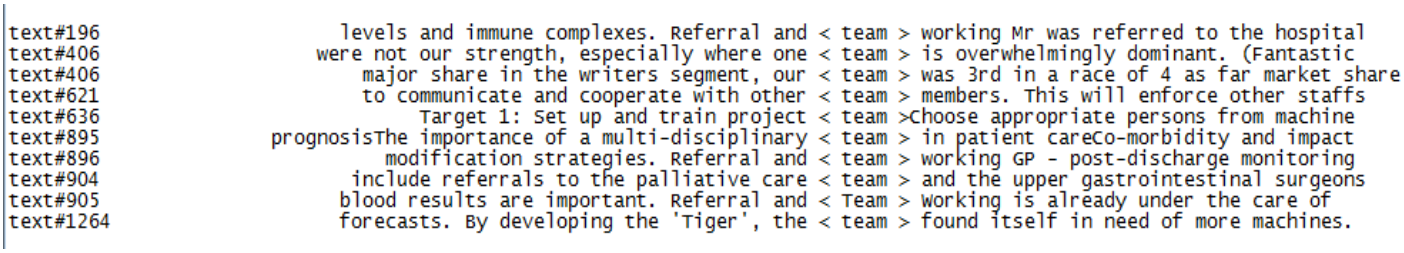
In the case of the word “response” both subcorpora present different uses of this word. It is possible to notice from the concordance lines that Brazilian students tend to use this word when related to “time + response” or “name of a procedure + response”. In addition, the use of “response” on these fixed structures is more frequent in the field of LS than in PS, which suggests that these fixed expressions are more relevant in this field of study. Furthermore, as can be seen in the table below, if we take into account only the subcorpus of LS Brazilian students actually presents an overuse the word “response”. This result signals that there is a difference in the use of academic words across different disciplines.

Table 9 - Response: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	Response	191	Response	100	-7
PS	Response	271	Response	61	13.65

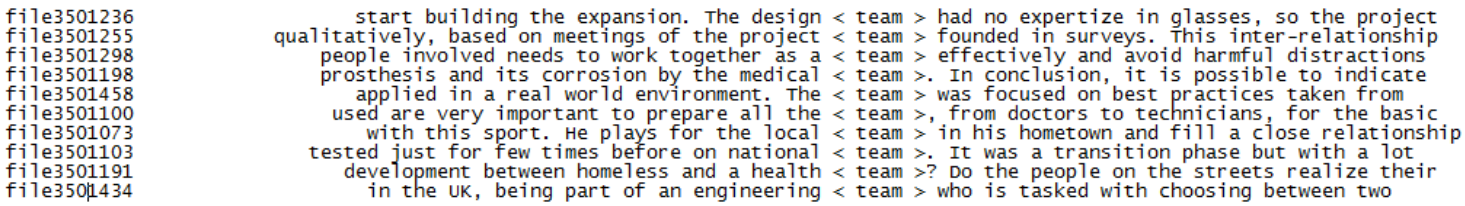
Concordance lines - team

Figure 9 - Team: BAWE



text#196	levels and immune complexes. Referral and	< team >	working Mr was referred to the hospital
text#406	were not our strength, especially where one	< team >	is overwhelmingly dominant. (Fantastic
text#406	major share in the writers segment, our	< team >	was 3rd in a race of 4 as far market share
text#621	to communicate and cooperate with other	< team >	members. This will enforce other staffs
text#636	Target 1: Set up and train project	< team >	Choose appropriate persons from machine
text#895	prognosisThe importance of a multi-disciplinary	< team >	in patient careCo-morbidity and impact
text#896	modification strategies. Referral and	< team >	working GP - post-discharge monitoring
text#904	include referrals to the palliative care	< team >	and the upper gastrointestinal surgeons
text#905	blood results are important. Referral and	< Team >	working is already under the care of
text#1264	forecasts. By developing the 'Tiger', the	< team >	found itself in need of more machines.

Figure 10 - Team: Brazilian students



file3501236	start building the expansion. The design	< team >	had no expertize in glasses, so the project
file3501255	qualitatively, based on meetings of the project	< team >	founded in surveys. This inter-relationship
file3501298	people involved needs to work together as a	< team >	effectively and avoid harmful distractions
file3501198	prosthesis and its corrosion by the medical	< team >	. In conclusion, it is possible to indicate
file3501458	applied in a real world environment. The	< team >	was focused on best practices taken from
file3501100	used are very important to prepare all the	< team >	, from doctors to technicians, for the basic
file3501073	with this sport. He plays for the local	< team >	in his hometown and fill a close relationship
file3501103	tested just for few times before on national	< team >	. It was a transition phase but with a lot
file3501191	development between homeless and a health	< team >	? Do the people on the streets realize their
file3501434	in the UK, being part of an engineering	< team >	who is tasked with choosing between two

In the case of the word “team” it is possible to assume that the underuse is related to text genre, as the occurrences of this word are more frequent in the genre of case studies in medical sciences in BAWE. The assignments prompts of these case studies usually contain one section entitled “referral and team working”, which encourage students to describe their experience working on a team. Thus, explaining the high frequency of occurrences of “team” on BAWE corpus. There are not many assignments written by Brazilian students that represent this same genre structure. Furthermore, as can be seen in the table below, there is also an accentuated difference when comparing the subcorpora of PS and LS.

Table 10 - Team: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	Team	159	Team	39	6.15
PS	Team	520	Team	87	58.49

Briefly taking into consideration the use of these three most frequent academic words in BAWE, we can observe that there is a significant difference in the use of academic vocabulary between the corpus of LS and PS, especially regarding frequency. In the next paragraphs, I will compare the use of the words that were overused in Brazilian students' assignments when compared to their frequencies in the BAWE corpus.

4.3.1.2 High frequency word in the corpus of Brazilian students

Considering the words in Table 07, there are 34 items that are proportionally more frequent in the corpus of Brazilian students than in BAWE, 25 of these words are statistically significant (LL 6,43 or more), these words are underlined in that table. Ten concordance lines of the top three of these words – "energy", "construction" and "concentration" - are presented below along with a brief discussion to the reason why they are more frequent in the corpus of Brazilian students.

Concordance lines - energy

Figure 11 - Energy: BAWE

text#118	developed than other forms of renewable	< energy >	, and is more capable than the main contenders
text#183	the tails and the desire to minimise free	< energy >	. Marrink and Mark confirm the formation
text#346	evapotranspiration in all instances. This	< energy >	to drive evapotranspiration comes from
text#352	holds, if RS remain the same. Question 2	< Energy >	Balance of a Sheltered and Exposed Crops
text#352	water, □ = 2450 J/g, this gives:	<energy >	to evaporate 1 mm depth of water over 1
text#623	using , So, . 8. Gibbs free	< energy >	determination and some comment. Gibbs
text#686	and experimenting with to find the lowest	< energy >	conformer of menthol- a relatively simple
text#687	Show your working to calculate the bond	< energy >	. Show your working to calculate the
text#691	and this is seen in a larger activation	< energy >	(compared to water solvent). The
text#754	pattern may have been causing his lack of	< energy >	and consequent anhedonia. A short-term

Figure 12 - Energy: Brazilian students

file3501224	with the possibility of using renewable	< energy >	sources. The prototypes developed at the
file3501225	station, the airflow will transfer some	< energy >	to the turbine, and this energy will be
file3501227	PV is measured in kilowatts peak (kWp) (	<Energy >	Saving Trust, 2014). Brazil is the world's
file3501228	requires 10 percent of the total amount of	< energy >	used in transport to be provided by renewables
file3501242	the use of 20% of renewable energy in EU	< energy >	consumption by 2020. It seems that wind
file3501248	generation projects as a premier choice in	< energy >	design. There are always more people waiting
file3501249	trends to raise cell viability. Also, laser	< energy >	(8J-24J) does not show to be a major issue
file3501252	order to allow this source of renewable	< energy >	. Alternatively, the recycling of CDW in
file3501239	questions based on the book "Sustainable	< Energy >	- Without the Hot Air", the first reasonably
file3501271	Building insulation is a way of improving the	< energy >	efficiency rating of buildings in a cost

Analysing the concordance lines, it is possible to claim that the main difference in the use of "energy" by Brazilian students is that these students usually use words related to sources of energy, such as "renewable", "solar" and "wind" as modifiers of the word "energy", as can be seen from the word sketch of "energy" in the corpus of Brazilian students presented on the table below.

Table 11 - Modifiers of “energy: Brazilian students

Modifiers of "energy"	
Total Freq.	231
renewable	26
solar	17
wind	30
kinetic	10
tidal	10

The use of modifiers is slightly different in BAWE, as they refer to other modifiers, such as “potential” and “interaction” which, in this context, are being used as physical terms. This can be seen on table 12 below.

Table 12 - Modifiers of “energy”: BAWE

Modifiers of "energy"	
Total Freq.	271
activation	14
kinetic	15
potential	46
interaction	9
fermi-	6

It seems that these differences might be associated to the theme developed on the assignments, especially because the corpus of Brazilian students is mainly composed of students who were SwB scholars. This might have influenced the topics developed on their assignments and, as a consequence, the use of the word “energy” in their assignments. In addition, the table below shows that the overuse of this word is associated with the subcorpus of PS, this result corroborates the hypothesis presented in the preceding section that LS and PS subcorpora use academic words in contrasting ways.

Table 13 - Energy: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	Energy	239	Energy	78	1.17
PS	Energy	684	Energy	672	-301.7

Concordance lines - construction

Figure 13 - Construction: BAWE

```
text#656      Integration, co-ordination and supervision of < construction > and operation, including contingency plans
text#657      design. Other sub-contractors included Eiffel < Construction > Metallique (responsible for Steel Construction
text#679         effort invested in aerodynamics, composite < construction > and engines it is easy to forget that tyres
text#919          development phase concentrated on design, < construction > and testing of simple devices to demonstrate
text#1132     Interest - Under GAAP, interest that finances < construction > is capitalized into the cost of assets
text#1259      Cross, AISC (The American Institute of Steel < Construction >) who states that he believes that a "plateau
text#1307          Gantt (1861-1919) during his work on the < construction > of Navy ships during world war one.
text#1340         This laboratory session was regarding the < construction >, working and use of pneumatic circuits
text#1346          rock joint was overestimated but also in < construction >. The visual estimation of the JRC is the
text#1875          its products had strong links within the < construction > industry. Apart from the partnership links
```

Figure 14 - Construction: Brazilian students

```
file3501224   distribution structure that supports the demand < construction > of new plants for electric power generation
file3501225          of more expensive materials, raising the < construction > cost. Since the low pressure turbine moves
file3501226           , aside from the employment required for < construction > and other peripheral employment. The development
file3501236      malls across Brazil, but it does not have a < construction > team, so they contracted a construction
file3501242      includes foundations, grid connection, road < construction > and wind turbines costs. The variable costs
file3501244          in wide range areas, including farming, < construction >, mining, surveying, package delivery and
file3501247        on the entire chord, but will require the < construction > of a bridge. Thus, this would the most
file3501248        to the topology. Offshore wind farms have < construction > costs that can be more than three times
file3501252     Alternatively, the recycling of CDW in roads < construction > is likely to come up with soon positive
file3501265       software. Moreover, describes its physical < construction > and measures, with their respective analysis
```

It is possible to notice, from the concordance lines above, that there is no evident difference in the use of “construction” in both corpora. Therefore, the overuse of this word by Brazilian students might be associated to the use of multi-word units, such as “construction company”, “construction sector” and “construction industry”. These three expressions appear 35 times in the corpus of Brazilian students, while they only occur three times in BAWE. This suggests that maybe Brazilian students are translating these recurrent expressions in Portuguese into English.

