UNIVERSIDADE FEDERAL DE MINAS GERAIS

ANDRESSA RODRIGUES GOMIDE

Processing a learner corpus to identify differences: the influence of task, genre and student background

Belo Horizonte

2016

ANDRESSA RODRIGUES GOMIDE

Processing a learner corpus to identify differences:

the influence of task, genre and student background

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais, como requisito parcial para obtenção de título de MESTRE em Linguística Aplicada

Área de Concentração: Linguística Aplicada

Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras

Orientadora: Prof. Deise Prina Dutra

Belo Horizonte

Faculdade de Letras da UFMG

2016

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

G633p Processing a learner corpus to identify differences [manuscrito]: the influence of task, genre and student background / Andressa Rodrigues Gomide. – 2016.
107 f., enc. : il., (color), (p&b).

Orientadora: Deise Prina Dutra.

Área de concentração: Linguística Aplicada.

Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 98 -102.

Apêndices: f. 103 -106.

1. Língua inglesa – Estudo e ensino – Falantes estrangeiros – Teses. 2. Aquisição da segunda linguagem – Teses. 3. Linguística de corpus – Teses. 4. Língua inglesa – Gramática – Teses. I. Dutra, Deise Prina. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD: 420.7

ACKNOWLEDGEMENTS

Prof. Deise Prina Dutra

Gabi ♥

Mãe, pai, irmã

Fred, Bruno, Giulia

Cláudia

Geraldinho, Luiza

Jacque, Lu, Picinin, Laís

FAPEMIG

"It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

Sherlock Holmes in Sir Conan Doyle's *A Scandal in Bohemia*

ABSTRACT

This master thesis deals with the technical and methodological aspects in creating, cleaning and processing a Brazilian university level learner corpus, the Corpus do Inglês sem Fronteiras (CorIsF) v 1.0. The two main goals of this study consist of making the processing of CorIsF replicable and in investigating and describing the variation of some linguistic characteristics across different learner groups, tasks and genres. The procedure was carried in R, a free software environment for statistical computing and graphics, and was divided in four parts: dataset compilation and pre-processing; dataset processing; extraction of the key features; and data visualization. The first step deals with the method used to collect the data and to do the first cleaning process, such as eliminating unwanted data and keeping the relevant ones. In the following step, CorIsF was subset in five small corpora covering different learner profiles, two different tasks, and on genre, and annotated with a part-of-speech (POS) tagger. In the third step the variability of POS within subcorpora, the frequency of types and tokens, and the usage of n-grams were investigated. In the final step some exploratory data visualization were performed with the creation and analysis of plots and wordclouds. After the preparation of the data, the language used in each subcorpora was contrasted and analysed, suggesting that task, genre and student background are likely to influence learners' written production.

Keywords: learner corpus; corpus design; English for Academic Purpose

RESUMO

Esta dissertação trata dos aspectos técnicos e metodológicos na criação, limpeza e processamento de um corpus de nível universitário de aprendizes brasileiros, o Corpus do Inglês sem Fronteiras (CorIsF) v 1.0. Os dois principais objetivos deste estudo consistem em tornar replicável o processamento do CorIsF e em investigar e descrever a variação de algumas características linguísticas em diferentes perfis de alunos, tarefas e gêneros. O procedimento foi realizado com auxílio da ferramenta R, um ambiente de software livre para computação estatística e gráfica, e foi dividido em quatro partes: a compilação e o pré-processamento do conjunto de dados; o processamento do corpus; a extração de principais aspectos; e a visualização de dados. O primeiro passo lida com os passos utilizados para coletar os dados e fazer o primeiro processo de limpeza, tais como a eliminação de dados indesejados e manutenção de informações relevantes. No passo seguinte, CorIsF foi subdivido em cinco pequenos corpora que cobrem diferentes perfis de alunos, tarefas e gênero e anotado com um etiquetador de classes gramaticais. No terceiro passo, a variabilidade de classes gramaticais em cada subcorpus, a frequência de *types* e *tokens*, e a utilização de n-gramas foram investigados. Na etapa final algumas visualizações como nuvens de palavras e gráficos foram geradas para análise dos dados. Após a preparação dos dados, a linguagem utilizada em cada subcorpora foi contrastada e analisada, sugerindo que a tarefa, o gênero e o perfil aluno são propensos a influenciar a produção escrita dos alunos.

Palavras-chave: corpus de aprendiz; desenho de corpus; inglês para fins acadêmicos

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BAWE	British Academic Written English
CLC	Cambridge University Press Learner Corpus
CIA	Contrastive Interlanguage Analysis
CorIsF	Corpus do Idioma sem Fronteiras
DDL	Data-driven learning
DPU	Delayed Pedagogical Use
ESP	English for Special Purposes
EDA	Exploratory Data Analysis
L1/L2	First / Second language
HKUST	Hong Kong University of Science and Technology
IsF	Idiomas/Inglês sem Fronteiras
IPU	Immediate Pedagogical Use
ICLE	International Corpus of Learner English
IELTS	International English Language Testing System
KWIC	Key word in context
LC	Learner Corpus
LDD	Learning-driven data
LLC	Longman Learner Corpus
MI	Mutual Information
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part-of-speech
SLA	Second Language Acquisition
TOEFL	Test of English as a Foreign Language
FRIDA	The French Interlanguage Database
LINDSEI	The Louvain International Database of Spoken English Interlanguage

TABLE OF CONTENTS

1 INTRODUCTION

Learner Corpus Research (LCR) has started in the late 1980s as a branch of Corpus Linguistic Research and has been gaining ground since then. The recent emergence of easy access to cloud architectures, open source software and commodity hardware has made the use of Learner Corpus (LC) reach a wider public, including small research groups and even a single researcher. The compilation of a corpus, which would once be a rather expensive and time-consuming project, is now feasible even for users that are not computer savvy.

The users of LC are also not restricted to a specific group, and LRC has proved to be an interdisciplinary useful for a great diversity of public. For instance, regarding Second Language Acquisition, LC can help in the description of different developmental stages, while for Natural Language Processing it can aid the development of automatic text scoring and error detection. Considering the field of Second Language Education, LC can be handy in informing pedagogy and assisting language teaching.

To what English for Academic Purposes is concerned, learner data have been massively collected in projects such as the International Corpus of Learner English (ICLE) (GRANGER ET AL., 2009), and BATMAT Corpus (LINDGRÉN, 2013). Considering the growing tendency of internationalization within the Brazilian academic context and the steady flow of learner English production, the main motivation underlying this work stems from the fact this production can be harnessed for the compilation of a constantly growing corpus.

What is aimed with the present work is to provide two main contributions to learner corpora methodology. The first one deals with the process of systematically gathering and processing the textual production of learners of English. The second contribution

is to offer to the scientific community a well-documented and accessible corpus with the written production of Brazilian learners of English.

1. 1 Goals and Objectives

1.1.1 Goals

1. Clean and process the CorIsF, tag it for POS and make the process replicable.
2. Verify and describe how some linguistic characteristics vary when different learner groups and different tasks or genres are considered.

1.1.2 Objectives

Objectives associated with the goals stated in number 1:

1. Anonymize the learners.
2. Simplify the data by excluding extra information such as answers to the multiple-choice questions present in the tests[1].
3. Tag the corpus for parts-of-speech (POS) using the package OpenNLP (HORNIK, 2014) for R.

Objectives associated with the goals stated in number 2:

1. Subset the data of CorIsF in the following 5 small corpora: production from
    a. (1) integrated and (2) independent tasks;
    b. production of learners from different faculty courses separating them according to (3) high and (4) low demand courses;

[1] the data collection of CorIsF is presented to students in the format of a test. They have first to answer some multiple-choice questions related to a text and a video and later they write a text following a prompt.

   c. one corpus with the (5) summaries produced on the academic writing course.

2. Identify the most frequent types, tokens, POS and n-grams across different subcorpora;
3. Compare the n-grams in the different subcorpora;
4. Make scripts available, so that the study can be replicable as the corpus grows.

1.2 Organization of the thesis

This study is organized as follows. Chapter 2 presents some of the main applications of learner corpora by discussing its role in Second Language Acquisition, language description and pedagogical uses. The chapter also presents some of the learner corpora around the world, and provides an outline of the most common learner corpus design criteria.

Chapter 3 is a discussion on earlier and current methods used for linguistic features extraction and a description of some most commonly used corpus analytic techniques. The chapter first presents a brief description of the approach to corpus taken by neo-firthian and functionalist linguists. It then outlines some of the most frequently linguistic features extracted from a corpus. A final section addresses the tools used for corpus exploration.

Chapter 4 is dedicated to the methodological procedures and resources used for the construction of the dataset and for the subsetting of the small corpora. The first section deals with the method utilized to collect the data and do the first cleaning process, such as eliminating unwanted data and keeping the relevant ones. The following section presents the subsetting the annotation procedures. The third section describes the process of identifying the variability of parts-of-speech (POS); the

frequency of types and tokens; and the n-grams. In the final section the scripts developed to easily generate data visualization are presented.
In chapter 5, the analysis and interpretation of the data are presented. The analysis is divided in three parts, addressing the following linguistic features for all the aforementioned subcorpora: types and tokens frequency; the POS usage; and the distribution on n-grams. Chapter 6 summarizes the main achievements of this study.

2 LEARNER CORPORA

Native speaker corpora have been widely used to inform second language teaching. These corpora have proved to be useful especially for revealing some language aspects which would be harder identified through intuition. However, when second language teaching is concerned, the use native language description alone is not sufficient. It is also necessary to identify what the main gaps and difficulties of learners of second language are (NESSELHAUF, 2004). This chapter will thus discuss the main applications of learner corpora, present some of the learner corpora around the world, and review some corpus design criteria.

2.1 Some Learner Corpora

Learner corpora have been around for quite some time now and more and more projects are focusing on learner production. Considering learner corpora in a more typical sense, it is believed that their compilations started in the late 1980s with the Longman Learner Corpus and became stronger in the 1990s (NESSELHAUF, 2004). It is worth emphasizing, though, that the idea of gathering learner data was not new back then. Several learner languages had already been collected, especially in the 1960s and 1970s, for the purpose of error analysis (GRANGER, 1998).

Differently from what happened with the earlier corpora, error analysis is far from being the main purpose of current learner corpora. Learner corpora are commonly described as "systematic computerized collections of texts produced by learners" (NESSELHAUF, 2007, p. 40) and the data collected in the 60s and 70s were neither systematic nor a considerably large computerized collection.

More than four decades after the compilation of the first learner corpora, the most common medium of learner corpora remains being the written production of learners of English, despite the advent of new technologies to collect spoken language. This

prevalence of written medium and English as a target language is utterly visible when we observe Dumont and Granger's list of learner corpora around the world[2] (table 1). From the 146 corpora displayed in the list, only 43 are uniquely spoken and other 14 are composed of written and spoken language. As for the target language, more than half of the listed corpora are compiled with English as a second language.

Among the most frequent target languages, Spanish, French and German come after English on the same list. There are 16, 15 and 13 corpora for each of these languages, respectively. These corpora are not as big as the English one, but they also offer interesting features. The French Interlanguage Database (FRIDA), for instance, is a corpus of French as a foreign language which counts with an error-tagging system especially developed for French interlanguage (GRANGER, 2003).

A prominent example of spoken corpora is The Louvain International Database of Spoken English Interlanguage (LINDSEI), which counts with 130 hours of recording and has been widely used for research purposes (GILQUIN; DE COCK; GRANGER, 2010). For instance, Aijmer (2011) identifies an overuse of *well* in the Swedish subcorpus of LINDSEI and stresses the need of discussing pragmatic markers in learning environments, while Brand and Götz (2011) discuss the correlation between fluency and temporal variables in spoken learner language and observe that some lexical and grammatical categories, such as tense agreement, are especially error-prone.

It is also worth to discuss the so-called commercial corpora. These corpora are created and used for commercial purposes, such as textbook and dictionary design. Some examples are the Longman Learner Corpus (LLC) and the Cambridge

[2] The list constantly updated by Amandine Dumont and Sylviane Granger from The Centre for English Corpus Linguistics of Université catholique de Louvain and available at https://www.uclouvain.be/en-cecl-lcworld.html Last accessed on February 16th, 2016

University Press Learner Corpus (CLC). These commercial corpora have the advantage of counting with very large datasets, most usually compiled with the production of test-takers of proficiency exams such as the *Test of English as a Foreign Language* (TOEFL) and the *International English Language Testing System* (IELTS). The CLC, for instance, has approximately 55 million words of English taken from exams scripts written by learners from 203 different nations and 138 L1s at all levels of proficiency (HAWKINS; BUTTERY, 2010).

Conversely, the non-commercial learner corpora exhibit a more moderate number of words. There are a few of these corpora that also count with a great number of words, such as the Hong Kong University of Science and Technology (HKUST) Learner Corpus, which has about 25 million words (PRAVEC, 2002). However, the data from these big corpora usually come from speakers of the same mother tongue, as it is the case with HKUST corpus. In this corpus, all the contributors share the same first language (L1) background (Chinese); being Cantonese the most frequent one.