In addition, as with the other words, there is a significant difference between the use of this word in PS and LS, which can be observed on table 14 below.

Table 14 - Construction: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	Construction	26	construction	13	-0.69
PS	Construction	147	construction	181	-112.65

Concordance lines - concentration

Figure 15 - Concentration: BAWE

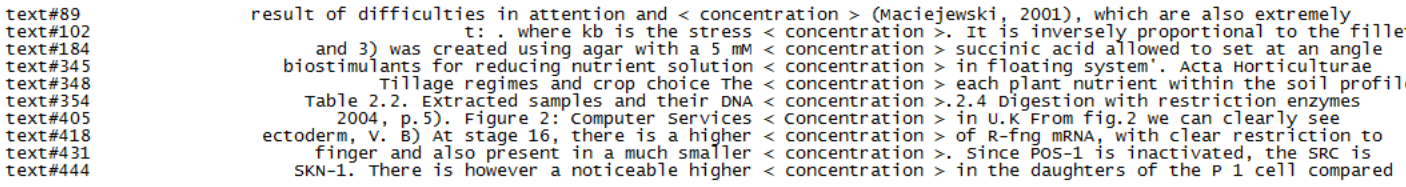
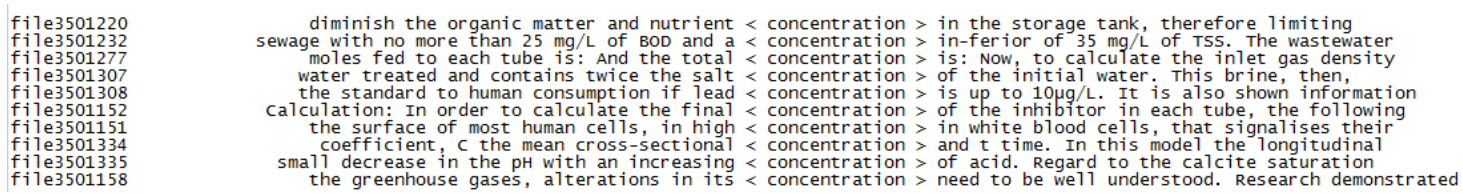


Figure 16 - Concentration: Brazilian students



The concordance lines of “concentration” are an indicative of the disciplines represented in the corpus of Brazilian writing. While BAWE is more equally balanced, the corpus compiled for this study contains more texts from disciplines such as medicine, chemistry, engineering, and physics. Therefore, the overuse of the word “concentration” can be associated with this overrepresentation of these disciplines. Furthermore, the table below shows that in the corpus of PS this word is more representative than in LS.

Table 15 - Concentration: BAWE compared to Brazilian students

	BAWE	Freq.	Brazilian students	Freq.	LL
LS	concentration	217	concentration	122	-12.22
PS	concentration	130	concentration	159	-98.12

Finally, taking into account the use of the 40 most frequent academic words in both corpora, it is possible to notice that the differences in the use of some academic words can be attributed to fixed expressions used by students in BAWE that are not present in the corpus of Brazilian students, such as "evidence base*" or expressions that Brazilian students might be transferring from Portuguese, such as "construction industry". Furthermore, the use of academic words in both subcorpora, LS and PS, is very distinct. None of the words analysed here presented roughly equal LL results in the two subcorpora when compared to BAWE.

The aim of this section was to investigate which academic words Brazilian students select in comparison to students in the BAWE corpus and analyse their concordance lines. The next section will look at the word forms used in the first sublist of the AWL.

4.3.2 Differences on the first sublist

According to Coxhead (2000) the words in the first sublist represent around 50% of the academic words in a corpus, therefore I have chosen to explore which word forms Brazilian students use (or not) when compared to students in BAWE based on this sublist. There are 60 node words with 453 word forms in this sublist; Brazilian students use 235 word forms, while 400 word forms are used in BAWE. The table in appendix 05 presents the frequency of occurrence of these words in both corpora. This table is organized alphabetically, rather than by frequency, in order to allow for comparisons related to word forms.

The AWL is organized in node words and their derived forms, these derivations are created through the addition of suffixes and prefixes, the next two sections explore these derivations. For the purpose of the discussion presented below prefixes are considered any addition before a node word. While suffixes are considered any addition after the node word as represented in the AWL.

In other words, “re” is not considered a prefix of “research” because the word “research” is a node word in the AWL.

4.3.2.1 Suffixes

When a suffix is added to a word it can change its grammatical function, *-tion* for example is used to create nouns from verbs. However, the AWL only depicts word forms, regardless of their grammatical function. In other words, some forms in the AWL use the suffix *-es/s*, for instance, which can be used to mark the third person singular in a verb or the plural of a noun. Therefore, the analysis below will not take into account grammatical functions.

As can be seen in the table in appendix 05, texts written by Brazilian students tend to show less processes of suffixation than texts in BAWE. While in BAWE 59 node words are used, in the corpus of Brazilian students 56 node words are used. Nevertheless, as previously mentioned these node words are translated into only 235 word forms in the corpus of Brazilian students, while in BAWE this represents 400 word forms. Thus, the difference in the total amount of academic words from the first sublist can be associated to the use of derivation processes. For instance, while students in BAWE use nine different inflections for the node word “concept”, as presented in the table below, Brazilian students use only three.

Table 16 - Word forms of the word “concept”

Brazilian students		**BAWE**	
concept	95	concept	1743
conception	5	conception	361
concepts	64	concepts	620
conceptual	13	conceptual	167
		conceptualisation	28
		conceptualise	12
		conceptualised	36
		conceptualises	5

		conceptualising	7
		conceptually	17

In addition, Garcia and Schwindt (2010) have studied the production of the suffixes - *ing, -ful -less, -ly,* and -*ment* by Brazilian students. In their study, -*ing* and -*ful* were more likely to be produced by Brazilian students than -*ly* and -*ment*. Considering the corpus compiled for this study, the table below depicts the occurrence of each suffix in different word forms and the total frequency of the suffix in the corpus. It is possible to notice that the suffixes -*ful* and -*less* do not occur in either corpora.

Table 17 - Suffixes in BAWE versus Brazilian students

Brazilian students			**BAWE**			
Suffix	Nº word forms	frequency	Suffix	Nº word forms	frequency	LL
-*ing*	19	479	-*ing*	43	6,004	+1,637.57
-*ly*	8	171	-*ly*	20	2,569	+801.64
-*ment*	5	229	-*ment*	8	2,709	+712.97

As can be seen in the table above, in the study presented here -*ment* is produced more frequently by Brazilian students than -*ly*. Finally, it is possible to claim that there is a considerable statistical difference among the use of these suffixes in both corpora. The results presented in this section suggest that Brazilian students do not use suffixes as frequently as it is found in BAWE. The next section investigates the use of prefixes by these students.

4.3.2.2 Prefixes

In the table below, we can see that the prefixes used by students in the BAWE corpus are: *dis-, i-, in-, mis-, over-, re-* and *un-*. Some of these occur more frequently, like *un-* and *re-*, for instance. Nevertheless, Brazilian students only selected the prefix *i-*.

Table 18 - Prefixes in BAWE x Brazilian students

BAWE		Brazilian Students		LL
dissimilar	54			+34.42
illegal	109	illegal	8	+31.88
illegality	27			+17.21
illegally	8			+5.1
inconsistencies	48			+30.59
inconsistency	65			+41.43
inconsistent	98			+62.46
insignificant	132			+84.13
invariably	34			+21.67
misinterpret	7			+4.46
misinterpretation	25			+15.93
misinterpreted	25			+15.93
overestimated	26			+16.57
reassess	13			+8.29
reassessed	8			+5.10
reassessment	16			+10.20
recreate	18			+11.47
recreated	7			+4.46
redefine	32			+20.39
redefined	25			+15.93
redefining	14			+8.92
redistribute	16			+10.20
redistributed	15			+9.56
redistributing	9			+5.74
redistribution	74			+47.16

reformulated	7			+4.46
reformulation	5			+3.19
reinterpret	8			+5.10
reinterpretation	9			+5.74
reoccurring	5			+3.19
restructure	24			+15.30
restructured	5			+3.19
restructuring	82			+52.26
unavailable	46			+29.32
unconstitutional	11			+7.01
undefined	13			+8.29
underestimate	22			+14.02
underestimated	67			+42.70
underestimates	12			+7.65
underestimating	5			+3.19
unidentifiable	7			+4.46
unresponsive	11			+7.01
unstructured	27			+17.21

As it is shown in the table above, students in BAWE have used 43 word forms with prefixes, while Brazilian students only use one word form with a prefix. Given the analysis presented in this section, it is possible to notice that Brazilian students are underusing suffixes and prefixes in comparison to students in BAWE. Furthermore, this results question Bauer and Nation (1993:253) claim that "base word and all its derived and inflected forms can be understood by a learner without having to learn each form separately", since Brazilian students use the node word, which shows that they understand its uses, but they do not use most of the derived forms. We can also hypothesize that prefixes like *un-*, *under-* and *over-*, which are not derived from Latin might not be transparent for Brazilian students. However, further studies would be necessary to confirm this hypothesis.

Resuming the last research question, after the analysis of the word forms in the first AWL sublist, we can see that the main difference in the use of academic words between Brazilian students and students represented in BAWE is the fact that the first ones do not modify the words using processes of affixation. Therefore, Brazilian students present a lower lexical variety than students represented in BAWE.

The aim of this section was to analyse the word forms privileged by Brazilian students in the first word list as this can contribute to the understanding of which aspects of academic vocabulary could be addressed in EAP classes for Brazilian students. The next section presents conclusion and implications of this study.

5. Conclusion

5.1 Introduction

In this section, I conclude this dissertation by summarizing the findings of this study and discussing some limitations of the investigation conducted here. In addition, I present this study's contributions and pedagogical implications to the teaching of EAP, especially in the context of the programme Languages without Borders in Brazil. Finally, I suggest future research that could be developed on the field EAP written by Brazilian students.

5.2 Summary of key findings

The aim of this research was to investigate the use of academic vocabulary by Brazilian students. More specifically, this research has explored the AWL coverage in a corpus of Brazilian students and compared how Brazilian students use the words in the AWL with how students represented in the BAWE corpus use these words. For the purposes of this study, I have conducted a literature review of previous investigations that examined the AWL coverage in academic corpora, I have also compiled a corpus of Brazilian student's written assignments, with 380 texts from four fields of expertise, and explored the use of academic words using Range and Sketch Engine.

The analysis of the AWL coverage showed that Brazilian students use academic words to the same extent as other academic corpora presented in the literature review. It also revealed a significant difference between the lexical profile of Life Sciences and Physical Sciences, especially if we take into account the use of off-list words. Life Sciences texts relied more on discipline specific words, corroborating Hyland and Tse's (2007) view that the division between academic and discipline specific vocabulary is not clear.

The investigation of the AWL words selected by Brazilian students and students in BAWE suggested that fixed expressions influence the underuse or overuse of academic words by Brazilian students, in some cases these students transfer expressions from Portuguese into English or do not use common expression in their field of study in their texts. Furthermore, the comparative analysis of the top frequency AWL words between the corpus of LS and PS suggested that these two areas of study use these academic words in distinct ways.

Finally, the study of the word forms selected by Brazilian students has showed that these students are not using as many process of affixation as students in BAWE. Taking into account

prefixes, this study showed that Brazilian students scarcely use them in their academic writing. In addition, the results of this section shows that even though Brazilian students use roughly the same amount of node words as students represented in BAWE they tend to repeat the same word form.

5.3 Research evaluation and limitations

The main limitation of this study is the size and representativeness of the corpus used in this investigation. Although tests of statistical significance were possible in this investigation, the corpus was not balanced. Therefore, the disciplines of Physical Sciences and Life Sciences were overrepresented. Furthermore, the corpus compiled for this study could be expanded to contain more assignments, however, due to time constraints, this was not possible.

Time and space constraints also hindered further analysis of all the AWL sublists. The investigation of the word forms in all sublists could have exhibited other language aspects that Brazilian students use differently than students in BAWE. In addition, the first part of the research could have been sounder if the AWL coverage of BAWE was verified as well, since the comparison of the AWL coverage would, then, be between two similar corpora.

In addition, in section 4.3.1 the presentation of only 10 concordance lines might be seen as a limitation, however this selection was necessary due to space constraints. Finally, I consider the use of the AWL as one of the limitations of this study. Although this academic word list is the most widely used in other academic research, I believe that the results of the last section might have been more instructive if the word list used as a reference was based on lemmas, rather than word forms. A word list based on lemmas could provide some insights related to the use of academic words in each part of speech, which is not possible with the AWL.

5.4 Pedagogical implications

The research findings suggest that Brazilian students use academic vocabulary to the same extent as other academic corpora, however, it was possible to notice some peculiarities in the use of academic vocabulary by Brazilian students. Considering these peculiarities, some pedagogical recommendations for EAP teaching for Brazilian students are presented below. Nevertheless, the recommendations given here take into consideration mainly the Languages without Borders programme.

As a first point, the results indicate that there is a considerable difference in the lexical profile of texts in the fields of Life Science and Physical Sciences. This difference is reflected not

only on the percentage of academic words used, but also on how these words are used in these fields of study. This suggests that LS and PS students would benefit more from their EAP classes if they took separate lessons as it is argued by Hyland and Tsé (2007). Considering the context of the programme Languages without Borders, this could be implemented by dividing students according to their field of study, rather than enrolling students based only in their level of English.

In addition, in classes with mixed disciplines, teachers could group students from the same field of study to complete the same activity, this way classroom practices could be more focused on the academic vocabulary of those specific disciplines. Furthermore, teachers could ask students to bring examples of texts of their field of study to the classroom and based on a corpus analysis point out to the students the different lexical choices in these texts. The main point is that teachers need to be aware that there is a considerable difference in the use of academic vocabulary between LS and PS and this should be taking into consideration when preparing their lessons.

Furthermore, the results on the second part of this study suggest two topics that should be addressed in EAP classes for Brazilian students. The first one is the use of discipline specific fixed expressions. This could be done by exploring with students the lexical bundles in BAWE for the student's discipline, or by creating pedagogical materials based on Hyland (2008) or Simpson-Vlach and Ellis (2010) lists of fixed expressions recurrent in academic texts.

The second topic that should be emphasized in EAP lessons is affixes, as it is clear from the results that Brazilian students are not using prefixes and suffixes to form new word forms as productively as their BAWE counterparts. EAP teachers can practice in class the functions and uses of affixes in academic texts by bringing examples from the BAWE corpus and asking students to compare with their own texts, for instance.

The final pedagogical contribution is the corpus compiled for this study. This corpus can be used by teacher to show examples of Brazilian academic writing to students and it can also be used by researchers to explore other linguistic aspects of Brazilian written EAP.

5.5 Further research suggestions

In addition to the pedagogical implications presented above, I would like to suggest some ideas for future research exploring Brazilian EAP.

The first one would be to compare the coverage of academic vocabulary in BAWE and the corpus of Brazilian students. This would allow the researcher to draw more informed conclusions related to the coverage of academic vocabulary in Brazilian students' writings. The study presented in this dissertation compared the academic vocabulary coverage with professional academic texts,

therefore further study comparing Brazilian students written EAP with BAWE students written EAP could complement this research. Furthermore, future studies should rely on the AVL, since this list is organized based by lemmas it, then, would allow for the analysis of academic vocabulary according to their parts of speech.

A second aspect to be explored in the future is the differences in the use of Academic Vocabulary across genres. From table 06 presented in the results section, we could notice that some genres, such as Proposals and Critique, present higher AWL coverage than others, such as Literature Survey, Explanation and Methodology Recount. However, the study presented in this dissertation did not explore genre differences in the use of the AWL.

In addition, another topic that could be addressed is the lexical bundles in Brazilian students' academic writing. Many researchers (Simpson-Vlach and Ellis, 2010; Hyland, 2008; Biber, Conrad and Cortes, 2004) claim that lexical bundles are the fabric of language and some of them have explored the use of lexical bundles in academic texts. Thus, it might be interesting to investigate the similarities and differences in the use of these units of language by Brazilian students.

References

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

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6, 253–279.

Bauer, L. (1993). *Manual of information to accompany the Wellington Corpus of Written New Zealand English.* Wellington, New Zealand: Victoria University of Wellington.

Baumann, J. & Graves, M. (2010). What is academic vocabulary?. J*ournal of Adolescent and Adult Literacy*, 54 (1), 4 - 12.

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction.* New York: Guilford.

Bennett, P. & Stoeckel, T. (2015). The new academic word list test (NAWLT). Retrieved from http://www.newgeneralservicelist.org/ngsl-levels-test/

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

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...:Lexical bundles in university teaching and textbooks. *Applied Linguistics,* 25 (3), 371–405.

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

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

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36, 1–22.

Browne, C., Culligan, B. & Phillips, J. (2014). *A New Academic Word List.* Retrieved from http://www.newacademicwordlist.org

Browne, C., Culligan, B. & Phillips, J. (2914). *A New General Service List*. Retrieved from http://www.newgeneralservicelist.org/

Campion, M. & Elley, W. (1971). *An Academic Word List*. Wellington New Zealand Council for Educational Research

Carter, R. (2012). *Vocabulary: Applied Linguistic Perspectives*. London: Routledge.

Chen, Q. & Ge, C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes*, 26, 502–514.

Chung, T. & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a foreign language*, 15(2), 103 - 116.

Chung, T. (2009). The newspaper word list: A specialised vocabulary for reading newspapers. *JALT Journal*, 31(2), 159–182.

Clark, M & Ishida, S. (2005). Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for academic purposes*, 4, 225 - 238.

Cobb, T. (n.d.). *The compleat lexical tutor.* Retrieved from http://www.lextutor.ca/ on 24 March 2016.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Amsterdam, the Netherlands: John Benjamins.

Cobb, T. & Horst, M. (2015). Learner Corpora and Lexis. In. S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Connor, U., & Upton, T. (2004). Introduction. In U. Connor & T. Upton (Eds.), *Discourse in the professions*: *Perspectives from corpus linguistics* (pp.1-8). Amsterdam: John Benjamins Publishing Company.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22,* 75-95.

Corson, D. (1997). The learning and use of academic English words. *Language Learning*, 47(4), 671–718.

Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, 8(4), 389-400

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34 (2), 213–238.

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252–267). Cambridge: Cambridge University Press.

Coxhead, A. & Hirsh, D. (2007). A pilot science word list for EAP. *Revue Francaise de linguistique appliquee*, XII (2) 65–78.

Coxhead, A. (2011). The Academic Word List 10 years on: research and teaching implications. *TESOL Quarterly*, 4(2), 355 - 362.

Dang, T. & Webb, S. (2014). The lexical profile of academic spoken English. *English for specific purposes*, 33, 66 - 76.

Dayrell, C. & Aluisio, S. (2008) *Using a comparable corpus to investigate lexical patterning in English abstracts written by non-native speakers*. In: Proceedings of the LREC 2008 Workshop Building and Using Comparable Corpora, Marrakech.

Durrant, P. (2014). Discipline and Level Specificity in University Students' Written Vocabulary. *Applied Linguistics*, 35(3), 328 - 356.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing?. *English for Specific Purposes,* 43, 49–61

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. J*ournal of Second Language Writing,* 4(2), 139 – 155.

Engels, L.(1968). The fallacy of word-counts. *International Review of Applied Linguistics in Language Teaching*, 6(3), 213–231.

Farid, A. (1985). *A vocabulary workbook: Prefixes, roots, and suffixes for ESL students.* Englewood Cliffs, NJ: Prentice Hall.

Farrell, P. (1990). Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary. *CLCS Occasional Paper* No. 25 Trinity College.

Fisher, D., & Frey, N. (2008). *Word wise and content rich: five essential steps to teaching academic vocabulary.* Portsmouth, NH: Heinemann.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System,* 21, 231 - 244.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. Upton (Eds.), *Discourse in the professions*: *Perspectives from corpus linguistics* (pp. 11-36). Amsterdam: John Benjamins Publishing Company.

Francis, W. & Kucera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Freitas, A. (2016) Proficiência escrita em Inglês especializado: estudo de corpus de abstracts em Medicina, Nutrição e Farmácia. (PhD Theses) Universidade Federal do Rio Grande do Sul.

Garcia, G. (1991). Factors influencing the English reading text performance of Spanish-speaking Hispanic children. *Reading Research Quarterly,* 26(4), 371–392.