Probably the largest and non-commercial learner corpus with participants from different L1 background is the International Corpus of Learner English (ICLE) (NESSELHAUF, 2004). The corpus, which has a total of 4,251,714 words, is divided in 16 sub-corpora, each one featuring the written production (mainly argumentative essays) of learners from different L1 backgrounds (GRANGER et al., 2009). One of the sub-corpora to be included in the next version of ICLE is its Brazilian sub-section (Br-ICLE)[3], a university level corpus of written production. With 200,000 words, the corpus was compiled with argumentative essays written by English major students at several universities in the country.

[3] The compilation of this subcorpus was done by Tony Berber Sardinha (PUC-SP)

Although several researchers have pointed that a high number of words is crucial for corpus studies (e.g. SINCLAIR, 1991), other aspects such as the variability of L1 should also be noted. The next section addresses how learner corpora are designed and presents the most commonly used criteria.

Table 1 - examples of learner corpora

Corpus	**L1**	**Size in words**	**Medium**	**Institution**
The Hong Kong University of Science & Technology learner corpus (HKUST)	Chinese (mostly Cantonese)	≈ 25,000,000	written	Hong Kong University of Science & Technology, Hong Kong (John Milton)
The Uppsala Student English Corpus (USE)	Swedish	≈ 1,200,000	written	Uppsala University, Sweden (Ylva Prytz and Margareta Axelsson)
The International Corpus of Learner English (ICLE)	14 different L1 backgrounds	≈ 3,000,000	written	Centre for English Corpus Linguistics Université catholique de Louvain (Sylviane Granger)
The Cambridge Learner Corpus (CLC)	various	≈ 50,000,000	written	Cambridge University Press and Cambridge ESOL, UK (Commercial)
The Longman Learners' Corpus	160 different L1 backgrounds	≈ 10,000,000	written	Longman (Commercial)
The Louvain International Database of Spoken English Interlanguage (LINDSEI)	11 different L1 backgrounds	≈ 800,000	spoken	Centre for English Corpus Linguistics Université catholique de Louvain (Gaëtanelle Gilquin and Fanny Meunier)
French Interlanguage Database (FRIDA)	various	≈ 200,000	written (French)	Centre for English Corpus Linguistics Université catholique de Louvain (Sylviane Granger)

Source: table adapted from: https://www.uclouvain.be/en-cecl-lcworld.html

2.2 Learner corpus applications

Learner corpora, as well as native speaker corpora, are commonly applied to second language education in two different forms. One first application of this kind of corpus is to describe learners' language and to aid the identification of its prominent aspects. A second and more elaborated form would be to inform Second Language Acquisition (SLA) through, for example, the study of developmental sequences. A third application, which has not been fully explored yet, would be to use learner corpora directly in classroom (NESSELHAUF, 2004).

A different division of learner corpus application is drawn by Granger (2009), which describes the pedagogical application as divided in two groups: the Delayed Pedagogical Use (DPU) and the Immediate Pedagogical Use (IPU). The former is the most commonly used approach and it is used as a resource for generic material design, such as dictionaries, grammar books and textbooks. The DPU corpora are generally compiled by research groups, and the learners from whom the data is collected are not, in most cases, immediately beneficiaries of such data-based activities. These corpora, which tend to be bigger than the IPU ones, are usually collected within a specific context and then applied to similar groups, where learners show a similar profile such as first language (L1) background, age and social group.

Regarding the IPU corpora, they are usually collected by teachers in their own teaching environment and the learners do not only produce the collected data, but are also users of the corpus. The corpus may also be later used within other similar context. However, due to the considerable small size of these corpora, it may not be representative of a very comprehensive population (RAGAN, 2001). Nonetheless, the inferences made from IPU corpora can be more relevant for a given group. In addition, allowing learners to identify their own inadequacies is effective in raising language awareness and autonomy (MUKHERJEE; ROHRBACH, 2006).

Considering the aforementioned application, the subsections that follow will discuss the relation between SLA and Learner Corpus (LC); some studies on interlanguage description based on corpus research; and, finally, some more practical pedagogical applications of LC.

2.2.1 Second Language Acquisition

Although research on Second Language Acquisition and Learner Corpus both deal with learner interlanguage, IL[4] (SELINKER, 1972), there is still no full agreement between these two areas. One reason for the gap between the two approaches can be given by the different goals they have. Callies and Paquot (2015), for example, point to the fact that researchers in SLA generally seek to identify whether second language (L2) learners permit certain types of construction, while LC researchers focus on the constructions made by learners themselves. Another important justification refers to the distinct knowledge that each area presents. Granger (2009) states that researchers in SLA ignore the potential of corpus tools, and researchers in LC, conversely, do not have a deeper knowledge of SLA. She stresses, in this way, the need of shared knowledge between these two fields in order to increase their interaction.

Nevertheless, studies have shown the relevance of LC studies for SLA. Tono (1998) and Meunier (2010), for instance, present some possible contributions LC can make to SLA studies. They are: description of the developmental stages of the interlanguage; studies on the effects of L1 transfer; identification of overuse and underuse of language resources; identification of the differences between universal and specific errors of each L1; and distinction between the production of native speakers and learners. Such applications can be seen in studies such as Cobb (2003), which

[4] Learners' production in a second language

highlights the influence of L1 on the acquisition of L2, and Hasko (2013), which investigates the developmental process of L2.

2.2.2 Language Description

Despite its importance in different fields, learner corpora have been mainly used in academic contexts, more specifically in describing and investigating English written productions. This trend may be due to the fact that such kind of production is more easily collected than the oral one. Furthermore, there are more English corpora available that can be used as reference when comparisons are to be made, than of any other target language.

A common method of analysis that has been widely used when learner corpora is concerned is the Contrastive Interlanguage Analysis (CIA). In this method of analysis, native language and learner language – sometimes learner language from two distinct L1 backgrounds – are compared quantitatively and qualitatively (GRANGER, 1996). It is worth emphasizing that the choice of the native corpus, also known as reference corpus, should be made cautiously, once it may influence negatively the analysis (GILQUIN; PAQUOT, 2007).

Through CIA or any other method, the use of learner corpora has provided research on L2 with a strong empirical basis in large datasets. Such amount of texts allows learners' production to be seen in many different ways, with great focus being given to the lexicon, phraseology and genre variety (GRANGER, 2009). Several studies have investigated these linguistic features of learners' production. Cumming et al. (2006), for instance, showed differences in the discourse according to the task type and the level of proficiency. Some of these characteristics are the length of the response, lexical diversity, clause length, and grammatical accuracy. Considering the lexical features it is worth to highlight Grant and Ginther´s (2000) study, whose findings demonstrate that the more proficient students are, the higher the lexical

specificity will be. In terms of grammatical and syntactic features, Jarvis et al. (2003) demonstrated, by using cluster analysis, that it is possible to notice differences in the use of linguistic features across different proficiency levels when they are considered together, rather than as individual units. When phraseology is concerned, several studies have focused on the use of formulaic language by learners. Dutra and Berber Sardinha (2013), for instance, analysed the written production of Brazilian learners of English and identified a need to enhance the learners' use of referential expressions in their production.

2.2.3 Immediate Pedagogical Application

As mentioned earlier, the study of corpora have been of great use for pedagogical purposes. Such applications have been noticed mainly when material design is concerned. For instance, back in 1987 the Collins COBUILD English Language Dictionary was published and six years later the Longman Learner Corpus was used to compile the first dictionary based on learner corpus analysis, the Longman Language Activator (1993). However, there has been a growing interest in applying learner corpus data immediately with the corpus participants.

Seidlhofer (2002), for instance, argues that learning-driven data (LDD)[5] can be extremely beneficial for language pedagogy, once it provide learners with language awareness, autonomy and authenticity. The author also adds that pedagogy should be designed to address specific needs of a given setting, which can be achieved with the use of local corpora.

A later study that also advocates for the use of own local learner corpora by teachers and learners is done by Mukherjee and Rohrbach (2006). The authors share the same

[5] the term was coined by Seidlhofer (2002) to refer to an approach which uses learner corpora for language teaching purposes

perspective of Seidflhofer (2002) that there is a need for more classroom-based corpus-linguistic action research and they also claim that the focus should be first on teachers developing activities based on the local corpus and only then the exploration should be taken by the learners themselves.

A more recent study has compared the benefits of using data-driven learning (DDL) within two groups - one group accessed only a native-speaker corpus while the other group relied on the combination of native-speaker and learner corpora (COTOS, 2014). The study demonstrated that the group that used both corpora showed better results than the one that only used the native speaker corpus. The author results corroborate with other studies (e.g. BERNARDINI, 2002; FLOWERDEW, 2012) which demonstrate that DDL has several benefits such as to expose learners to real language, to make them aware of a wider diversity of syntactic structure and to draw attention to form, meaning and function. Cotos' (2014) study has also demonstrated that the use of LDD has the extra advantage of making learners more cognitively involved, once their own data is used. This higher involvement of learners, as shown in the results, has increased learning drive and has also facilitated learning, as students internalized the new learnt structure better.

2.3 Design Criteria

As presented in section 2.1, learner corpora have commonly been defined as "systematic computerized collections of texts produced by learners" (NESSELHAUF, 2007, p. 40). They may be collected for a specific study or for a more general and broader use (NESSELHAUF, 2004). This author adds that this "systematic" collection should be based on defined criteria, mostly external, which can then lead to various analysis on language learning, such as L1 and gender influence. Furthermore, as Sinclair (1991, p. 9) has pointed out, 'the results are only as good as the corpus'. Consequently, corpus design should have very clearly established criteria. Ellis (1994)

and Granger (1998) divide the variables of a learner corpus in two groups, one related to the language and another to the learner situation (table 2). Granger (1998) highlights that topic choice is rather important for its directly influence on lexical choice, while the overall learner output is considerably affected by the genre choice. As for the learner variable, Granger emphasizes that first language background is crucial for corpus analysis, once it varies considerably according to different mother tongues.

Table 2 - Language and learner variables

Language	**Learner**
Medium	Age
Genre	Sex
Topic	Mother tongue
Technicality	Region
Task setting	Other foreign languages
	Level
	Learning context
	Practical experience

Source: GRANGER, S. **Learner English on Computer**. London: Longman, 1998. p. 8

Although there is a common agreement of which variables should be considered in the compilation of learner corpora, Granger stresses that the design adopted can vary according to the study, as long as they are clearly set and yield "soundly based conclusions, making it not only possible but indeed legitimate to make comparisons between different studies" (ENGWALL, 1994, p. 49 apud GRANGER, 1998).

The already mentioned corpus ICLE illustrates how these criteria can be usually set. In its description, ICLE variables are divided into task and learner variables (table 3). Some of these variables are shared across all the texts, as it is the case of genre (academic essay) and learning context (undergraduate students of English), while other variables differ within subcorpora (e.g. gender) or across subcorpora (e.g. mother tongue) (GRANGER et al., 2009).

Table 3 - ICLE task and learner variable

Task Variables	**Learner Variables**
medium *writing*	age *young adults*
genre *academic essay (mainly argumentative)*	gender *76% are female*
field *general English (rather than ESP)*	mother tongue *16 different L1*
length *500 - 1,000 words*	region
topic *list of suggested topics*	other FLs
task setting *coordinators decide (timing, reference tool, part of an exam)*	stay in English-speaking country *no / 3 months or less / 3 months or more*
	proficiency level *advanced (C1/C2)*
	learning context *undergraduate students / EFL*

Source: adapted from Granger et al. (2009)

When the design criteria of a commercial to a non-commercial corpus are compared, not much difference is observed. Nevertheless, it is worth pointing out that the former type of corpus is usually accompanied of extra information obtained from the proficiency exams from which the data is drawn. The data in CLC, for instance, is derived from the many Cambridge proficiency exams[6], granting access to candidates' proficiency in English.

Although the further procedures to which a corpus goes through, such as annotation and lemmatization, are not necessarily part of the corpus design, they should also be taken into account. ICLE's essays, for instance, were lemmatized and POS tagged

[6] CPE, CAE, FCE, PET, KET (general purpose) and BEC Higher, BEC Vantage, BEC Preliminary (business English)

with the non-open source software CLAWS (GARSIDE; SMITH, 1997). The corpus digital version also has a built-in concordancer and user-friendly search interface which allows the user to search for word forms, lemmas, multiword units, part-of-speech tags and regular expressions (GRANGER et al. 2009). The next chapter will address some of the corpus linguistic methods and tools used for language analysis.

3. CORPUS LINGUISTICS: EARLIER AND CURRENT METHODS

The previous chapter has addressed the importance and the use of learner corpora. The data obtained from learners of a second language can indeed be advantageous for linguistic studies. However, without the necessary tools for data analysis, this process can be extremely time-consuming and error prone. For this reason, there are several tools which handle corpus and datasets in a faster and easier way. This chapter will, therefore, present the most commonly used corpus analytic techniques.

The chapter is structured as it follows. Firstly, a brief description of the approach to corpus taken by neo-firthian and functionalist linguists will be presented. In the second section, some of the most frequently used techniques to extract information from a corpus will be described. The third and final section will address the tools used for corpus exploration.

3.1 Neo-Firthian and Functionalist Linguistics

The linguistic features observed and extracted from a corpus may vary according to the field and needs of the study. While there is a distinction between corpus-as-theory and corpus-as-method, the approach to a corpus also varies according to theoretical linguistic background. McEnery and Hardie (2012) discuss how neo-Firthian scholars and functionalist linguists interact with corpus linguistics.

The neo-Firthian scholars are said to follow J. R. Firth language approach, which was incorporated to corpus linguistics by John Sinclair. Perhaps the term most associated to these scholars is ‘collocation’. These scholars have contributed immensely to an empirical research on collocation and to the way lexis and grammar are seen as interrelated. Although there are several different definitions, there is an agreement that ‘collocation’ denotes that meaning is conveyed not in isolation, but in association with the words that they co-occur (MCENERY; HARDIE, 2012). A

collocation is, therefore, a pattern of two items that co-occur in proximity, but not necessarily adjacently. Sinclair (1991, 2004) call these items node and collocate, being the first the unit under examination and the latter the units found in a given span.

McEnery and Hardie (2012) present two techniques used by neo-Firthians to study collocations: collocation-via-concordance and collocation-via-significance. The former, also known as "hand and eye" technique, consists of a manually scanning, and, although it is still currently adopted, dates from earlier studies as Sinclair (1991) and Stubbs (1995). The latter approach relies on statistical techniques, such as chi-squared, log-likelihood, t-score, z-score and mutual information. In both approaches frequency and language use are essential when an argument is built, and the use of statistics enhances the analysis.

Two other terms frequently used by neo-Firthians are 'discourse' and 'semantic prosodic'. Despite the multiple definitions for 'discourse', for the neo-Firthians, the term is associated with the structure of a text itself, rather than any political or social meaning, as it is the case in Critical Discourse Analysis[7]. What is aimed by neo-Firthians researchers is the understanding of how a sentence relates to its neighbours to build cohesion and coherence. The second term, 'semantic prosody' (e.g. SINCLAIR, 1991; LOUW, 1993; PARTINGTON, 1998), refers to the positive or negative meaning of a word or phrase has according to the unit it co-occurs. For instance, the word 'happen' alone does not convey a negative or positive meaning, but, since most of its frequent collocates carry a negative meaning, the word is said to have a negative semantic prosody to this word (MC ENERY; HARDIE, 2012).

[7] "CDA typically studies how context features (such as the properties of language users of powerful groups) influence the ways members of dominated groups define the communicative situation in "preferred context models" (VAN DIJK, 2003, p. 358)

As for the functionalist linguists, McEnery and Hardie (2012) stress that the emphasis that these scholars give to language in use makes their research compatible to corpus linguistic techniques. The authors add that functionalists have been relying more and more on corpus tools. Conversely, corpus linguists as Douglas Biber and Stefan Gries have relied on functionalist theoretical framework to develop new corpus investigation techniques. For instance, Biber's (1992) multi-dimensional approach to text type variation aims at identifying functional explanation for grammatical variations, which is a similar concern functionalists have. The following section will address some of these investigation techniques in more detail.

3.2 Investigation Techniques

Several studies have investigated linguistic features of learners' production. Cumming et al. (2006), for instance, showed differences in the discourse according to the task type and the level of proficiency. Some of these characteristics are the length of the response, lexical diversity, clause length, and grammatical accuracy. Considering the lexical features it is worth to highlight Grant and Ginther's (2000) study, whose findings demonstrate that the more proficient the students are, the higher the lexical specificity (type/token ratio and average word length) will be.

In terms of grammatical and syntactic features, Jarvis et al. (2003) demonstrated, by using cluster analysis, that it is possible to notice differences in the use of linguistic features across different proficiency levels when they are considered together, rather than as individual units. Several studies have also focused on the use of formulaic language. Cortes (2004) and Hyland (2008), for instance, approach the use of lexical bundles[8] and academic written texts. The former demonstrates that the use of

[8] "simply sequences of word forms that commonly fo together in natural discourse" (BIBER et al., 1999, p. 990)

bundles by learners and professional writers differ significantly, while the latter discusses how their usages differ by discipline. In the subsection that follows, three main possible analyses with corpus tools will be addressed.

3.2.1 Word frequency list

In spite of its simplicity, this basic analysis is a helpful tool for researchers, from lexicographers to material and language syllabus designers, since a word frequency list is the basis for other corpus linguistic analysis. For instance, when used with statistical tests such as the Mutual Information (MI), t-score and z-score, they are the source of information to identify collocations. These frequency lists can also be compared among different corpora and, when a larger corpus is used as a 'reference' or 'benchmark', keywords can be identified. This technique is useful for verifying whether the frequency of a word in a given corpus matches the expectation. Keywords can be either positive, when they are frequent, or negative, when they are infrequent in the corpus being analysed (EVISON, 2011).

One main concern with this list, however, is related to the definition of 'word'. Evison (2011) highlights two main issues: units of words and lemmatization. The first refers to units such as *I'll* and whether they should be counted as one or two words. The second issue is whether words such as *smile* and *smiled* should be taken as one or as distinct units. One common definition of word, usually adopted by computer programmers and adopted in this study is "a sequence of alphabetic (or alphanumeric) characters uninterrupted by whitespace (i.e. spaces, tabs, and newlines)" (GRIES, 2009, p. 13). Following this definition, "smile" and "smiled" are counted as different words. Gries (2009) also stresses that it is necessary to distinguish types (which words are present) from tokens (how many words are present).

3.2.2 Part-of-speech taggers

As the name says, part-of-speech (POS) tagging refers to the assignment of POS label to each token in a corpus. The list of POS labels may vary according to the used tagger. For instance, the tagger CLAWS7[9] list contains 137 different tags, while the Apache OpenNLP[10] (Appendix C) counts with only 36 labels.

The use of part-of-speech tool, according to Anthony (2013), is somehow debatable. Some of the corpus-driven approach followers argue that the annotation may contaminate the original data and prevent the researchers from observing new linguistic patterns. Sinclair (2004), for instance, argues that

> "The interspersing of tags in a language texts is a perilous activity, because the text thereby loses its integrity, and no matter how careful one is the original text cannot be retrieved (...) In corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly and then the patterns of this uncontaminated text are able to be observed." (SINCLAIR, 2004, p. 191)

However, the most recent corpus tools allow the tagging to be visible or omitted, which might make this debate irrelevant. Instead, a growing number of corpus annotation supporters can count on free and easy accessible taggers, as it is the case with CLAWS7[11], which allows text fragments (up to 100,000 words) to be annotated online.

POS tagging has proved to be effective to identify features that would not be immediately observable in a raw corpus. By adding another layer of information to the corpus, new features of the object being analysed can be revealed (ANTHONY, 2013). For instance, Biber et al. (1999), when discussing complex noun phrases

---

[9] http://ucrel.lancs.ac.uk/claws/

[10] https://opennlp.apache.org/documentation/manual/opennlp.html#tools.postagger

[11] http://ucrel.lancs.ac.uk/claws/trial.html

identify that academic language is packed with nouns, differently from other registers. Without the POS annotation, this type of analysis would not be possible.

3.2.3 n-grams

The last investigation technique here described is the n-grams. Differently from the collocations, which are words that co-occur in proximity, the n-grams are an adjacent sequence of two or more words. The term, which is commonly adopted in computational linguistics, refers to a simply sequence of tokens, or "a sequence of N words" (JURAFSKY; MARTIN, 2008, p. 94). Other terms as multi-word units, cluster and lexical bundles have also been used, and the choice for a term or another depends on the assumptions made in the study being carried. For instance, Biber et al. (1999, p. 990) define lexical bundles as "simply sequences of word forms that commonly go together in natural discourse" and add that they "can be regarded as lexical building blocks that tend to be used frequently by different speakers in different situations".

Regardless the term adopted by researchers and the assumptions they carry, the knowledge of formulaic sequences has proved to be extremely significant for language fluency processing (ROBINSON; ELLIS, 2008). They have been used in several studies on academic language description (e.g. BIBER; CONRAD; CORTES, 2004; HYLAND, 2008; SIMPSON-VLACH; ELLIS, 2010) and second language acquisition (STAPLES et al., 2013, VINCENT, 2013). However, their extraction can be somehow problematic when working with learner corpora. These types of corpora are usually created with texts written by learners after a given topic, which leads to a large number of topic-related sequences of words that can inflate the final results (BOHÓRQUEZ, 2015). Nonetheless, keeping the topic-related grams may also shed some light on how learners of a second language deal with assignment prompts.

Regarding the process of n-grams extractions, Gries (2009) emphasizes that it is worth knowing which strategies are employed by the chosen method. Gries alerts that most ready-made programs (e.g. kfNgram[12], MLCT concordancer[13] and Collocate 0.5.4) generate n-grams list without considering end of sentences and punctuation. This means that the last word of a sentence will be associated to the first word of the following one forming a 'false' n-gram. Since this kind of n-grams might be undesirable and problematic, it may be worth studying the corpus tools which better fit one's research purpose. Some of these corpus tools will be discussed in the next section.

3. 3 Corpus Linguistic Tools

Although term 'corpus' is commonly misused as being the data itself and the tool used for its analysis, the difference between corpus, database, and corpus tools should be outlined. Corpus may be defined as

> "a machine-readable collection of (spoken or written) texts that were produced in a natural communicative setting, and the collection of texts is compiled with the intention (1) to be representative and balanced with respect to a particular linguistic variety or register or genre and (2) to be analyzed linguistically." (GRIES, 2009, p.7)

A database, in turn, is simply an organised collection of data, which is not necessarily compiled to represent any linguistic event. As for the corpus tool, it is the software, environment or any other tool with which a corpus is analysed. As Hunston (2002, p. 20) points out, "a corpus by itself can do nothing at all, being nothing more than a store of used language". Likewise, a corpus tool alone cannot process anything. This

[12] http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

[13] https://sites.google.com/site/scottpiaosite/software/mlct

section covers some of the tools most used by researchers to observe and handle language data.

3.3.1 Four generations of concordancers

The constant evolution that corpus tools have been going through has made them more accessible to the general public. Nowadays, researchers from the most varied fields rely on corpus tools to carry their studies. These tools can now be easily installed on a personal computer and they usually present a considerably user-friendly interface.

Nonetheless, to reach this current usability, corpus tools have been evolving into more powerful and accessible versions. Regarding more specifically the concordances, McEnery and Hardie (2012) divide them into four generations, as described in table 4. Although the first generation of concordances could not do more than produce key word in context (KWIC)[14], the authors attribute to them the importance of developing concepts which still serve as foundation for the current corpus tools. One example of a first-generation concordance is CLOC (REED, 1978), which was used at The University of Birmingham, where the COBUILD project was held.

As regards the second generation of concordancers, the authors claim that there was not much difference in terms of functions. The concordancers were still limited to providing KWIC, with the addition that now word lists could be generated and the concordance lines could be sorted alphabetically according to the context to the left or right. One benefit that came with the second generation, though, was the democratising effect of the new tools. This second generation rose in the 1980s and 1990s, when personal computers became accessible for the ordinary users. For this

[14] A KWIC is probably the most common format for concordance lines and it is formed by sorting and aligning the word(s) that is being analysed in a given corpus.

reason, researchers could work on small-scale studies with the aid of concordances such as Kaye Concordancer (KAYE, 1990) and Micro-OCP (HOCKEY, 1988).

The concordancers classified by McEnery and Hardie (2012) as third-generation ones encompass the tools that are currently most used by researchers. Some examples are *WordSmith Tools* (SCOTT, 2016) and *AntConc* (ANTHONY, 2014). These tools came with the benefits of being able to process large datasets and to support a wide range of writing systems. There are also several functions that they can perform, such as extraction of collocations and n-grams, and statistical analysis. Despite the many improvements the third generation concordancers show, the authors claim that there are still some limitations in these tools. According to them, there are still a considerable number of techniques for corpus analysis which have been developed but not incorporated to the concordancers yet, as it is the case with the collostructions analysis, which measures the level of repulsion or attraction of a word to any syntactic string. (STEFANOWITSCH; GRIES, 2003). Another weakness is that some the softwares need to be installed to the computer. This can be problematic for two reasons: the software may not run in all operating systems, as it is the case with *WordSmith Tools*, and the processing may be slow, depending on the computer processor.