Garcia, G. & Schwindt, L. (2010). Afixos em L2: um estudo preliminar sobre a aquisição de sufixos em línguainglesapor falantes de portuguêsbrasileiro. Paper presented at IX Encontro do CELSUL: Universidade do Sul de Santa Catarina.

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

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

Gass, S. & Selinker, L. (2008) *Second language acquisition: an introductory course.* New York: Routledge/Taylor and Francis Group.

Gavioli, L. (2002). Some thoughts on the problem of representing ESP through small corpora. In B. Ketteman & G. Marko (Eds), *Language and Computers: Studies in Practical Linguistics*(pp. 293–303). Amsterdam: Rodopi.

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

Ghadessy, P. (1979). Frequency counts, words lists, and materials preparation: A new approach. *English Teaching Forum*, 17, 24–27.

Granger, S. & Paquot, M. (2010) *The Louvain EAP Dictionary (LEAD).* In Proceedings of the XIV EURALEX International Congress, Leeuwarden, The Netherlands, 6-10 July 2010, 321-326.

Harmon, J.M., Wood, K.D., & Hedrick, W.B. (2008). Vocabulary instruction in middle and secondary content classrooms: Understandings and direction from research. In A.E. Farstrup & S.J. Samuels (Eds.), *What research has to say about vocabulary instruction* (pp. 150–181). Newark, DE: International Reading Association.

Hirsh, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a Foreign Language*, 8(2), 689 - 696.

Hoey, M. (2007). Lexical priming and literary creativity. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 7-30). London: Continuum.

Hsue-Cho, M. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, 13(1), 403 - 430.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Huntley, H. (2006). *Essential academic vocabulary.* Cengage Learning.

Hutchinson, T. & Waters, A. (1987). *English for specific purposes: A learning-centered approach.* Cambridge: Cambridge University Press.

Hyland, K. & Tse, P. (2007). Is there "an academic vocabulary"?. *TESOL Quarterly*, 41 (2), 235 - 253.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4–21

Johansson, S. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Oslo, Norway: University of Oslo, Department of English.

Kaur, J. & Hegelheimer, V. (2007). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, 18 (4), 287 - 310.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *Estudios de Linguistica Inglesa Aplicada*, 7, 79–102.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In. C. Lauren & M. Nordman (Eds.), *Special Language: from human thinking to thinking machines.* (pp. 316 - 323). Clevedon (England): Multilingual Matters Ltd.

Lei, L. & Liu, D. (2016). A new medical academic word list: a corpus based study with enhanced methodology. *Journal of English for academic purposes*, 22, 42 - 53.

Leki, I. & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly*, 31(1), 39 - 69.

Li, E. & Pemberton, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary. In *Joint seminar on corpus linguistics and lexicography* (pp. 183 - 196). Hong Kong: HKUST Language Center.

Li, Y. & Qian, D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System,* 38, 402–411.

Lynn, R. (1973). Preparing word lists: a suggested method. *RELC Journal,* 4(1), 25–32.

Martınez, I., Beck, S. & Panza, C. (2009). Academic vocabulary in agriculture research articles. *English for Specific Purposes*, 28, 183–198

McCarthy, M., O'Keeffe, A. & Walsh, S. (2009). *Vocabulary Matrix: understanding, learning and teaching*. Andover: Heinle, Cengage Learning.

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

Mifflin, H. (2006). *English for academic success series.* College Reading.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* New York, NY: Routledge.

Nagy, W. & Townsend, D. (2012). Words as tools: learning academic vocabulary as language acquisition. *Reading research quarterly,* 47 (1), 91 - 108.

Nation, P. (1983). Testing and learning vocabulary. *Guidelines,* 5(1), 12–25.

Nation, P.& Heatley, A. (2002). *Range: A program for the analysis of vocabulary in texts*. Retrieved from <http://www.victoria.ac.nz/lals/resources/range>.

Nation, P. (2013). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nesi, H. (2008). *Corpora and EAP*. In: LSP: Interfacing Language with other Realms: Proceedings of the 6th Languages for Specific Purposes International Seminar, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

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

Richards, J. (1974). Word lists: Problems and prospects. *RELC Journal*, 5(2), 69–84.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, & S. Gozdz-Roszkowski (Eds.), *Corpora and ICT in language studies.* (pp. 127 - 140). Frankfurt am Main, Germany: Peter Lang.

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. London & New-York: Continuum.

Park, J. (2012). *Needs analysis of ELI 721* (Undergraduate). University of Hawaii at Manoa.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22(1), 69 - 90.

Sardinha, A. (2000). Linguistica de corpus: historico e problematica. *D.E.L.T.A.*, 16 (2), 323 – 367.

Sarmento, S. & Kirsch, W. (2015). Inglês sem fronteiras: uma mirada ao contexto de prática pelo prisma da formação de professores. *Ilha do Desterro*, 68(1), 47 - 59.

Sarmento, S., Scortegagna, B. & Silva, L. (2014). *O corpus LILE: abstracts das áreas de linguística e literature*. In: Proceedings of the I International Seminar on Language Acquisition, Pontífice Universidade Católica do Rio Grande do Sul Porto Alegre.

Savage, A. & Mackey, D. (2010). *Read This!*. Cambridge: Cambridge University Press.

Schmitt, N. Schmitt, D. and Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18 (1), 55 - 88.

Schmitt, D. & Schmitt, N. (2011). *Focus on Vocabulary: mastering the academic word list.* (2nd Ed). New York: Pearson Education.

Scott, M. & Johns, T. (1993). *MicroConcord*. Oxford: Oxford English Software.

Shaw, P. (1991). Science research students' composing processes. *English for specific purposes*, 10, 189 - 206.

Simpson-Vlach, R. & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4), 487 - 512.

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

Snow, C.E., Lawrence, J. & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness*, 2(4), 325–344.

Snow, C. (2010). Academic Language and the Challenge of Reading for Learning About Science. *Science, New Series*, 328 (5977), 450 - 452.

Stahl, S & Fairbanks, M. (1986). The Effects of Vocabulary Instruction: A Model-Based Meta-Analysis. *Review of Educational Research*, 56(1), 72-110.

Stubbs, M. (2007). On texts, corpora and models of language. In M. Hoey, M, Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, discourse, and corpora*: *Theory and analysis* (pp. 127-162). London: Continuum.

Thorndike, E. & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College Press.

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

Townsend, D., Filippini, A., Collins, P. & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal*, 112 (3), 497 - 518.

Townsend, D. & Kiernan, D. (2015). Selecting academic vocabulary words worth teaching. *The reading teacher*, 69(1), 113 - 118.

Trimble, L. (1985). *English for science and technology. a discourse approach*: Cambridge: Cambridge University Press.

Valcourt, G. & Wells, L. (1999). *Mastery: A University Word List reader*. Ann Arbor: The University of Michigan Press.

Vongpumivitch, V., Huang, J. Y., & Chang, Y. C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28, 33 - 41.

Wang, J., Liang, S. & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27, 442–458

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28, 170–182.

West, M. (1953). *A general service list of English words.* London: Longman.

Xue, G. & Nation, P. (1984). A university word list. Language Learning and Communication, 3, 215–229.

Yang, M. (2015). A nursing academic word list. *English for specific purposes*, 37, 27 - 38.

Yorkey, R. (1981). *Checklists for vocabulary study.* New York: Longman.

Zimmerman, C., Burgemeier, A., Zwier, L., Robin, B. & Richmond, K. (2012). *Inside Reading: the academic word list in context* (2nd Ed). New York: Oxford University Press.

Appendices

Appendix 1 - Email sent to Brazilian Students

Bom dia

Eu estou fazendo meu mestrado em Linguística Aplicada na Universidade de Warwick na Inglaterra. Eu sou professora do Inglês sem Fronteiras na UFRGS e minha pesquisa aqui é sobre Inglês acadêmico escrito por Brasileiros, o objetivo dessa pesquisa é analisar textos escritos por brasileiros para poder, no futuro, melhor preparar outros Brasileiros que vão entrar em universidades no Reino Unido.

Para continuar a pesquisa, eu preciso da ajuda de Brasileiros que estejam ou estiveram em alguma universidade do Reino Unido. Se você quiser participar desse estudo você só precisa enviar qualquer texto que você escreveu em inglês e submeteu para avaliação na faculdade, como essays, assignments, lab reports, research proposals, etc de preferência em documento DOC ou DOCXaté o dia 10/06/2016. Todos os textos serão analisados em conjunto e a sua participação será anônima.

Abaixo vocês encontram o formulário de consentimento em Inglês qualquer dúvida, por favor, mande e-mail para l.goulart-da-silva@warwick.ac.uk

Research information and consent form

Title of Research: Analysing written productions of Brazilians studying in UK universities.

Background

As a Brazilian studying in an undergraduate, masters or PhD program in the UK, you are invited to participate in my MA dissertation research. My MA is on English Language Teaching at the University of Warwick and my research goal is to study academic English in written texts produced by Brazilian students as part of their degrees at their universities in the UK.

What your participation would involve

If you choose to participate in this research, your collaboration would involve sending me any piece of written work that you have submitted for evaluation to your university in the UK. These might be different kinds of text, for example, lab reports, essays, reviews etc.

These texts will be used to analyse language features. Thus, improving the understanding of how Brazilians use English in academic contexts and, in the future, helping English teachers to prepare students for their future life in a University in the UK.

Risks and benefits to you

There are no risks to you in participating in this research; all the texts will be processed anonymously. Therefore, even if there are issues, such as plagiarisms, misspellings or inappropriate use of language, these will not be referred back to you. I believe that participating in this research will benefit you as the results can help you develop your academic English skills.

Feedback to you

At the completion of the study, the research report will be available for those interested in improving their academic English skills.

Implications of giving or withholding consent

Your decision on whether or not to participate in this project will be kept confidential and will in no way affect any other aspect of your relationship with your university in the UK.

Further information

If you would like any further information to help you to decide whether or not to participate in this project, please contact me by e-mail: l.goulart-da-silva@warwick.ac.uk

Appendix 2 - Top 50 off-list words in Life Sciences

AL	887
ET	875
PATIENTS	355
CELLS	291
DNA	230
CANCER	225
CELL	195
SPECIES	170
ML	166
GENES	152
DRUG	126
PROTEIN	126
DRUGS	124
BACTERIA	118
PROTEINS	112
GENE	106
CLINICAL	102
INFECTION	100
TISSUE	99
CHRONIC	98
SYMPTOMS	93
DIAGNOSIS	87
BIOLOGICAL	84
GENETIC	84
ACTIVATION	81

BRAZIL	81
CLIMATE	81
MALARIA	74
CALCIUM	71
REFERENCE	69
HCV	68
PCR	67
MEMBRANE	65
MUTATIONS	65
TISSUES	65
LIVER	64
MUSCLE	64
X	64
DIET	61
CARBON	59
MBC	59
VITAMIN	56
MOLECULES	55
MARINE	54
ORGANISMS	54
VIRUSES	54
DIABETES	53
PATHWAY	53
CORRELATION	52
THERAPY	52

Appendix 3 - Top 50 off-list words in Physical Sciences

AL	388
ET	386
TURBINE	216
URBAN	171
MM	147
CARBON	138
SOFTWARE	126
SOLAR	126
GRAPH	124
LASER	111
CIRCUIT	110
BRAZIL	107
VOLTAGE	107
SPECIES	106
X	103
CONCRETE	101
FIG	96
FLUID	95
STORAGE	95
ARCHITECTURE	92
UNILEVER	92
UK	91
PLOT	87
BIOMASS	86

COEFFICIENT	84
ORGANIC	83
RENEWABLE	83
TORQUE	82
LONDON	81
PEAK	80
DIAGRAM	79
HUGE	77
GENERATOR	76
COMPETITIVE	75
DIAMETER	75
COLUMN	74
DENSITY	74
AXIS	73
MATLAB	72
CLIMATE	71
PITCH	71
TECHNOLOGIES	71
CO2	70
ROTOR	69
TRAFFIC	69
PH	68
VELOCITY	68
CELLS	67
LINEAR	67
HEIGHT	65

Appendix 4 - AWL words in BAWE and the corpus of Brazilian students

Brazilian students		BAWE	
process	792	data	1829
energy	709	process	1529
data	707	found	1235
project	628	project	1054
found	538	design	1047
analysis	523	method	967
area	493	analysis	931
design	449	required	880
method	340	research	857
factors	328	function	853
specific	293	structure	816
function	286	energy	776
concentration	281	area	748
research	266	significant	734
stress	260	evidence	636
environment	254	factors	631
areas	251	available	617
environmental	249	ratio	609
site	248	period	607
range	244	environment	602
methods	238	range	600
obtained	235	potential	599
equation	231	methods	592
processes	230	similar	591

output	229	team	567
region	225	issues	563
maximum	220	areas	550
structure	213	individual	503
required	208	normal	498
significant	206	section	492
final	206	output	484
complex	206	role	464
approach	201	stress	463
negative	200	specific	462
technology	199	strategy	451
potential	198	response	444
available	197	approach	439
positive	192	processes	438
achieve	192	physical	435
construction	191	reaction	426
similar	190	appropriate	426
reaction	190	financial	423
constant	187	series	420
defined	186	involved	418
analysed	186	phase	415
aspects	184	major	412
factor	182	factor	408
transfer	180	previous	404
phase	171	impact	401
elements	169	final	401
previous	168	source	400

identify	168	distribution	400
physical	165	theory	398
major	164	obtained	396
distribution	163	constant	384
issues	162	occur	383
created	162	resources	382
response	161	complex	382
error	161	layer	378
volume	160	maximum	377
create	160	defined	375
analyse	158	positive	374
initial	156	error	366
located	155	technology	365
impact	154	input	364
access	153	negative	359
ratio	151	site	354
section	150	techniques	352
implementation	148	relevant	348
parameters	145	concentration	340
individual	145	approximately	338
capacity	139	ensure	337
technique	138	environmental	331
source	137	region	328
input	137	initial	326
benefits	135	code	326
variable	134	image	323
normal	133	element	318

involved	133	achieved	316
period	132	capacity	315
affect	131	components	312
resources	130	elements	311
focus	130	equation	310
contact	130	target	297
components	130	contact	297
techniques	128	affect	295
equipment	128	computer	294
code	128	medical	290
scenario	127	features	288
role	127	communication	286
economic	127	require	283
simulation	126	overall	281
sources	124	access	280
issue	124	requirements	278
regions	123	occurs	275
chemical	123	variables	273
occurs	122	chemical	273
variation	119	parameters	272
transport	117	achieve	271
procedure	117	hence	270
transmission	116	significantly	268
task	116	indicates	268
identified	116	technique	264
interaction	115	benefits	262
achieved	115	affected	262

features	114	identified	258
relevant	113	culture	257
cycle	113	created	257
strategy	112	consumers	257
occur	112	functions	256
image	112	income	250
designed	110	requires	249
team	109	policy	249
primary	109	accurate	249
theory	107	volume	246
reactor	107	via	246
mechanism	107	stable	244
device	107	indicate	243
projects	106	individuals	237
objective	105	file	236
target	104	transfer	235
previously	104	network	235
generated	104	create	235
affected	103	aspects	235
generate	102	variable	229
cultural	102	task	229
evidence	101	annual	229
evaluate	101	shift	228
approximately	101	processing	226
requirements	100	identify	226
global	100	issue	224
medical	99	involves	221

consists	99	implementation	221
conducted	99	designed	221
goal	98	displayed	220
consumption	98	images	219
alternative	98	component	218
layer	95	estimated	217
culture	95	investment	214
concept	95	structures	208
equations	94	primary	208
aspect	94	assessment	208
community	93	indicated	206
indicate	92	random	201
network	91	interaction	201
impacts	91	errors	199
feature	91	alternative	199
element	91	focus	198
sector	90	projects	197
requires	90	maintain	197
obtain	90	sites	195
ensure	90	internal	195
communication	90	external	195
percentage	89	removed	194
conversion	88	previously	194
accurate	88	minimum	194
infrastructure	87	prior	193
consequence	87	whereas	192
sequence	86	procedure	191

release	86	established	191
interactions	86	variation	190
implemented	86	tasks	190
functions	86	interactions	190
authors	86	channel	188
maintain	85	investigation	184
location	85	entity	184
devices	85	benefit	182
hypothesis	84	occurred	181
researchers	83	majority	181
methodology	82	community	181
individuals	82	corresponding	180
appropriate	82	economic	178
accuracy	82	parallel	177
variables	81	sources	176
transition	81	psychological	176
minimum	80	generated	176
consequences	80	evaluation	176
component	80	equipment	174
tasks	79	trend	173
sustainable	79	resource	173
reliable	79	global	173
exposure	79	scheme	172
demonstrated	79	selected	171
compounds	79	consumption	171
cited	79	sector	170
guarantee	78	cycle	170

internal	77	simulation	169
extraction	77	ratios	168
context	77	predicted	168
vary	76	define	167
significantly	76	option	166
overall	76	intensity	166
intervention	76	display	166
estimated	76	sufficient	165
intensity	75	status	165
file	75	principle	165
external	75	designs	165
creating	75	accessed	163
consequently	74	framework	161
stable	73	mental	160
option	72	schemes	158
instance	72	percentage	158
indicates	72	mechanism	158
dimensions	72	compounds	155
vehicles	71	investigated	154
reactions	71	investigate	154
panels	71	consumer	154
mental	71	sequence	153
maintenance	71	purchase	153
define	71	strategies	152
strategies	70	detected	152
challenges	70	concept	152
adequate	70	structural	151

professionals	69	edition	151
display	69	mode	150
capable	69	items	150
stability	68	traditional	149
displacement	68	obtain	149
mechanisms	67	implemented	149
identification	67	consistent	149
goals	67	location	148
images	66	resolution	147
fundamental	66	monitor	147
exposed	66	initially	147
established	66	mechanisms	146
core	66	intervention	146
generation	65	definition	146
evaluation	65	construction	146
sites	64	injury	144
policy	64	generate	144
normally	64	formula	144
monitoring	64	varying	143
equivalent	64	security	143
concepts	64	assumed	143
affects	64	accuracy	143
vehicle	63	beneficial	142
researches	63	technical	141
principles	63	generation	140
evaluated	63	conducted	140
criteria	63	attributes	140

promote	62	aspect	140
indicated	62	indicating	138
detection	62	context	138
crucial	62	regions	137
benefit	62	compound	137
assess	62	guidelines	136
creation	61	enable	135
processing	60	domain	135
investment	60	approaches	134
whereas	59	selection	133
security	59	reliable	133
predict	59	monitoring	133
errors	59	assess	133
selection	58	summary	132
options	58	instance	132
focused	58	demonstrated	132
facilities	58	purchasing	131
theoretical	57	link	131
majority	57	document	131
assessment	57	distributed	131
structures	56	normally	130
considerable	56	principles	129
computer	56	aware	129
finally	55	requirement	128
approaches	55	equations	128
sum	54	core	128
removed	54	stresses	127

involves	54	equivalent	127
definition	54	objective	126
contribute	54	coding	126
selected	53	affects	126
participants	53	participants	125
investigation	53	involve	125
conclusion	53	integral	125
topic	52	obvious	124
statistical	52	consists	124
require	52	attached	124
procedures	52	promote	123
parameter	52	criteria	123
distinct	52	credit	123
media	51	affecting	123
implement	51	specification	122
establish	51	adequate	122
enhance	51	published	120
variations	50	proportion	120
sectors	50	procedures	120
resource	50	medium	120
published	50	feature	120
principal	50	sections	119
layers	50	crucial	118
expansion	50	communities	118
economy	50	route	117
via	49	modified	117
series	49	integrated	117

job	49	depression	117
analysing	49	cited	117
visual	48	vary	116
responses	48	theoretical	116
priority	48	stability	115
despite	48	recovery	115
constructed	48	predict	115
brief	48	parameter	115
traditional	47	domestic	114
structural	47	sustainable	113
scenarios	47	subsequent	113
reliability	47	options	113
released	47	brief	113
outcomes	47	analysed	113
investigate	47	responses	112
transportation	46	nuclear	112
linked	46	functional	112
generating	46	duration	112
functional	46	awareness	112
enable	46	survival	111
confirmed	46	detection	111
conclude	46	panel	110
survival	45	conventional	110
style	45	proportional	109
occurred	45	expansion	109
medium	45	dimensions	109
manual	45	creating	109

detected	45	linked	108
version	44	contribute	108
varies	44	conclusion	108
status	44	transferred	107
scheme	44	regulation	107
route	44	psychology	107
policies	44	integration	106
perspective	44	export	105
incidence	44	cultural	105
financial	44	funds	104
diversity	44	release	103
decades	44	reactions	103
challenge	44	indication	103
assumed	44	exposure	103
technical	43	economy	103
sexual	43	transition	102
requirement	43	text	102
periods	43	professional	102
extracted	43	inputs	102
evolution	43	visible	101
estimate	43	located	101
depression	43	implications	101
analytical	43	finally	101
analyses	43	diversity	101
acquired	43	derived	101
predicted	42	underlying	100
interact	42	policies	100