Differently from the previous generation, the fourth generation tools are web-based and, therefore, neither require local processing of the data nor a specific operating system. Added to this, the web-search interface also comes with the advantage of being more user-friendly and ready for immediate use. Some examples of fourth-generation concordancers are CQPweb (HARDIE, 2012), SketchEngine (KILGARIFF, 2013), and corpus.byu.edu (DAVIES, 2004-). These concordancers have also partially solved the problem of copyright. Corpora which would otherwise have its distribution restricted to licensed users, can now be accessed through these online tools, since a concordance line does not contain more words than what is

considered 'fair use' under copyright law[15]. While this makes more corpora be accessible to more people, it also has the drawback of not allowing full access to the texts.

Table 4 - Four generations of corpus tools

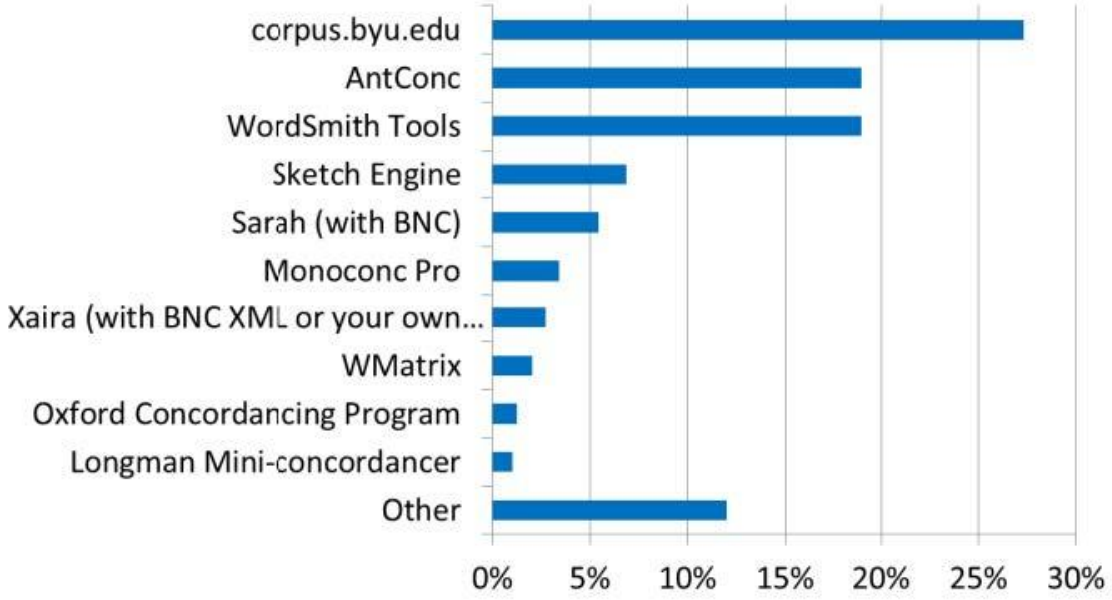
G	examples	functions	benefits	limitations
1st	CLOC (REED, 1978)	KWIC	- term development - replicability	could process only Roman alphabet
2nd	- Kaye Concordancer (KAYE, 1990) - Longman Mini-concordancer (CHANDLER, 1989) - Micro-OCP (HOCKEY, 1988)	- KWIC - word (types and tokens) list - sort concordances alphabetically	democratising effect	could not process many words (few tens of thousands)
3rd	- WordSmith Tools (SCOTT, 2016) - AntConc (ANTHONY, 2014) - MonoConc (BARLOW, 2000)	- KWIC - frequency lists - n-grams - collocations - statistical analysis	support large datasets and a wide range of writing system	- recent data analytics have not been incorporated yet - some of them run only in Windows
4th	- CQPweb (HARDIE, 2012) - SketchEngine (KILGARIFF, 2013) - corpus.byu.edu (DAVIES, 2004-)	- KWIC - frequency lists - collocations - statistical analysis	- installation is not required - faster processing - ready-made corpora	restricted access to the available corpora

Source: Table based on McEnery and Hardie (2012)

[15] Fair use is a legal doctrine that promotes freedom of expression by permitting the unlicensed use of copyright-protected works in certain circumstances. Section 107 of the Copyright Act provides the statutory framework for determining whether something is a fair use and identifies certain types of uses—such as criticism, comment, news reporting, teaching, scholarship, and research—as examples of activities that may qualify as fair use. (http://copyright.gov/fair-use/more-info.html)

The third and fourth generation tools are the currently most used ones and the advantages of one over the other vary according to the need of the research. A survey carried by Tribble (2012) has shown that corpus.byu.edu, AntConc and WordSmith Tools are the favourite ones (figure 1) among researchers, and it has also revealed that users are becoming more demanding and interested in software development and coding. The section that follows will, therefore, discuss the use of programming language such as Python and R for language processing and text mining.

Figure 1 - Survey results in response to the question "Which computer programs do you use for analysing corpora?"



Source: TRIBBLE, C. Teaching and language corpora: quo vadis? In: TEACHING AND LANGUAGE CORPORA CONFERENCE (TALC), 10., Warsaw. Anais 2012.

3.3.2 Programming languages for corpus analysis

The previous section has shown some concordance softwares and online query system for corpus analysis. Although these tools are considered to be more user friendly than relying on programming language, a growing number of linguists interested in developing their own tools have been noticed.

Several researchers have been encouraging linguists for a long time now to acquire some basic programming skills. Biber, Conrad and Reppend (1998) claim that

> "concordancing packages are very constrained with respect to the kinds of analyses they can do, the type of output they give, and, in many cases, even the size of the corpus that can be analyzed (…) Computers are capable of much more complex and varied analyses than these packages allow, but to take full advantage of a computer's capability, a researcher needs to know how to write programs." (BIBER; CONRAD; REPPEN, 1998, p. 254)

The authors also argue that with the aid of computer programming, the analysis can be adapted to suit specific research needs. In addition, large data can be dealt with and, when compared to a concordancer software, the analysis is done faster and more accurately.

Stefan Gries (2009), another enthusiast of programming languages for corpus analysis, presents some reasons why linguists should engage on learning a programming language. First, he argues that learning and using a programming language are not as time consuming as it may appear. According to him, once the first scripts are developed and some basic abilities are mastered, it will be possible to reuse the scripts, which makes the process as fast as using a concordancing software. In addition, the processing time required is considerably lower than the time required by concordance tools such as AntConc.

A second reason presented by Gries is the fact that the final users are the ones in control and do not depend on the software developer. Therefore, they can make decisions to fit a specific goal, such as defining what a word is. Gries presents as the third advantage the fact that there are open source programming language, such as R and Python, which makes the packages transparent and continually updated by users from the entire world. Not only that, the scripts developed by one researcher can be easily reproduced in other studies. The fourth point indicated by Gries, is related to the numerous tasks that can be done in the same environment, in contrast to the

tasks carried by concordance tools. For instance, R allows the performance of statistical evaluation, annotation, data retrieval, graphical representation and data processing using only its own environment.

Although R has been the programming language of choice for linguists such as Gries (2009, 2013) and Baayen (2008), there are several other options for data analysis. Some examples are Java, Perl, Python and R. Each programming language has its benefits and disadvantages, and its choice depends on the researcher main goal. Java, for instance, is a very sophisticated language and allows complex operations, yet it can also be intimidating for novice programmers. Perl was used in the development of earlier concordancers, as it was the case with the first versions of AntConc (ANTHONY, 2002). The most recent version of AntConc, however, was developed in Python, which makes processing faster and allows more advanced statistic modules (ANTHONY, 2013). Python has become increasingly popular in the linguistic community, once it has an easy to understand syntax and has mainstream open source software libraries like the Natural Language Toolkit (NLTK)[16]. Likewise, R (R Core Team, 2014) is among the most popular programming languages and counts with several packages for natural language processing. Some of these packages and main features of R will be covered in the following section.

3.4 Text mining in R

Differently from Python, which is a general-purpose programming language, R is a specialized statistical language or, more specifically, "a free software environment for statistical computing and graphics"[17]. Among the various reasons to use R for linguistic studies, Baayen (2008) claims that

---

[16] http://www.nltk.org

[17] definition available at https://www.r-project.org

> "(...) it is a carefully designed programming environment that allows you, in a very flexible way, to write your own code, or modify existing code, to tailor R to your specific needs." (BAAYEN, 2008, p.xi)

Baayen emphasizes that this flexibility is made possible due to R's elegant and consistent environment, which makes the data management easier. Not only that, R has the advantage of being freely available under the GNU General Public License[18] and counting with great community support through *The Comprehensive R Archive Network* (CRAN)[19]. This network makes a vast number of libraries and packages available for numerous applications. When filtering the current packages on CRAN website[20], for instance, 36 packages are listed, not to mention the immense variety of graphical facilities that R offers for exploratory data analysis (EDA)[21]. However, a text mining analysis demands more than analytical techniques, as section 3.4.1 will demonstrate.

3.4.1 Text mining analysis

When text mining is concerned, one may first imagine standard analysis techniques, such as clustering and classification. However, there are several other steps that should be taken before these final analyses. Feinerer, Hornik and Meyer (2008) describe in four steps the whole process that a text mining analyst has to go through. Firstly, it is necessary to import the texts which will be worked with to a computing environment, such as R. These texts often come from a highly heterogeneous input which makes an organization necessary. After organizing and structuring the texts, the second step is to tidy up the data so that it has a more appropriate

---

[18] http://www.gnu.org/licenses/gpl.html

[19] https://cran.r-project.org

[20] https://cran.r-project.org/web/views/NaturalLanguageProcessing.html (accessed on Feb 11th, 2016)

[21] a method which allows datasets to be analysed and summarised mainly through data visualization

representation. Some of these pre-processing include stopwords and whitespace removal and stemming procedures. On a third step, the preprocessed data has to be transformed into structured formats. According to the authors, the "classical" text mining format adopted is the term-document matrix, which is a matrix with the frequency with which terms occurs in a collection of texts. Only after all these three steps are taken, the analyst will be able to work with the data analysis itself, generating frequency lists, identifying collocations and so forth.

There are several packages available for R which can assist this whole process. The *tm package* (FEINERER; HORNIK; MEYER, 2008)[22], for instance, provides methods for text mining in R, such as data import, corpus handling, preprocessing and metadata management. Some of the transformations that can be done with *tm* are to eliminate extra whitespace, punctuation and specific words. The package also offers statistical analysis, such as the function *findAssocs*, which returns all the words that are associated to a term, given a minimal correlation.

Another library that is useful when dealing with a corpus is the Apache OpenNLP[23] library (HORNIK, 2015), which is a toolkit for Natural Language Processing (NLP) written in Java. Some of the tasks that this library supports are tokenization, sentence segmentation, part-of-speech tagging and chunking. Also written in Java, the library RWeka (HORNIK; BUCHTA; ZEILEIS, 2009) offers a wide range of tools for NLP, from data pre-processing to data visualization.

3.4.2 Tidy data

As it was mentioned in the previous section, the data has to be pre-processed before conducting the analysis. In fact, when dealing with data analysis, a great part of the

[22] https://cran.r-project.org/web/packages/tm/index.html
[23] https://cran.r-project.org/web/packages/openNLP/index.html

time is spent on data preparation and cleaning (e.g. DASU; JOHNSON, 2003; WICKHAM, 2014). For this reason, a main concern when dealing with data analysis is that the data is stored as tidy data, i.e.,

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table. (WICKHAM, 2014, p. 4)

When this framework is adopted, the datasets are more easily manipulated, modelled and visualised, once a smaller set of tool is necessary to deal with the data (WICKHAM, 2014). Some of the tools that deal with tidy data in R are the packages *dplyr* (WICKHAM, FRANCOIS, 20115), *plyr* (WICKHAM, 2011), and *data.table* (DOWLE et al., 2014).

The *data.table*[24] package is commonly used to work with data frames. This powerful and useful tool offers functions such as subsetting, ordering, and grouping data. However, the use of functions may not be very intuitive for linguistics with little knowledge on programming languages. More suitable package for simple operations would than be the *dplyr*[25] and the *plyr* package[26]. The *dplyr* package provides the user with functions which are easily interpreted and efficient.

The combination of the right tools, such as the ones mentioned above and of a structured dataset, can considerably reduce the time spent on the data pre-processing. Furthermore, a tidy dataset is more easily accessible by other user who might be interested in working with the same data. The next chapter will present a detailed description of the method used in this study, including all the steps, from pre-processing to the final data analysis.

---

[24] https://cran.r-project.org/web/packages/data.table/index.html
[25] https://cran.r-project.org/web/packages/dplyr/
[26] https://cran.r-project.org/web/packages/plyr/plyr.pdf

4 MATERIALS AND METHODS

This chapter focuses on the methodological and technical aspects involved in the construction, processing and analysis of the CorIsF dataset, a dataset retrieved from evaluative activities of *Idiomas sem Fronteiras* courses[27]. The procedures that are here described require a certain familiarization with the programming language R. Although this requirement may discourage less computer-savvy readers, it is worth reminding that one of the objectives of this study is to develop replicable scripts so that other language researchers can handily access the CorIsF dataset or similar ones.