innovation	42	decline	100
concluded	42	confirm	100
compound	42	transmission	99
assessed	42	despite	99
topics	41	assignment	99
targets	41	appendix	99
outputs	41	remove	98
innovative	41	ethical	98
highlight	41	enables	98
deviation	41	confirmed	98
contrast	41	regulations	97
categories	41	networks	97
availability	41	bond	97
panel	40	theories	96
methodologies	40	sum	96
inserted	40	significance	96
complexity	40	roles	96
analyzed	40	reliability	96
visible	39	layers	96
shift	39	specified	95
prediction	39	removal	95
precise	39	occurring	95
nuclear	39	hypothesis	95
investments	39	evaluate	95
facilitate	39	chart	95
distributed	39	devices	94
resolution	38	communications	94

removal	38	circumstances	94
occurrence	38	aid	94
link	38	periods	93
environments	38	outcome	93
cycles	38	maintenance	93
assignment	38	implement	93
achieving	38	displacement	93
sex	37	survey	92
sequences	37	innovation	92
sections	37	estimate	92
phenomenon	37	restricted	91
minimize	37	perspective	90
dynamic	37	highlighted	90
duration	37	goals	90
derived	37	concluded	90
computational	37	sought	89
assembly	37	restriction	89
adults	37	manual	89
unique	36	maintained	89
transmitted	36	job	89
regulation	36	evolution	89
phases	36	device	89
inputs	36	detect	89
domain	36	varied	88
demonstrate	36	establish	88
varying	35	dynamic	88
specifically	35	assessed	88

potentially	35	analyse	87
mode	35	visual	86
indicating	35	unique	86
highlighted	35	impacts	86
detect	35	finance	86
creates	35	consequences	86
awareness	35	revealed	85
sufficient	34	minimise	85
interventions	34	finite	85
integration	34	demonstrate	85
focusing	34	conclusions	85
consistent	34	apparent	85
aware	34	accurately	85
utilized	33	strategic	84
suspended	33	respond	84
remove	33	resolved	84
psychological	33	potentially	84
proportion	33	labour	84
professional	33	enhance	84
involving	33	index	83
generates	33	identity	82
document	33	identification	82
diverse	33	identical	82
consumers	33	demonstrates	82
category	33	percent	81
attached	33	indicator	81
valid	32	deviation	81

technological	32	monitored	80
principle	32	induced	80
identifying	32	implies	80
hence	32	fundamental	80
guidelines	32	concepts	80
flexibility	32	codes	80
dimension	32	assumptions	80
channel	32	select	79
bias	32	links	79
adapt	32	induction	79
variance	31	defines	79
triggered	31	varies	78
trend	31	trends	78
transferred	31	phases	78
reverse	31	involvement	78
prior	31	inspection	78
perceived	31	goal	78
partners	31	establishment	78
jobs	31	adults	78
involve	31	acquisition	78
investigated	31	outcomes	77
incentive	31	legal	77
format	31	involving	77
eliminate	31	creative	77
cultures	31	contrast	77
conventional	31	chemicals	77
contributes	31	categories	77

construct	31	targets	76
considerably	31	specifically	76
confirm	31	schedule	76
conduct	31	modification	76
communities	31	files	76
chart	31	extraction	76
author	31	primarily	75
assumptions	31	orientation	75
affecting	31	illustrated	75
adjusted	31	distinct	75
protocol	30	variance	74
neutral	30	processed	74
monitor	30	investigations	74
interval	30	conversion	74
integrated	30	constraints	74
annual	30	availability	74
alterations	30	adjusted	74
transform	29	vision	73
sustainability	29	revenue	73
significance	29	researchers	73
reinforced	29	promotion	73
parallel	29	eventually	73
injury	29	selective	72
identity	29	maintaining	72
files	29	corporate	72
consisted	29	perceived	71
automated	29	obviously	71

simulated	28	media	71
reactive	28	margin	71
ratios	28	logical	71
precision	28	discrete	71
maintained	28	dimensional	71
enabling	28	authority	71
disposal	28	altered	71
decline	28	transmitted	70
contribution	28	partner	70
assigned	28	minor	70
virtual	27	format	70
submitted	27	flexible	70
specified	27	evolutionary	70
simulate	27	designing	70
schedule	27	channels	70
proportional	27	transformation	69
outcome	27	participation	69
maintaining	27	intensive	69
framework	27	exposed	69
formula	27	computing	69
finite	27	adapted	69
corresponding	27	undertaken	68
transformed	26	scope	68
text	26	outputs	68
strategic	26	incidence	68
networks	26	contract	68
locations	26	considerable	68

isolated	26	category	68
interpretation	26	author	68
ethical	26	transport	67
dynamics	26	sequential	67
displayed	26	minimal	67
chemicals	26	isolated	67
acquire	26	focused	67
academic	26	facilities	67
validate	25	complexity	67
theories	25	achieving	67
specifications	25	version	66
routes	25	unstable	66
restriction	25	promoting	66
respond	25	intermediate	66
residents	25	concentrated	66
residential	25	communicate	66
recovery	25	capable	66
random	25	assumption	66
minimal	25	assume	66
labour	25	methodology	65
intense	25	constructed	65
illustrate	25	automatically	65
evident	25	teams	64
consumer	25	statistical	64
concentrated	25	illustrates	64
anticipation	25	challenge	64
analyze	25	authorities	64

adapted	25	attributed	64
summary	24	attribute	64
qualitative	24	shifts	63
modified	24	radical	63
hypotheses	24	evident	63
domains	24	variations	62
contributing	24	transform	62
consumed	24	topic	62
conclusions	24	subsequently	62
administration	24	scenario	62
utilization	23	retention	62
survey	23	professionals	62
select	23	phenomenon	62
restricted	23	estimation	62
modes	23	displays	62
migration	23	creates	62
investigations	23	consisting	62
initiatives	23	rely	61
extract	23	consequence	61
enhanced	23	bonds	61
contemporary	23	welfare	60
codes	23	secure	60
uniform	22	requiring	60
revealed	22	induce	60
react	22	dominant	60
percent	22	considerably	60
partnership	22	assembly	60

margin	22	abnormal	60
involvement	22	technological	59
invest	22	regional	59
inhibition	22	precise	59
induced	22	cycles	59
implementing	22	contribution	59
flexible	22	adult	59
facility	22	sustainability	58
emerged	22	logic	58
economical	22	implementing	58
dominant	22	generating	58
colleagues	22	gender	58
attitudes	22	funding	58
accumulation	22	flexibility	58
abandoned	22	facilitate	58
vision	21	assigned	58
variability	21	subsidies	57
targeted	21	estimates	57
specificity	21	entities	57
perception	21	ensuring	57
motivation	21	converted	57
modifications	21	conclude	57
label	21	authors	57
injuries	21	validity	56
inhibit	21	released	56
induction	21	perception	56
indicator	21	offset	56

functioning	21	item	56
expanded	21	creation	56
differentiate	21	assistance	56
comprehensive	21	thereby	55
beneficial	21	instruction	55
aggregate	21	highlights	55
trigger	20	extract	55
sustained	20	environments	55
randomly	20	decades	55
promoting	20	computational	55
preliminary	20	vehicle	54
plus	20	ultimately	54
passive	20	symbol	54
labelled	20	preliminary	54
integrating	20	identifying	54
indicators	20	exclusion	54
income	20	contributed	54
eventually	20	consequently	54
defining	20	bias	54
constantly	20	attitudes	54
compatible	20	analytical	54
capability	20	analyses	54
bond	20	valid	53
assume	20	sectors	53
aggregates	20	regime	53
adjust	20	paradigm	53
abnormal	20	motivation	53

volumes	19	insight	53
visualization	19	functioning	53
substitution	19	dramatic	53
shifted	19	corresponds	53
schemes	19	alter	53
requiring	19	allocated	53
regulations	19	protocol	52
registered	19	participate	52
ranging	19	investments	52
processed	19	insertion	52
logic	19	facility	52
integrity	19	exhibit	52
integral	19	enhanced	52
insight	19	creativity	52
index	19	contacts	52
excluded	19	adaptive	52
ensuring	19	participatory	51
differentiation	19	injuries	51
demonstrates	19	emphasis	51
definitions	19	diverse	51
criterion	19	specify	50
corresponds	19	sex	50
consist	19	ranges	50
commitment	19	radicals	50
alternatives	19	investors	50
aid	19	insufficient	50
abstract	19	innovative	50

validation	18	empirical	50
utilised	18	commentary	50
transformation	18	adjustments	50
survive	18	acquired	50
similarities	18	terminal	49
seeking	18	randomly	49
seek	18	modifications	49
scope	18	indicative	49
revolution	18	incorporated	49
relevance	18	evaluated	49
rejection	18	erosion	49
predictions	18	coherence	49
philosophy	18	classical	49
phenomena	18	assessing	49
normalized	18	adjustment	49
monitored	18	sequences	48
investigating	18	priority	48
interpreted	18	modify	48
interacting	18	economies	48
insertion	18	defining	48
focuses	18	debate	48
evolve	18	consist	48
estimates	18	conflicts	48
defines	18	bonding	48
cycling	18	acquisitions	48
attributed	18	sufficiently	47
approximation	18	simulated	47

altered	18	plus	47
adaptation	18	locations	47
temporary	17	interpretation	47
specification	17	inhibition	47
selecting	17	indicators	47
roles	17	fluctuations	47
residence	17	exceed	47
researcher	17	differentiation	47
participation	17	colleagues	47
orientation	17	assist	47
occurring	17	approximate	47
negatively	17	accumulation	47
modify	17	simulate	46
minimise	17	prediction	46
legislation	17	legislation	46
implications	17	journal	46
illustrated	17	extracted	46
highlights	17	definitely	46
gender	17	contracts	46
experts	17	computers	46
expand	17	capability	46
exclusive	17	assuming	46
evolutionary	17	utilised	45
debate	17	temporary	45
coordinates	17	margins	45
convert	17	manipulation	45
controversial	17	interventions	45

classic	17	interval	45
channels	17	integrity	45
attributes	17	ensures	45
assure	17	documented	45
assessing	17	coupling	45
alteration	17	construct	45
accessible	17	concentrate	45
accessibility	17	assurance	45
validity	16	appropriately	45
undertaken	16	whereby	44
terminal	16	variability	44
structured	16	reverse	44
stressful	16	purchases	44
retention	16	portion	44
relies	16	interact	44
regulators	16	institutions	44
reactors	16	inhibit	44
psychology	16	illustrate	44
modification	16	experts	44
imagery	16	dimension	44
illustrates	16	decade	44
funding	16	survive	43
expertise	16	structured	43
enhancing	16	specifications	43
empirical	16	promotes	43
emphasis	16	maximise	43
domestic	16	injured	43