The procedure was divided in four parts:

1. dataset compilation and pre-processing;
2. dataset processing;
3. extraction of the key features;
4. and data visualization.

Section 4.1 deals with the method utilized to collect the data and to do the first cleaning process, such as eliminating unwanted data and keeping the relevant ones. The following section presents the procedures adopted to subset CorIsF in small corpora and to annotate it with a part-of-speech (POS) tagger. The third section describes the process of identifying the variability of POS, the frequency of types and tokens, and the n-grams. The final section presents scripts that easily generate data visualization for the aspects observed in section 4.3.

[27] The English without Borders Program (*Inglês sem Fronteiras – IsF*) was launched on December 18th, 2012. In November 2014, the Program expanded to Language without Borders (*Idiomas sem Fronteiras*) which aims at promoting the learning of French, Spanish, German, Italian, English, Japanese, Mandarin and Portuguese for foreigners.

4.1 Data Compilation and Pre-processing

4.1.1 Data Collection

The dataset has been compiled with the written production of students enrolled in the Brazilian programme Language without Borders - English (*Idiomas sem Fronteiras - Inglês*). In this programme, the learners, who are in its majority students of Brazilian federal universities and in, a very small number, civil servants of the same institutions, are enrolled in face-to-face courses which can be 16, 32 or 64 hour-long. Although each course varies in its specific goals, all of the courses are aimed at improving English skills of Brazilian learners inserted in the academic context.

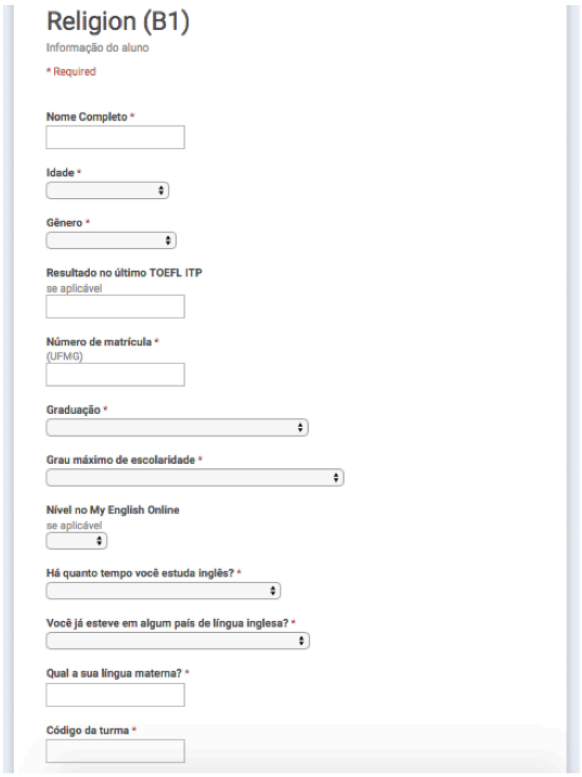
Throughout these courses, the students are required to do at least one graded assessment. Harnessing from this task, an online structure was developed in order to collect the learners' production, and two types of tasks were set in order to meet two main needs of the learners: getting a language proficiency certificate and developing skills such as writing, reading and listening, which are necessary for academic success.

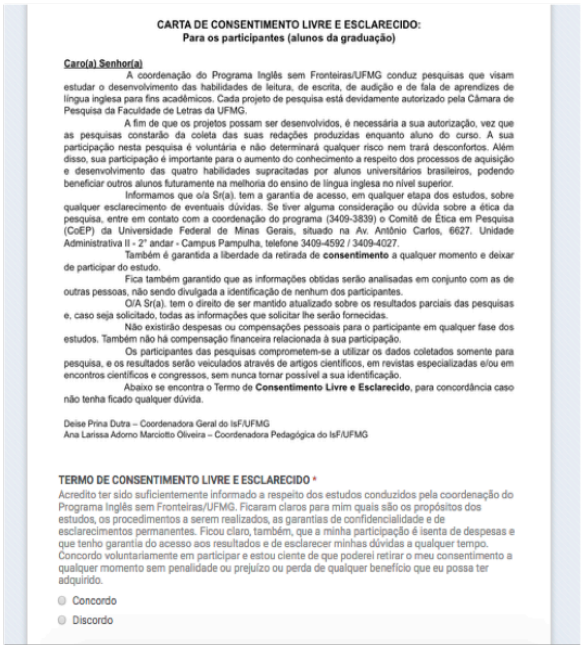
The first test type was based on the structure of language proficiency tests, such as International English Language Testing System (IELTS) and Test of English as Foreign Language (TOEFL). The online tests were created featuring three sections: reading, listening and writing. For the first two sections, students have to answer multiple-choice questions about a text and a video that are given to them. In the last part, students are required to write a text, which can be an integrated or an independent task. For the former, the learners have to either watch a video, or analyse a graph or a written text and, then, write an average of 200 words. For the latter, students are given a question and then they have to write, approximately, 300 words on the given topic.

The second type of assessment developed covers four academic writing genres: academic e-mails, statement of purpose, summary and abstract. These genres are approached in in-class activities and, at the end of each module, students are required to submit their own written production.

Added to the textual production itself, the dataset also features some learner and task variable. Before starting the task, the participants are asked to fill in a digital form through *Google Forms* with their information and to read a consent form (figure 2) for their participation in the research, with which they may choose to agree or disagree. At this point, the participants should indicate their age, gender, undergraduate major, highest degree or level of school completed, time spent studying English, time spent in an English-speaking country, mother tongue, last TOEFL score and they should also indicate whether the activity is being done with or without supervision.

Figure 2 - Online form (left) and consent form (right)





CARTA DE CONSENTIMENTO LIVRE E ESCLARECIDO:
Para os participantes (alunos da graduação)

Caro(a) Senhor(a)

A coordenação do Programa Inglês sem Fronteiras/UFMG conduz pesquisas que visam estudar o desenvolvimento das habilidades de leitura, de escrita, de audição e de fala de aprendizes de língua inglesa para fins acadêmicos. Cada projeto de pesquisa está devidamente autorizado pela Câmara de Pesquisa da Faculdade de Letras da UFMG.

A fim de que os projetos possam ser desenvolvidos, é necessária a sua autorização, vez que as pesquisas constarão da coleta das suas redações produzidas enquanto aluno do curso. A sua participação nesta pesquisa é voluntária e não determinará qualquer risco nem trará desconfortos. Além disso, sua participação é importante para o aumento do conhecimento a respeito dos processos de aquisição e desenvolvimento das quatro habilidades supracitadas por alunos universitários brasileiros, podendo beneficiar outros alunos futuramente na melhoria do ensino de língua inglesa no nível superior.

Informamos que o/a Sr(a). tem a garantia de acesso, em qualquer etapa dos estudos, sobre qualquer esclarecimento de eventuais dúvidas. Se tiver alguma consideração ou dúvida sobre a ética da pesquisa, entre em contato com a coordenação do programa (3409-3839) o Comitê de Ética em Pesquisa (CoEP) da Universidade Federal de Minas Gerais, situado na Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Campus Pampulha, telefone 3409-4592 / 3409-4027.

Também é garantida a liberdade da retirada de **consentimento** a qualquer momento e deixar de participar do estudo.

Fica também garantido que as informações obtidas serão analisadas em conjunto com as de outras pessoas, não sendo divulgada a identificação de nenhum dos participantes.

O/A Sr(a). tem o direito de ser mantido atualizado sobre os resultados parciais das pesquisas e, caso seja solicitado, todas as informações que solicitar lhe serão fornecidas.

Não existirão despesas ou compensações pessoais para o participante em qualquer fase dos estudos. Também não há compensação financeira relacionada à sua participação.

Os participantes das pesquisas comprometem-se a utilizar os dados coletados somente para pesquisa, e os resultados serão veiculados através de artigos científicos, em revistas especializadas e/ou em encontros científicos e congressos, sem nunca tornar possível a sua identificação.

Abaixo se encontra o Termo de **Consentimento Livre e Esclarecido**, para concordância caso não tenha ficado qualquer dúvida.

Deise Prina Dutra – Coordenadora Geral do IsF/UFMG
Ana Larissa Adorno Marciotto Oliveira – Coordenadora Pedagógica do IsF/UFMG

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO *

Acredito ter sido suficientemente informado a respeito dos estudos conduzidos pela coordenação do Programa Inglês sem Fronteiras/UFMG. Ficaram claros para mim quais são os propósitos dos estudos, os procedimentos a serem realizados, as garantias de confidencialidade e de esclarecimentos permanentes. Ficou claro, também, que a minha participação é isenta de despesas e que tenho garantia do acesso aos resultados e de esclarecer minhas dúvidas a qualquer tempo. Concordo voluntariamente em participar e estou ciente de que poderei retirar o meu consentimento a qualquer momento sem penalidade ou prejuízo ou perda de qualquer benefício que eu possa ter adquirido.

Concordo
Discordo

For this study the version 1.0 of CorIsF[28], compiled from August 2014 to December 2015, was adopted. This version is composed of 145,043 words distributed across five genres and written by students from four universities, as described in table 5. The dataset is not collected following the strict rules that keep a corpus representativeness. What is aimed instead, is to gather learners' textual production throughout its constant growth in a way that it is kept well-documented and accessible for further analysis. By doing so, a great variety of subcorpora can then be handily derived from the dataset. Since the data can be used by to address different research questions, the processing necessary to make the data ready for analysis is fully described in the next section.

Table 5 - Word distribution in CorIsF v1.0

	integrated	**independent**	**e-mail**	**SOP**	**summary**	**total**
UFMG	42,950	68,852	1,364	7,222	14,921	135,309
UFLA	1,386	969	---	---	---	2,355
UFSJ	4,726	---	---	---	---	4,726
TOTAL	49,062	67,168	1,364	7,222	14,921	**142,390**

4.1.2 Data loading and cleaning

This section describes the process of loading, cleaning and processing the CorIsF dataset. Since one important goal of this study is to make this process reproducible by other researchers, all the steps were carried through the use of replicable and expandable scripts on the programming language R.

All the process is here divided in seven steps. Some functions were created to eliminate repetitive work and to make the process faster, since the dataset is

[28] For this research, only the texts produced at UFMG and from the integrated, independente and summary tasks will be used.

continuously growing. Therefore, this section provides a detailed description of these steps, which can be generally taken by anyone working with our dataset.

4.1.2.1 Clearing workspace and setting work directory

The first step to be done is to clear any previous work and to set the work directory. The path to the work directory should be modified to suit the user's own path.

```
# clear workspace
rm(list=ls(all=TRUE))
# set work directory (substitute the path accordingly)
setwd("/Users/username/Documents/Corpus/")
```

4.1.2.2 Loading the Packages

The data gathered with *Google Forms* is stored as a comma-separated file (*.csv*), which can be read in R as a data frame. A commonly used package to work with data frames is the `data.table`. This powerful and useful package offers functions such as subsetting, ordering, and grouping data. However, the use of functions may not be very intuitive for linguistics with little knowledge on programming languages. A more suitable package for simple operations would than be the `dplyr`. It provides the user with functions which are easily interpreted and efficient. Other packages that deal specifically with text mining and natural language processing, such as `tm`, `RWeka` and `openNLP` are also used in this research.

In order to use the packages, they should be previously installed in the library. The installation can be done with the function `install.packages` as in `install.packages("data.table")` . The packages should then be loaded to the system with the function `library` or `require`. To load the package `data.table`, for instance, the code `library("data.table")` should be used.

4.1.2.3 Loading the files

All the files can be loaded at once or it can be done individually by using `read.csv` as in the following code:

```
ufmg2015coffee <- read.csv("UFMGcoffeeA1mar2015.csv")
```

Loading files individually can be beneficial if there is interest in naming the objects differently from the files or if not all the files are to be used. Loading all of them at the same time, however, is not as time-consuming and avoids mistakes, such as loading the wrong files.