coordination	16	evolved	43
contributed	16	dynamics	43
complement	16	dramatically	43
circumstances	16	cooperation	43
challenging	16	conduct	43
appendix	16	collapse	43
adult	16	coded	43
thereby	15	bulk	43
statistics	15	widespread	42
specify	15	symbols	42
shifts	15	reactive	42
shifting	15	psychologists	42
seeks	15	predictions	42
rigid	15	isolation	42
protocols	15	ignored	42
overlap	15	expand	42
invested	15	derivative	42
intrinsic	15	consuming	42
injured	15	conflict	42
infinite	15	challenges	42
infer	15	attitude	42
induce	15	attachment	42
inadequate	15	neutral	41
implies	15	initiated	41
expanding	15	induces	41
evolved	15	focuses	41
discrimination	15	correspond	41

diminish	15	consumed	41
designs	15	adjacent	41
coupling	15	seek	40
converted	15	removing	40
consensus	15	philosophy	40
chapter	15	intrinsic	40
bulk	15	intervals	40
assumption	15	instructions	40
adjustments	15	infrastructure	40
widespread	14	generates	40
visualize	14	eliminate	40
substitute	14	displaying	40
subsequently	14	cyclic	40
removing	14	volumes	39
rely	14	utilized	39
releases	14	targeting	39
regulated	14	qualitative	39
register	14	prospects	39
promotion	14	principal	39
promotes	14	investigating	39
participate	14	inhibits	39
participant	14	exporting	39
obviously	14	expertise	39
normality	14	coupled	39
intermediate	14	contributing	39
intelligent	14	comprehensive	39
insufficient	14	alternatives	39

initially	14	statistics	38
incident	14	retain	38
identifies	14	relevance	38
expose	14	regulated	38
expert	14	priorities	38
establishing	14	prime	38
entity	14	oriented	38
enhancement	14	occurrence	38
energetic	14	instability	38
enables	14	expanded	38
economically	14	excluded	38
designing	14	documents	38
designers	14	administration	38
couple	14	virtual	37
consistency	14	undergo	37
computers	14	reveal	37
civil	14	participant	37
briefly	14	obtaining	37
approximate	14	minimize	37
adaptive	14	initiator	37
validated	13	highlight	37
unstable	13	focusing	37
undergo	13	estimations	37
summarize	13	eliminated	37
substituted	13	cultures	37
similarity	13	capabilities	37
restrictions	13	uniform	36

registers	13	targeted	36
reconstruction	13	summarised	36
portion	13	style	36
perspectives	13	selecting	36
partner	13	schematic	36
minor	13	positively	36
links	13	issued	36
justified	13	initiation	36
items	13	infinite	36
institutions	13	incorporate	36
initiative	13	expanding	36
incorporated	13	establishing	36
hierarchy	13	detector	36
globally	13	definitions	36
formulation	13	currency	36
exhibitions	13	challenging	36
distributions	13	analysing	36
disposed	13	sourcing	35
constants	13	shifted	35
conceptual	13	partners	35
authorities	13	migration	35
assuming	13	mature	35
analyzing	13	guarantee	35
allocated	13	consume	35
adjustment	13	consultation	35
adjusting	13	constraint	35
adaptations	13	constantly	35

accurately	13	adaptation	35
tension	12	academic	35
targeting	12	transmit	34
surveys	12	purchased	34
regime	12	promoted	34
ranges	12	predominantly	34
priorities	12	physically	34
manually	12	perceptions	34
isolation	12	invest	34
isolate	12	interpret	34
investing	12	inadequate	34
initiated	12	enabling	34
imposed	12	contributes	34
founded	12	transformed	33
foundation	12	sustained	33
evaluating	12	similarly	33
equipped	12	ranging	33
dramatically	12	mediated	33
decade	12	labelled	33
credit	12	interactive	33
creative	12	integrate	33
correspond	12	individually	33
constructions	12	incident	33
constraints	12	expert	33
classical	12	domains	33
clarify	12	disposal	33
automatic	12	displaced	33

adapting	12	demonstrating	33
acquisition	12	convert	33
achieves	12	charts	33
varied	11	assumes	33
trends	11	unaffected	32
transmit	11	sexual	32
schematic	11	regulatory	32
regional	11	recover	32
pursuit	11	persistent	32
promoted	11	passive	32
principally	11	liberalisation	32
primarily	11	intense	32
predicting	11	inherent	32
obtaining	11	imposed	32
nevertheless	11	financing	32
mutual	11	enhancing	32
minimizing	11	enabled	32
minimising	11	distortion	32
maximize	11	confirms	32
item	11	adjust	32
intervals	11	adapt	32
integrate	11	vehicles	31
instructions	11	restrict	31
instruction	11	regulate	31
insecurity	11	phenomena	31
individually	11	interpreted	31
facilitates	11	exports	31

ethics	11	evaluating	31
distributor	11	elimination	31
distinctive	11	couple	31
displays	11	consultant	31
differentiated	11	compatible	31
detector	11	analyze	31
demonstrating	11	allocation	31
cooperation	11	validation	30
contrary	11	shifting	30
consume	11	revolution	30
confirming	11	restrictions	30
conducting	11	respondents	30
comprises	11	recovered	30
complementary	11	predicting	30
committed	11	partnership	30
automatically	11	maturity	30
authority	11	investing	30
assemble	11	encountered	30
apparent	11	dominated	30
achievement	11	designer	30
underlying	10	constrained	30
transported	10	compute	30
traditions	10	automatic	30
theoretically	10	analyzed	30
teams	10	utility	29
sustain	10	ultimate	29
summarized	10	transportation	29

summarised	10	researched	29
substituting	10	relies	29
stresses	10	norms	29
stabilized	10	modes	29
revenue	10	max	29
relaxed	10	inhibited	29
regulatory	10	ethnic	29
quoted	10	designers	29
publication	10	criterion	29
projected	10	converting	29
positively	10	constants	29
persistent	10	commitment	29
persistence	10	automated	29
paradigm	10	utilization	28
occupation	10	underwent	28
military	10	trace	28
manipulation	10	substitution	28
legal	10	submitted	28
justify	10	revenues	28
interpretations	10	regulating	28
instability	10	precision	28
insert	10	orientated	28
incentives	10	occupied	28
identical	10	lecture	28
highlighting	10	inserted	28
grade	10	inappropriate	28
foundations	10	fund	28

environmentally	10	emerged	28
enormous	10	coordination	28
diversified	10	consisted	28
dimensional	10	coherent	28
detecting	10	briefly	28
cyclic	10	assure	28
converting	10	approximation	28
coherent	10	alternating	28
coding	10	supplement	27
briefing	10	succession	27
attitude	10	react	27
assist	10	quoted	27
whereby	9	predicts	27
visualise	9	justified	27
transformations	9	intelligent	27
symbols	9	institutional	27
suspension	9	imply	27
sufficiently	9	exceeds	27
subsequent	9	economical	27
styles	9	distinction	27
simulating	9	appreciate	27
secure	9	annually	27
retained	9	virtually	26
restoration	9	utilise	26
responds	9	supplementary	26
removes	9	rigid	26
rejected	9	restraints	26

regimes	9	prospective	26
recover	9	practitioners	26
predominantly	9	incorporating	26
physically	9	incomes	26
perceptions	9	fees	26
participated	9	extracts	26
modifying	9	differentiate	26
lecture	9	denoted	26
institute	9	assignments	26
inferred	9	altering	26
imply	9	accessible	26
ignored	9	trigger	25
expands	9	sustain	25
emerging	9	stressed	25
emergence	9	stabilisation	25
elimination	9	solely	25
eliminated	9	rejected	25
documents	9	regulates	25
designer	9	publication	25
consuming	9	proceed	25
constitute	9	pose	25
biased	9	mutual	25
assumes	9	manipulate	25
approximated	9	maintains	25
alter	9	locate	25
administrative	9	jobs	25
achievements	9	excluding	25

accessed	9	exclude	25
welfare	8	diversification	25
versions	8	arbitrary	25
underwent	8		

Appendix 5 - First sublist of the AWL

BAWE		Brazilian students		
analyse	513	analyse	158	4.88
analysed	520	analysed		0.32
analyser	40			
analysers	8			
analyses	204	analyses		13.32
analysing	285	**analysing**	49	30.39
analysis	3429	analysis	523	454.93
analyst	38			
analysts	47			
analytic	40			
analytical	206	analytical	43	13.84
analytically	21			
analyze	93	analyze	25	2.33
analyzed	93	analyzed		-0.51
analyzes	12			
analyzing	71	**analyzing**	13	6.63
approach	2750	approach	201	806.74
approachable	7			
approached	95	approached		27.71
approaches	663	approaches		177.13
approaching	62	**approaching**	7	69.62
area	2416	area	493	172.84
areas	2100	areas	251	393.11
assess	454	assess	62	71.4
assessed	253	assessed		28.88

assesses	35			
assessing	197	**assessing**	17	50.98
assessment	618			151.11
assessments	67	assessment	57	
assume	452	assume	20	174.43
assumed	600	assumed		175.61
assumes	224	assumes		89.92
assuming	212	**assuming**	13	69.52
assumption	511	assumption	15	228.35
assumptions	498	assumptions	31	161.87
authoritative	68			
authorities	427	authorities	13	188.73
authority	874	authority	11	467.21
availability	312	availability	41	51.76
available	2010	available	197	464.88
beneficial	541	beneficial	21	220.08
beneficiaries	43			
beneficiary	35			
benefit	911	benefit	62	208.28
benefited	103	benefited		38.14
benefiting	41			
benefits	1189	benefits	135	236.25
concept	1743	concept	95	609.73
conception	361	conception	5	190.2
concepts	620	concepts	64	136.32
conceptual	167	conceptual	13	46.84
conceptualisation	28			

conceptualise	12			
conceptualised	36			
conceptualises	5			
conceptualising	7			
conceptually	17			
consist	177	consist	19	37.38
consisted	149	consisted		12.05
consistency	153	consistency	14	37.67
consistent	605	consistent	34	208.26
consistently	167	consistently	7	65.92
consisting	175		14	74.44
consists	418		10	18.57
constituencies	15		11	
constituency	22		76	
constituent	46		8	7.71
constituents	43		35	
constitute	162		10	56.11
constituted	89			30.64
constitutes	154			
constituting	24	**constituting**	6	0.86
constitution	104			
constitutional	89			
constitutionally	11			
constitutions	13			
constitutive	37	constitutive	7	3.21
context	1401	context	77	487.97
contexts	234	contexts	5	113.56

contextual	78			
contextualise	7			
contextualised	10			
contextualising	6			
contextualized	8			
contract	617	contract	7	334.63
contracted	43			
contracting	63			
contractor	22			
contractors	44			
contracts	196			
create	1473	create	160	307.23
created	1360	created		255.62
creates	463	creates		132.65
creating	732	**creating**	75	162.16
creation	569	creation	61	120.34
creations	19			
creative	203	creative	12	67.97
creatively	13			
creativity	146	creativity	8	50.93
creator	43	creator	5	8.31
creators	12			
data	4523	data	707	575.62
definable	9			
define	573	define	71	102.63
defined	1301	defined		191.29
defines	292	defines		95.46