To load all the files at the same time, the following code can be used:

```
#list the files in the directory which ends in .csv
temp <- list.files(pattern="*.csv")
#load dplyr
library(dplyr)
#read all the listed csv files listed as tbl_df, and data.frame
for (i in 1:length(temp))
        assign(temp[i], tbl_df(read.csv(temp[i])))
```

The previous code not only reads the files, but also stores them as *tbl_df*, a necessary format when using the *dplyr* package. The function `class` prints the classes to which the object belongs, as in:

```
#check class
class(UFMGcoffeeA2ago2015.csv)
## [1] "tbl_df"     "tbl"        "data.frame"
```

4.1.2.4 Cleaning the files

As aforementioned, the forms used for data collection were developed in a way that, apart from the written task, the headings would be the same for all the files. However, if, for any reasons, such as corrupted files or form submitted incorrectly, the headings do not match, the names should be modified as in:

```
#change name of column 32
names(UFMGcoffeeA2ago2015.csv)[32] <- "Integrated.Task"
names(UFMGwaterB2nov2014.csv)[36] <- "Independent.Task"
```

Note that the name of the task varies according to the genre, but the word “Task” is kept in all the headings, to facilitate data retrieval. For this dataset there are 6 types of genres:

1. Integrated Task
2. Independent Task
3. e-mail Task
4. Statement of purpose Task
5. Summary Task
6. Abstract Task

Another issue to observe is that the field *A atividade está sendo realizada em sala de aula* was not present in the first developed forms. In order to add this information, it was first identified whether the activity was done with or without the teacher supervision and then the information was added to the data frame, with the following script:

```
#add the variable and its value. If the task was done with supervision,
use "Sim". Otherwise use "Não"
UFMGreligionB1ago2015.csv <- mutate(UFMGreligionB1ago2015.csv,
A.atividade.está.sendo.realizada.em.sala.de.aula. = "Sim")
```

Once the modifications are done to the data frames in the environment, the files should then be overwritten with these changes.

```
#save the modified object
write.csv(UFMGcoffeeA2ago2015.csv, file = "UFMGcoffeeA2ago2015.csv")
write.csv(UFMGwaterB2nov2014.csv, file = "UFMGwaterB2nov2014.csv")
```

4.1.2.5 Extracting texts and saving as .docx

Collecting the texts through Google Forms has its benefits, such as being free of cost and user-friendly. However, the students' texts are not stored in a convenient way to be marked and/or commented by teachers. A procedure that should hence be done is

to extract the written production and save it as .docx, which is a suitable format for text review.

Two main forms have been used by our team to execute this extraction: using the Google add-on *Save as doc*[29] and using a formula created specially for this purpose. Using the add-on can be valuable for users not familiarised with R, since it presents a user-friendly interface. However, the wanted cells should be selected manually, which is both time-consuming and error-prone. The formula described below is a better choice to deal with large volumes of data.

```
#Load necessary libraries
library(qdap)
library(ReporteRs)
#create a function to extract the texts
toDocx <- function(df, classCode, document) {
        extract <- df %>%
                filter(Código.da.turma == classCode) %>% #filter by class
code
                select(Nome.completo, e.mail, ends_with("Task"))  #select
'name', 'email' and 'written texts' column
        flexi <- FlexTable(extract)
        doqui = docx(title = "Texts")
        doqui <- addFlexTable(doqui, flexi)
        writeDoc(doqui, document)
}
```

Table 6 - toDocx arguments

Argument	Description
df	the data frame from which the texts will be extracted.
classCode	the code of the class for which the texts will be extracted.
document	the name of .docx file to be created.

This function extracts the written part of the exam, the name of the students and their e-mails and subsequently saves them as .docx. One of the arguments (table 6) of

[29] https://chrome.google.com/webstore/detail/save-as-doc/iekpcmcpnbgoldpmhfbioecljjjjnpap

the function is the 'class code', which allows the files to be grouped in batches of different classes.

For instance, to extract the texts of the class "16072" from the data frame "UFMGloveB1out2015.csv" to a file named *16072test2.docx* the following script should be used:

```
#extract texts of class 16072
toDocx(UFMGloveB1out2015.csv, 16072, "16072test2.docx")
```

Extracting the texts for all the classes can be time-consuming. Then, in order to extract all the texts from a specific batch, the *for loop* below can be used.

```
#generate a list with all the class codes
lista <- as.character(df$Código.da.turma) #df  is the object from where
the texts will be extracted
#execute  a for loop with the function toDocx and the elements of "lista""
for (i in 1:length(lista))
        toDocx(df, lista[i], paste(lista[i], ".docx", sep = ""))
```

4.1.2.6 Deleting unwanted data

Although all students were required to do the activities used for CorIsF compilation, their contribution to the corpus is optional and they can either agree or disagree with the consent form presented in the beginning of the test. Hence, the data from the students who do not wish to take part in the research should be removed from the dataset.

In order to remove these learners' production, the function `disagree` was created. When using this function, the rows of learners who select "disagree" when taking the tests are deleted and the previous table is overwritten by the new one.

An extra feature that this function has is to eliminate the user "Answer Key". This user is created so that the test open-ended questions can be automatically corrected. Since this information should not be included in the corpus either, it was decided to eliminate these two observations with the same function.

```
#create function to delete rows of students who do not wish to participate
#the output is saved with the same file name with the CSV extension or not
disagree <- function(df, isCSV = TRUE) {
        if (isCSV == TRUE) {
                file.name <- deparse(substitute(df))
        } else {
                file.name <- paste(deparse(substitute(df)), ".csv",
sep="")
        }
        write.csv(filter(df, TERMO.DE.CONSENTIMENTO.LIVRE.E.ESCLARECIDO ==
"Concordo" & Nome.completo != "Answer Key"), file = file.name)
}
```

The function above was created considering that all the objects can be loaded with the ".csv" extension in its name or not. The argument `isCSV` (table 7) indicates whether the name of the object contains the extension .csv in its name. If that is the case, the argument "value" should be TRUE. To clean the object `UFMGreligionB1ago2015.csv`, for instance, the function should be executed as follows:

```
disagree(UFMGreligionB1ago2015.csv, isCSV = TRUE)
```

Table 7 - disagree arguments

Argument	**Description**
df	the data frame to be cleaned
isCSV	logical. Indicates whether the extension .csv should be included in the file name (FALSE) or not (TRUE)

4.1.2.7 Selecting relevant information

Since the form for data collection was developed to meet not only methodological needs, but also pedagogical ones, some of the collected information may not be relevant for this data analysis. By eliminating such extra information, less processing time and storage will be demanded.

At this step, all the answers to the close-ended questions and learners' name and registration number will be eliminated. The variables to be kept are: age, gender,

undergraduate major, highest degree or level of school completed, time spent studying English, time spent in an English-speaking country, mother tongue, last TOEFL score, and written task. Since the variable for the written part varies according to the genre, the function `contains` was adopted to select all the variables whose name contains the word "Task". This function only adopts one argument and its execution will automatically overwrite the old file.

```
#create function to keep specific variables
relevant <- function(df) {
        file.name <- deparse(substitute(df))
        write.csv(select(df, Idade, Gênero, Resultado.no.último.TOEFL.ITP,
Número.de.matrícula, Graduação, Grau.máximo.de.escolaridade,
Nível.no.My.English.Online, Há.quanto.tempo.você.estuda.inglês.,
Você.já.esteve.em.algum.país.de.língua.inglesa.,
Qual.a.sua.língua.materna.,
A.atividade.está.sendo.realizada.em.sala.de.aula., contains("Task")), file
= file.name)
}
```

4.2 Dataset Processing

In what follows, the methods used for subsetting and annotating the dataset are presented. In section 3.1, the focus was on the procedures used to subset the corpora according the variables. This section also presents a subcorpus of the British Academic Written English[30] (BAWE) which will be compared to one the of CorIsF subcorpora. Section 3.2 deals with the parts-of-speech annotation, explaining our annotator choice and describing the tagging procedure.

[30] this corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

4.2.1 Subsetting the dataset

In order to describe the interlanguage of CorIsF in different learning contexts and with different students' profiles the design of five small corpora (table 8) were set as it follows:

1. two corpora to identify the difference between the integrated and the independent productions, the C-int and C-ind;
2. two corpora to contrast the production of learners from courses with higher and lower demand, the C-high and C-low;
3. and one corpus with the summaries produced during course on academic writing, the C-sum, to be compared with the subsection of literature review of the British Academic Written Corpus (BAWE) .

Table 8 - Words per subcorpora

	c-ind	**c-int**	**c-high**	**c-low**	**c-sum**	**c-bawe**
words	68,852	42,950	68,897	60,999	14,921	81,281

With this division it is intended to describe the differences in learners' written production considering three aspects: type of task, learner profile and contrast production with native speakers.

Before extracting the information, the data frames must all be merged together, in order to make data easily accessible. To bind all the rows together, the script below was used.

```
#if not done before, list all the .csv files in the folder
filenames <- list.files(pattern="*.csv", full.names=TRUE)
#load necessary library
#if 'dplyr' is loaded remove detach the package
detach("package:dplyr", unload=TRUE)
library(plyr)
#for each element of the list, apply a function. results are kept as list
import.list <- llply(filenames, read.csv)
#merge all data frame
```

```
allData <- Reduce(function(x, y) merge(x, y, all=T), import.list,
accumulate=F)
#unload 'plyr' and to load dplyr
detach("package:plyr", unload=TRUE)
library(dplyr)
#make data avaiable as tbl_df too
allData <- tbl_df(allData)
```

4.2.1.1 Integrated and independent (c-ind and c-int)

As mentioned earlier, the dataset is mainly composed of argumentative and descriptive essays, which are here named as independent and integrated, respectively. In order to separate these two groups, the rows for each task were filtered:

```
#extracting independent tasks
cind <- filter(allData, Independent.Task != "NA")
#save text as a character object
charCind <- as.character(cind$Independent.Task)
#extracting integrated tasks
cint <- filter(allData, Integrated.Task != "NA")
#save text as a character object
charCint <- as.character(cint$Integrated.Task)
```

4.2.1.2 Higher and lower demand (c-high and c-low)

In order to verify the differences and similarities between two distinct profiles, the popularity of the course was chosen as a threshold. The definition of "popularity" was derived based on the minimum grade necessary to be admitted at the university. Although the production of other universities is also included in the dataset, we used the entrance cut-off point of *Sistema de Seleção Unificada*[31] (SISU) 2015 for the *Universidade Federal de Minas Gerais* (UFMG), once it is the source of most of the texts in the dataset.

[31] Sisu is the computerized system of the Ministry of Education through which public institutions of higher education offer places to their candidates.

Table 9 - List of the top 10 courses with high demand at SISU-UFMG 2015

Course	**Cut-off grade**
Medicina	798.50
Engenharia Química	788.36
Engenharia Mecânica	774.50
Engenharia Elétrica	759.30
Engenharia Civil	757.64
Engenharia De Produção	752.16
Engenharia Aeroespacial	751.04
Direito	746.02
Arquitetura e Urbanismo	738.56
Engenharia de Controle e Automação	737.64

Table 10 - List of the top 10 courses with low demand at SISU-UFMG 2015

Course	**Cut-off grade**
Engenharia Florestal	546.74
Engenharia Agrícola E Ambiental	560.92
Biblioteconomia	600.76
Arquivologia	615.06
Radiologia	623.86
Turismo	630.94
Aquacultura	633.72
Filosofia	634.50
Pedagogia	634.80
Ciências Socioambientais	634.96

The cut-off grade for each course was retrieved from the university website[32] considering the two following aspects. There is more than one call for university admission as well as different cut-off grades according to the candidate profile. The grades adopted here were the ones necessary to be approved in the first admission

[32] https://www.ufmg.br/sisu/cursos-e-vagas/

call and for candidates that fit the criteria *ampla concorrência*[33]. A simple arithmetic mean was carried to calculate the average cut-off point. All the courses below this number were listed as low-demand, while the courses above this limit were considered to present a high-demand (tables 9 and 10).

Once the courses were divided in these two groups, the dataset could be divided in two with the script below:

```
#create an object with the name of the courses for High and Low demand
lowNames <-  c("Filosofia", "Arquivologia", "Agronomia", "Engenharia de
Alimentos", "Zootecnia", "Gest. de Serv. de Saúde", "Engenharia Ambiental,
Engenharia Ambiental", "Fonoaudiologia", "Curso Superior de Tecnologia em
Radiologia", "Ciências Atuariais", "Controladoria e Finanças", "Gestão
Pública", "Turismo", "Cinema de Animação e Artes Digitais", "Engenharia
Ambiental", "Biblioteconomia", "Teatro", "Música", "Antropologia",
"Aquacultura", "Design de Moda", "Artes Visuais", "Nutrição",
"Estatística", "Ciências Sociais", "Design", "Química Tecnológica",
"Enfermagem", "Educação Física", "Letras", "Química", "Geografia",
"Farmácia", "Ciências Biológicas", "Pedagogia")

highNames <- c("Ciências Contábeis", "Engenharia Aeroespacial, Engenharia
Aeroespacial", "Engenharia de Minas", "Matemática", "Rel. Econ.
Internacionais", "Odontologia", "Engenharia Química", "História",
"Medicina", "Eng. de Controle e Automação, Eng. de Controle e Automação",
"Engenharia Metalúrgica", "Engenharia Aeroespacial", "Administração",
"Ciências Econômicas", "Geologia", "Engenharia de Produção", "Comunicação
Social", "Engenharia de Sistemas", "Biomedicina", "Fisioterapia",
"Psicologia", "Direito", "Arquitetura e Urbanismo", "Eng. de Controle e
Automação", "Física", "Sist. de Informação", "Medicina Veterinária",
"Ciência da Computação", "Engenharia Mecânica", "Engenharia Elétrica",
"Engenharia Civil")
```

The two groups are then extracted from the batch with all data:

```
#filter allData according to the demand
clow <- filter(allData, Graduação %in% lowNames)
chigh <- filter(allData, Graduação %in% highNames)
```

[33] When applying to SISU, the candidate should specify whether he or she wants to apply for a university place in a broad competition or to the places reserved according to the Law nº 12.711/2012 (Lei de Cotas)

In order to create an object with only the written production, it is necessary to select only the task columns, as in the script below:

```
#select only the tasks from the dataset - clow
a <- clow$Independent.Task
b <- clow$Summary.Task
c <- clow$Integrated.Task
d <- clow$email.Task
e <- clow$Statement.of.purpose.Task
```

The objects *a, b, c, d, e* have some empty rows, since there is only one type of written task for each test taker submission. Therefore, if the column "Independent Task" contains one text, for instance, the other columns will be filled with the symbol NA (not available). Hence, the following scripts are used to select and gather all the texts that are not NA.

```
#remove NA and transform objects into character - clow
a2 <- as.character(a[!is.na(a)])
b2 <- as.character(b[!is.na(b)])
c2 <- as.character(c[!is.na(c)])
d2 <- as.character(d[!is.na(d)])
e2 <- as.character(e[!is.na(e)])
#merge all characters into one - clow
charClow <- c(a2, b2, c2, d2, e2)
#select only the tasks from the dataset - chigh
a <- chigh$Independent.Task
b <- chigh$Summary.Task
c <- chigh$Integrated.Task
d <- chigh$email.Task
e <- chigh$Statement.of.purpose.Task
#remove NA and transform objects into character - chigh
a2 <- as.character(a[!is.na(a)])
b2 <- as.character(b[!is.na(b)])
c2 <- as.character(c[!is.na(c)])
d2 <- as.character(d[!is.na(d)])
e2 <- as.character(e[!is.na(e)])
#merge all characters into one - chigh
charChigh <- c(a2, b2, c2, d2, e2)
#remove unnecessary vectors
rm(a, a2, b, b2, c, c2, d, d2, e, e2, highNames, lowNames)
```

4.2.1.3 Summaries (c-sum)

This last subcorpus of the dataset contains summaries written during an Academic Writing Course. The extraction of this corpus is done as shown below:

```
#filter only summary tasks
csum <- filter(allData, Summary.Task != "NA")
#transform into character
charCsum <- as.character(csum$Summary.Task)
```

This subcorpus will be compared to a subcorpus of the British Academic Written English (BAWE) corpus, which features the genre "Literature Survey". This genre was chosen due to its similarities to the summary genre. Since BAWE will be used for comparison, it also needs to be loaded. The first step is to load BAWE documentation[34] and extract the name of the files which belongs to the genre family literature survey.

```
#load the dataframe with the student ID and the texts specifications
baweList <- read.csv2("CORPUS_TXT/BAWE.csv")
#load dplyr package. If 'plyr' is loaded, detach it
#detach("package:plyr", unload=TRUE)
library(dplyr)
#make the df tbl_df
baweList <- tbl_df(baweList)
#filer only the text ID for the genre "literatyre survey"
summaryIDbawe <- filter(baweList, genre.family == "literature survey" |
genre.family == "Literature survey" | genre.family == "literature survey +
proposal")
#create chr string with all the text file names
litSurvey <- as.character(unique(summaryIDbawe$id))
#save the names of the wanted files
lista <- as.list(paste("f", litSurvey, ".txt", sep = ""))
```

The corpus is than loaded as a character and processed.

```
#rename files so that the name won't start with a number
#file.rename(list.files(pattern="*.txt"), paste("f",
list.files(pattern="*.txt"), sep = ""))
```

[34] Documentation available at
http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf

```
#set the wd where BAWE files are
setwd("/Users/andressarodriguesgomide/Documents/FALE/Corisf/Corpus/CORPUS_
TXT/")
#use lapply to read lines and store it as character
charCbawe <- as.character(lapply(lista, readLines))
#remove the tags inside < >
charCbawe <- gsub("<\\w*>|</\\w*>|<\\w*/>", " ", charCbawe)
#go back to previous wd
setwd("/Users/andressarodriguesgomide/Documents/FALE/Corisf/Corpus")
```

4.2.1.4 Final processing

An additional cleaning process that might be done is to transform the character from upper to lower case and to remove the punctuation, as this may affect the list of frequent words as well as its annotation. The function `finalClean` was created in order to group those steps into a single one.

```
#load tm package
library(tm)
#create function transform from upper to lower case, to remove punctuation
and white spaces
finalClean <- function(val) {
                a <- tolower(val)
                b <- removePunctuation(a)
                c <- stripWhitespace(b)
                return(c)
}
#apply function and save it with different name
charChighClean <- finalClean(charChigh)
#apply function and save it with same name
charClow <- finalClean(charClow)
```

As some of the data analysis can be performed with softwares such as AntConc 3.4.3 (ANTHONY, 2014) and WordSmith Tools version 6 (SCOTT, 2016) it is useful to have the files saved as .txt as well. The files can be saved by using the R function `write` as in `write(charChigh, file = "charChighFinalClean.txt")`

4.2.2 Parts-of-speech tagging

There are several parts-of-speech annotators such as the Stanford Tagger (MANNING, 2011) and CLAWS 7 POS tagger (GARSIDE, SMITH 1997). The Apache OpenNLP tagger (Apache Software Foundation, 2004) was chosen for this study for the following reasons:

1. it is an open source tagger,
2. its implementation is relatively easy, and
3. its processing time is considerably good. For instance, the annotation of a 68,914-word subcorpus of CorIsF carried on a 2,8 GHz Intel Core i7 computer was done in 3.46024 minutes.

The tagset used for the English POS model is the same as the Penn Treebank tag set[35]. This is another positive aspect, since this tagset is frequently used with other annotators such as the Twitter POS[36] and in some annotators within Sketch Engine[37].

In order to make the annotation easier, the function below was created:

```
#Load packages
library(NLP)
library(openNLP)
library(openNLPmodels.en)
library(tm)
library(stringr)
library(gsubfn)
library(plyr)
#create function which annotates text and save the output as txt or return
the result
POScor <- function(batch, save = TRUE) {
        cor <- as.String(batch)
        file.name <- paste(deparse(substitute(batch)), "Tagged.txt",
sep="")
        cortag <- lapply(list(cor), function(x){
```

[35] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[36] https://gate.ac.uk/wiki/twitter-postagger.html
[37] https://www.sketchengine.co.uk/xdocumentation/wiki/tagsets/penn

```
                sent_token_annotator <- Maxent_Sent_Token_Annotator()
                word_token_annotator <- Maxent_Word_Token_Annotator()
                pos_tag_annotator <- Maxent_POS_Tag_Annotator()
                y1 <- annotate(x, list(sent_token_annotator,
word_token_annotator))
                y2 <- annotate(x, pos_tag_annotator, y1)
                y2w <- subset(y2, type == "word")
                tags <- sapply(y2w$features, '[[', "POS")
                r1 <- sprintf("%s/%s", x[y2w], tags)
                r2 <- paste(r1, collapse = " ")
                if(save == TRUE) {
                        write(r2, file = file.name)
                }
                return(r2)
                } )
}
```

Once the subcorpus is subset as demonstrated in the previous section, it can than be annotated with the function `POScor`. The function admits two arguments, being the first the name of the batch to be annotated and the second a logical argument to determine whether the output should be saved as a .txt file or it should only be returned and assigned to a variable, as demonstrated in the script below.

```
#annotate the subcorpus and save it as txt
POScor(charCsum, save = TRUE)
#since the default for the argument 'save' is TRUE, the function action
will have the same result if the argument is not included, as in:
#POScor(charCsumClean)
#annotate the subcorpus and return the result, which should be assigned to
a variable
taggedCsum <- POScor(charCsum, save = FALSE)
```

To calculate the time spent for the tagging process, the following script was used.

```
start.time <- Sys.time()
POScor(charCsum)
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
```

4.3 Extraction of key features

Once the data is processed, the linguistic information can then be retrieved. The linguistic and discourse features are approached through the following aspects: variability of parts-of-speech (POS); frequency of types and tokens; and n-grams.

The scripts developed to identify this information, being some of them adapted from Gries (2009), are here described.

4.3.1 Types and tokens

The first step to compare the frequency of types and tokens was to verify the number of words in each subcorpus. The word count was first made through three different forms:

1. Using online word counters such as *Word Counter Net*[38]
2. Adopting the pattern "\W+" for the definition of word and using the function `strsplit`
3. Using the function `word_count` from the package `qdap`

The results achieved with all the three procedures were substantially similar, with minor variances, as the definition of word may vary. Taking in consideration that the online counters demand an environment different from R, and considering that loading the package `qdap` requires more processing time, the procedure described in item was 2 was adopted to calculate the number of words. The following function prints out the number of types and tokens.

```
#function for types and tokens
typeEtoken <- function(batch) {
        tokens <- length(unlist(strsplit(batch, "\\W+")))
        types <- length(unique(unlist(strsplit(batch, "\\W+"))))
```

[38] https://wordcounter.net

```
        return(data.frame(tokens = tokens, types = types))
}
#execute the function
typeEtoken(charCbawe)
##   tokens types
## 1  82734 11644
```

4.3.1.1 Frequency list

The formula `freqList` was created to generate a list with the real and the normalised frequency. The normalised frequency was calculated by using the ratio per thousand. Although the ratio per million is more commonly used, it was not adopted here due to the small size of the dataset. The formula takes four arguments as described in table 11 and the content can be assigned to a variable in R as in `sortedDFnorm <- freqList(charChigh, 10, stopls = TRUE, save = TRUE)` or saved as a .csv file on the set working directory.

Table 11 - freqList arguments

Argument	Description
batch	a character vector of the subcorpus
minfreq	the minimum value to be included in the list
stopls	logical, indicates if stop words should be removed or not
save	logical, indicates if the output should be saved as csv or not

```
#create function
freqList <- function(batch, minfreq, stopls = FALSE, save = FALSE) {
        words <- unlist(strsplit(batch, "\\W+")) #separate the words and
store as a vector
        tokens <- length(words) #save number of tokens
        if (stopls == TRUE) {
                stopList <- stopwords(kind = "en")
                words <- words[!(words %in% stopList)]
        }
        #create table with list and sort according to its frequency
        freq.list <- table(words)
        sorted.freq.list <- sort(freq.list, decreasing = TRUE)
        #sort list with minimum value
        sorted.freq.list <- sorted.freq.list[sorted.freq.list > minfreq]
        #paste word and freq
```

```
        sorted.table <- paste(names(sorted.freq.list), sorted.freq.list,
sep= "\t")
        #store result in a data frame
        sortedDF <- data.frame(do.call(rbind, strsplit(sorted.table,
split="\t")))
        #change names
        names(sortedDF)[1] <- "word"
        names(sortedDF)[2] <- "frequency"
        #convert to data table
        sortedDF <- tbl_dt(sortedDF)
        #add normalised frequency variable
        sortedDFnorm <- mutate(sortedDF, per1000 =
as.numeric(levels(sortedDF$frequency))[sortedDF$frequency] * 1000 /
tokens)
                if (save == TRUE) {
                file.name <- paste(deparse(substitute(batch)),
"freqList.csv", sep="")
                write.csv2(sortedDFnorm, file = file.name)
                }
        return(sortedDFnorm)
}
```

4.3.2 Parts-of-speech

To verify the difference in POS usage, the formula `posList` was created with the arguments below.

Table 12 - posList arguments

Argument	Description
textfile	a text file with the annotated corpus
file.name	a character string naming the output file
save	logical, indicates if the output should be saved as .csv file or not

```
#create formula
posList <- function(textfile, file.name, save = FALSE) {
        #read in tagged file
        taggedtext <- scan(file = textfile, what = "char")
        freq.list <- table(taggedtext)
        sorted.freq.list <- sort(freq.list, decreasing = TRUE)
        sorted.table <- paste(names(sorted.freq.list), sorted.freq.list,
sep= "\t")
        #store result in a data frame
```

```
        sortedDF <- data.frame(do.call(rbind, strsplit(sorted.table,
split= "/+")))
        #make new df splitting tags and frequency
        df1 <- data.frame(do.call('rbind',
strsplit(as.character(sortedDF$X2), "\t",fixed=TRUE)))
        #get the sum of each POS
        df2 <- aggregate(as.numeric(levels(df1$X2))[df1$X2],
by=list(X1=df1$X1), FUN=sum)
        #add a column with the percentage and arrange it in descending
order
        df3 <- df2 %>%
                mutate(percentage = 100* x /sum(x)) %>%
                arrange(desc(x))
        #change names
        names(df3)[1] <- "POS"
        names(df3)[2] <- "frequency"
        if (save == TRUE) {
                write.csv2(df3, file = file.name)
        }
        return(df3)
}
```

The output of this function is a data frame with three variables: the tag for the part-of-speech, the raw frequency and the normalised frequency. As for the word frequency list, the data frame can be either stored as a new variable or saved as a .csv file.

4.3.3 n-grams

The function to generate a list of the n-grams is similar to the one for token frequency and the arguments are described on table 13.