defining	230	**defining**	20	59.15
definition	1013	definition	54	358.43
definitions	256	definitions	19	74.3
derivation	40			
derivative	75	derivative	7	18.15
derivatives	36	derivatives	5	5.53
derive	115	derive	8	34.88
derived	513	derived		151.88
derives	57			
deriving	30			
dissimilar	54			
distribute	47			
distributed	276	distributed		41.3
distributing	28			
distribution	1134	distribution	163	165.37
distributional	10			
distributions	55	distributions	13	2.46
distributive	17			
distributor	10	distributor	11	-5.88
distributors	22	distributors	5	1.13
economic	3388	economic	127	1396.36
economical	144	economical	22	19.05
economically	192	economically	14	54.42
economics	206			
economies	475	economies	6	253.78
economist	28			
economists	99			

economy	1778	economy	50	804.53
environment	2160	environment	254	412.23
environmental	1205	environmental	249	47.65
environmentalist	7	environmentalists	6	
environmentalists	16			
environmentally	64	environmentally	10	8.15
environments	250	environments	38	33.36
establish	580	establish	51	147.78
established	1128	established		379.86
establishes	116			
establishing	261	**establishing**	14	93.06
establishment	360	establishment	7	178.32
establishments	57			
estimate	334	estimate	43	56.95
estimated	588	estimated		99.74
estimates	210	estimates		54.65
estimating	60	**estimating**	5	15.97
estimation	156	estimation	7	59.84
estimations	59			
evidence	3163	evidence	101	1377.31
evidenced	55	evidenced		13.62
evident	557	evident	25	213.64
evidential	14			
evidently	70			
export	335	export	7	163.39
exported	37	exported		5.91
exporter	27			

exporters	22			
exporting	67			
exports	260			
factor	1705	factor	182	362.23
factored	6			
factors	2718	factors	328	502.94
finance	324	finance	6	162.1
financed	39			
finances	64			
financial	1563	financial	44	707.03
financially	77			
financing	85			
formula	345	formula	27	96.37
formulae	43			
formulas	39	formulas	5	6.69
formulate	64	formulate	5	17.9
formulated	109	formulated		41.41
formulating	52			
formulation	104	formulation	13	18.42
formulations	13			
function	2304	function	286	411.73
functional	403	functional	46	79.6
functionally	30			
functioned	36			
functioning	189	**functioning**	21	38.47
functions	891	functions		209.06
identifiable	72			

identification	369	identification	67	35.1
identified	1181	identified		272.59
identifies	205	identifies		62.93
identify	954	identify	168	96.86
identifying	255	**identifying**	32	39.82
identities	200			
identity	1052	identity	29	481.63
illegal	109	illegal	8	31.88
illegality	27			
illegally	8			
income	1085	income	20	543.34
incomes	87			
inconsistencies	48			
inconsistency	65			
inconsistent	98			
indicate	686	indicate	92	110.68
indicated	488	indicated		84.66
indicates	783	indicates		191.97
indicating	377	**indicating**	35	91.65
indication	304	indication	6	150.11
indications	40			
indicative	152	indicative	8	54.13
indicator	220	indicator	21	52.15
indicators	169	indicators	20	32.01
individual	3049	individual	145	1139.77
individualised	24			
individualism	79			

individualist	22			
individualistic	29			
individuality	48			
individually	179	individually	11	58.62
individuals	1599	individuals	82	576.8
insignificant	132			
interpret	223	interpret	7	97.63
interpretation	823	interpretation	26	359.59
interpretations	236	interpretations	10	92.74
interpretative	17			
interpreted	349	interpreted		125.54
interpreting	94			
interpretive	26			
interprets	32			
invariably	34			
involve	416	involve	31	120.4
involved	1707	involved		478.41
involvement	458	involvement	22	170.42
involves	689	involves		192.26
involving	375	**involving**	33	95.48
issue	1546	issue	124	423.96
issued	102			
issues	2019	issues		553.52
issuing	43			
labor	146			
labour	1723	labour	25	901.05
labouring	14			

labours	17			
legal	1306	legal	10	740.81
legality	33			
legally	113			
legislate	9			
legislating	5			
legislation	505	legislation	17	216.14
legislative	130	legislative	5	-53.07
legislator	9			
legislators	7			
legislature	42			
major	1710	major	164	403.61
majorities	5			
majority	1060	majority	57	373.39
method	2850	method	340	534.71
methodical	5			
methodological	98			
methodologies	56	methodologies	40	-9.18
methodology	303	methodology	82	7.33
methods	1857	methods	238	318.52
misinterpret	7			
misinterpretation	25			
misinterpreted	25			
occur	1226	occur	112	302.31
occurred	724	occurred		235.54
occurrence	189	occurrence	38	14.07
occurrences	63	occurrences	5	17.42

occurring	337	**occurring**	17	122.54
occurs	890	occurs		139.23
overestimated	26			
percent	373	percent	22	125.05
percentage	567	percentage	89	71.64
percentages	56			
period	2633	period	132	960.27
periodic	56	periodic	8	8.24
periodical	9			
periodically	20			
periods	484	periods	43	122.25
policies	1135	policies	44	461.96
policy	2278	policy	64	1031.05
principle	943	principle	32	402.51
principled	17			
principles	885	principles	63	264.34
procedural	106	procedural	6	36.35
procedure	633	procedure	117	57.86
procedures	422	procedures	52	76.11
proceed	117	proceed	6	42.2
proceeded	26			
proceeding	47			
proceedings	107			
proceeds	72			
process	4408	process	792	428.93
processed	172	processed		35.23
processes	1350	processes		146.58

processing	576	**processing**	60	125.49
reassess	13			
reassessed	8			
reassessment	16			
recreate	18			
recreated	7			
redefine	32			
redefined	25			
redefining	14			
redistribute	16			
redistributed	15			
redistributing	9			
redistribution	74			
reformulated	7			
reformulation	5			
reinterpret	8			
reinterpretation	9			
reoccurring	5			
require	800	require	52	253.35
required	2350	required		595.46
requirement	390	requirement	43	80.05
requirements	743	requirements	100	119.28
requires	822	requires		169.99
requiring	182	**requiring**	19	39.56
research	3630			1063.02
researched	113	researched		36.83
researcher	405	researcher	17	159.77

researchers	509	researchers	83	60.08
researches	47	researches		-43.44
researching	61			
respond	348	respond	25	103.31
responded	99	responded		35.98
respondent	33			
respondents	117	respondents	5	45.82
responding	120			
responds	64	responds		9.65
response	1337	response	161	248.04
responses	484	responses	48	110.69
responsive	65	responsive	8	11.74
responsiveness	44	responsiveness	6	6.93
restructure	24			
restructured	5			
restructuring	82			
role	3091	role	127	1229.93
roles	757	roles	17	363.17
section	1908	section	150	531.19
sectioned	11			
sections	425	sections	37	109.19
sector	954	sector	90	228.59
sectors	263	sectors	50	22.52
significance	741	significance	29	300.51
significant	2637	significant	206	737.62
significantly	768	significantly	76	181.12
signified	49			

signifies	51			
signify	57			
signifying	26			
similar	2434	similar	190	681.21
similarities	283	similarities	18	90.81
similarity	157	similarity	13	42.01
similarly	202	similarly	8	81.55
source	1395	source	137	322.03
sourced	58			
sources	888	sources	124	135.24
sourcing	58	**sourcing**	8	8.99
specific	1893	specific	293	241.01
specifically	448	specifically	35	125.31
specification	270	specification	17	87.17
specifications	116	specifications	25	7.09
specificity	78	specificity	21	1.94
specifics	14			
structural	684	structural	47	209.14
structurally	42			
structure	2821	structure	213	808.88
structured	238	structured		73.8
structures	883	structures		283.85
structuring	36			
theoretical	671	theoretical	57	175.9
theoretically	93	theoretically	10	19.61
theories	1120	theories	25	538.08
theorist	42			

theorists	172			
theory	3841	theory	107	1742.7
unavailable	46			
unconstitutional	11			
undefined	13			
underestimate	22			
underestimated	67			
underestimates	12			
underestimating	5			
unidentifiable	7			
unresponsive	11			
unstructured	27			
variability	107	variability	21	8.48
variable	869	variable	134	113.19
variables	1144	variables	81	342.94
variance	206	variance	31	27.94
variant	23			
variants	43	variants	6	6.56
variation	549	variation	119	32.98
variations	242	variations	50	16.73
varied	311	varied		130.88
varies	231	varies	44	19.69
vary	389	vary	76	31.17
varying	406	**varying**	35	105.16

Appendix 6 – Ethical Form

Research Ethics Form

Centre for Applied Linguistics, University of Warwick

Ethical Review of Masters Research Projects

CAL is committed to ensuring that all research undertaken by it members, both staff and students, meets the highest possible ethical standards. You will already have been introduced to research ethics in the context of the MA/MSc Research Methods module, but now that you are about to embark on a research project it is essential that you consider very carefully the ethical issues that it might raise and that you discuss these with your supervisor. Please treat this not only as a means of ensuring that your research meets appropriate ethical standards but also as a learning opportunity.

Please note that data collection must not begin until this form has been completed and signed.

Name of student: Larisse Goulart da Silva

Name of supervisor: Sue Wharton

Title of proposed research project: Analysing written products of Brazilians studying in UK universities.

1. We confirm that we have read the following documents:

   The University of Warwick's 'Statement of the Ethical Conduct of Research'
   http://www2.warwick.ac.uk/services/ris/research_integrity/code_of_practice_and_policies/statement_ethical_conduct_research

   The British Association for Applied Linguistics 'Recommendations on Good Practice in Applied Linguistics'

   http://www.baal.org.uk/dox/goodpractice_full.pdf

2. We confirm that we have discussed the ethical implications of the proposed research project
and that it is consistent with the principles laid down in the above documents.

3. Where the research involves the analysis of textual data, we also confirm that we have
considered copyright issues and are satisfied that the project does not violate copyright laws.

4. We confirm that if the research involves vulnerable individuals (e.g. children), deception, or
any other practices that raise complex ethical issues, we have brought this to the attention of the Director of Graduate Studies.

5. We confirm that we have agreed a procedure for dealing with any ethical issues that might
arise during the course of the project.

Signed:

Student Laiissa Goulart da Silva Date 12/02/2016

Supervisor [signature] Date 12/2/16

Please return this form, signed by both parties, to the MA/MSc Secretary by no later than the end of May or before data collection begins, whichever is the sooner.