Table 13 - ngramList arguments

Argument	**Description**
batch	a character vector of the subcorpus
gramMin	the minimum value for the n-gram
gramMan	the maximum value for the n-gram
minfreq	the cut-off frequency
save	logical, indicates if the output should be saved as .csv file or not

```
#load necessary libraries
library(RWeka)
## Warning: package 'RWeka' was built under R version 3.1.3
library(dplyr)
#create function
ngramList <- function(batch, gramMin, gramMax, minfreq = 10, save = FALSE)
{
        gramas <- NGramTokenizer(batch, Weka_control(min = gramMin, max =
gramMax)) #break into ngrams
        tokensG <- length(gramas) #count the number of n-grmas
        freqList <- table(gramas) #create table with ngrams
        sortedls <- sort(freqList, decreasing = TRUE) #sort table
according to frequency
        sortedls <- sortedls[sortedls > minfreq]
        sortedTb <- paste(names(sortedls), sortedls, sep= "\t") #paste n-
gram and frequency
        sortedDF <- data.frame(do.call(rbind, strsplit(sortedTb,
split="\t"))) #save as data frame
        #change names
        names(sortedDF)[1] <- "ngram"
        names(sortedDF)[2] <- "frequency"
        #convert to data table
        sortedDF <- tbl_dt(sortedDF)
        #add normalised frequency
        sortedDFnorm <- mutate(sortedDF, per1000 =
as.numeric(levels(sortedDF$frequency))[sortedDF$frequency] * 1000 /
tokensG)
        if (save == TRUE) {
                file.name <- paste(deparse(substitute(batch)),
"ngramFreq.csv", sep="")
                write.csv2(sortedDFnorm, file = file.name)
        }
        return(sortedDFnorm)
}

#example of execution on c-high subcorpus for 3-grams and 4-grams
cHighGrams <- ngramList(charChighClean, 3, 4, 30)
cLowGrams <- ngramList(charClow, 3, 4, 30)
```

4.4 Data visualization

This final section describes the procedure adopted to derive visual information from the data obtained in the previous section. Although it is possible to carry the analysis

without visual aids, it is believed that the proper use of data visualization can amplify cognition and aid both information processing and pattern recognition (KIRK, 2012). The following three subsections describe the procedure to generate word clouds and plots which represent the frequency of types, POS and n-grams.

4.4.1 Word cloud

Generating a word cloud with the most frequent words allows a better visualization of differences in word usage. In order to simplify the implementation of the function `wordcloud` from the homonymous package, the function `cloudCor` was created. This function, which takes three arguments as described on table 14, generates a word cloud with a list of frequent words and save it as a .png file.

Table 14 - cloudCor arguments

Argument	Description
batch	a character vector of the subcorpus
file.name	the name of the png output file
scale	vector of length 2 indicating the range of the size of the words

```
#load libraries
require(wordcloud)
require(XML)
require(tm)
require(wordcloud)
require(RColorBrewer)
#create function
cloudCor <- function(batch, file.name, scale = c(4,.5)) {
        freqList2 <- function(batch) {
        words <- unlist(strsplit(batch, "\\W+")) #seperate the words and
store as a vector
        tokens <- length(words) #save number of tokens
        stopList <- stopwords(kind = "en")
        words <- words[!(words %in% stopList)]
        #create table with list and sort according to its frequency
        freq.list <- table(words)
        sorted.freq.list <- sort(freq.list, decreasing = TRUE)
        #sort list with minnimum value
        sorted.freq.list <- sorted.freq.list[sorted.freq.list > 2]
```

```
        #paste word and freq
        sorted.table <- paste(names(sorted.freq.list), sorted.freq.list,
sep= "\t")
        #store result in a data frame
        sortedDF <- data.frame(do.call(rbind, strsplit(sorted.table,
split="\t")))
        #change names
        names(sortedDF)[1] <- "word"
        names(sortedDF)[2] <- "frequency"
        #convert to data table
        sortedDF <- tbl_dt(sortedDF)
        return(sortedDF)
}
        a <- freqList2(batch)
        pal <- brewer.pal(8,"Dark2")
        png(file.name, width=480, height=480, pointsize= 20)
        wordcloud(a$word,
          as.numeric(levels(a$frequency))[a$frequency],
          scale=c(4,.2),
          min.freq=20,
          max.words=Inf,
          random.order=FALSE,
          rot.per=.15,
          colors= pal)
        dev.off()
}
#example
cloudCor(charChighClean, "cHighCloud.png")
```

4.4.2 Plotting POS difference

The function `posList` generates a data table with the POS usage. The function below prints a plot comparing the difference in POS frequency in two different data tables. The arguments for the function are described in table 15.

```
#create function
plotPOS <- function(POSlist1, POSlist2, name1, name2, file.name, tags =
c("NN", "IN", "DT", "NNS", "JJ", "VB", "RB", "CC", "VBZ", "VBP", "PRP",
"TO", "VBN", "CD", "MD")) {
        a <- filter(POSlist1, POS == tags)
        b <- filter(POSlist2, POS == tags)
        a$group <- name1
        b$group <- name2
        ab <- rbind(a, b)
        png(file.name)
```

```
        print(ggplot(ab, aes(POS, percentage, group=group,col=group),
environment = environment()) + geom_point() + labs(title = "POS
Frequency"))
        dev.off()
}
##example
#create POSlistChigh and POSlistClow
POSlistChigh <- posList("charChighCleanTagged.txt")
POSlistClow <- posList("charClowCleanTagged.txt")
plotPOS(POSlistChigh, POSlistClow, "high", "low", "high_lowPOS.png",
c("NN", "IN", "DT", "NNS", "JJ", "VB", "RB", "CC", "VBZ", "VBP", "PRP",
"TO", "VBN", "CD", "MD"))
```

Table 15 - plotPOS arguments

Argument	Description
POSlist1	a data frame with the percentage of POS generated with the function `posList`
POSlist2	a data frame with the percentage of POS generated with the function `posList`
name1	a character string with name of the first group
name2	a character string with name of the second group
file.name	the name of the png output file
tags	the parts-of-speech to be included in the plot. If not mentioned, the default list ("NN", "IN", "DT", "NNS", "JJ", "VB", "RB", "CC", "VBZ", "VBP", "PRP", "TO", "VBN", "CD", "MD") will be adopted.

4.4.3 Plotting n-gram usage

This last function plots a graph to compare the usage of given n-grams in two different subcorpora and adopts the arguments described in table 16.

```
#load package
library(ggplot2)
#create function
plotNgram <- function(ngramL1, ngramL2, name1, name2, grams, file.name) {
        a <- filter(ngramL1, ngram == grams)
        b <- filter(ngramL2, ngram == grams)
        a$group <- name1
        b$group <- name2
        ab <- rbind(a, b)
        g <- ggplot(ab, aes(ngram, per1000))
        png(file.name)
        print(g + geom_point(aes(color = group), size = 5, alpha = 1/2) +
```

```
labs(title = "n-gram Frequency"))
        dev.off()
}
#example
plotNgram(cHighGrams, cLowGrams, "c-high", "c-low", c("a lot of", "cup of
coffee", "in the world"), "grams.png")
```

Table 16 - plotNgram arguments

Argument	Description
ngramL1	a data frame with n-grams generated with the function `ngramList`
ngramL2	a data frame with n-grams generated with the function `ngramList`
name1	a character string with the name of the first group
name2	a character string with the name of the second group
file.name	the name of the png output file
grams	the n-grams to be compared

The present chapter dealt with the process of compiling and processing a dataset; deriving specific corpora from this data; and extracting and visualizing linguistic information from it. Although the scripts developed and described here are fairly simple and do not constitute cutting-edge programming, we hope to demonstrate with them that language researchers can profit from acquiring some basic programming skills. A detailed analysis of the information derived from the use of these scripts is presented in the next chapter.

5 Analysis and interpretation of the data

The procedures described in the previous chapter had the aim to present the tools to analyse the written production of IsF learners. This chapter presents the analysis and interpretation of the six subcorpora presented in chapter three. The analysis will be divided in three parts, addressing the following aspects for all the aforementioned subcorpora: types and tokens frequency; the POS usage; and the distribution of n-grams.

5.1 Word Frequency List

Some contrasting aspects related to type frequency, word choice and distribution are here described. However, the differences identified were rather subtle and did not yield much insight into the variations within groups.

The first analysis carried was related to the frequency of types and tokens, which were extracted with a minimum cut-off point of 1.4 words per thousand. This low cut-off point was adopted due to the small size of the corpus and to the fact that it covers approximately the 50 most frequent words in each subcorpora. The intent here was to grasp the main lexical choice of learners, differently from other studies such Coxhead (2000) and Gardner & Davies (2013) which have adopted more elaborated strategies for extracting these word lists. The two subcorpora with highest type/token ratio were c-sum and c-bawe (table 17). This result is possibly due to the presence of academic specific words in these subcoprora, since the texts in both of them were written after academic articles. The high frequency of types in c-bawe may also be explained by the high proficiency of the writers, who are mostly English native speakers.

Opposing to these two subcorpora, c-int presents the lowest type frequency. Such low frequency might be attributed to the reliance of learners on the essay prompt which leads to the repletion of the same words. On integrated tasks, learners tend to use words and phrases present on the essay commands and/or on the given graphs. For instance, the words 'coffee', 'caffeine' and 'cups' are among the ten most frequent words in c-int. These words are all present in the infograph of the test named 'Coffee' (appendix B). However, in this study the difference in frequency may also be due to a higher variety of independent tasks. Considering that there are five different topics for the independent tasks, but only two for the integrated ones (appendix A) a wider diversity of words is expected to be seen in the former type of task. A more detailed analysis for each group of subcorpora will be presented in subsections 4.1.1 to 4.1.3.

Table 17 - Number of tokens and types across subcorpora and their type/token ratio

subcorpus	**tokens**	**types**	**type/token ratio (%)**
c-high	68,897	6,070	8,81%
c-low	60,999	5,523	9,05%
c-ind	68,852	5,535	8,03%
c-int	42,950	3,162	7,36%
c-sum	14,921	2,273	15,23%
c-bawe	81,281	10,848	13,34%

5.1.1 c-high and c-low

When generating the word cloud and the list of frequent words for c-high and c-low, there is not a clear difference between these two groups. Table 18 lists the 45 most frequent words for C-high and C-low without the stop words. It is clear that these frequent words are related to the essay theme, which may indicate that learners from both groups strongly rely on the prompt. One difference that should be noted, however, is that, among the 45 most frequent words, the words 'religion' and 'water'

appear on c-high's list, but not on c-low's one. Such words are related to the prompt of tests given to students of C1 and C2 levels (appendix A), which led us to verify and confirm that there are more learners from the courses with a high demand enrolled in more advanced courses. Nonetheless, this is still a confirmation that students from both levels rely heavily on the prompts.

5.1.2 c-ind and c-int

As with the previous two subcorpora, the lists of frequent words for c-ind and c-int also feature a great number of words related to the essay topic. This usage is even more perceptible in integrated tasks. From the 50 most frequent words in c-int, only 13 are not directly related to the topic, while for c-ind, only 10 words are strictly associated with the essay theme (table 19)[39].

From this analysis two inferences can be made. First, those students tend to use their own words when writing an independent task, which makes this type of task suitable for a diagnostic activity (e.g. SPOLSKY, HULT; 2008). Another perspective is that, when submitted to an integrated task, learners are more prone to use words that may not common to their lexicon, but which are presented in the prompt. By being induced to use vocabulary to which they are less familiarised with, such as "psychoactive" and "caffeine", the students are more likely to internalize new words. Integrated tasks can be, hence, useful for formative activities. Furthermore, working with these two types of task prepare students for proficiency exams as TOEFL iBT and IELTS, which use both task types.

---

[39] The topic related words are highlighted

Table 18 - Most frequent words for c-high and c-low

	c-high			c-low		
	word	**freq**	**per 1000**	**word**	**freq**	**per 1000**
1	coffee	694	10,073	coffee	670	10,983
2	people	665	9,652	people	646	10,590
3	can	455	6,604	can	483	7,918
4	media	374	5,428	media	381	6,246
5	water	254	3,686	day	239	3,918
6	us	241	3,497	one	219	3,590
7	one	240	3,483	world	216	3,541
8	day	224	3,251	information	204	3,344
9	information	215	3,120	think	193	3,163
10	world	215	3,120	important	186	3,049
11	religion	193	2,801	us	182	2,983
12	important	190	2,757	per	171	2,803
13	person	180	2,612	person	166	2,721
14	many	179	2,598	first	159	2,606
15	per	171	2,481	many	159	2,606
16	like	165	2,394	like	156	2,557
17	think	165	2,394	will	152	2,491
18	will	164	2,380	need	151	2,475
19	need	162	2,351	year	151	2,475
20	first	158	2,293	time	148	2,426
21	news	154	2,235	religion	146	2,393
22	year	154	2,235	americans	143	2,344
23	time	152	2,206	news	139	2,278
24	reality	142	2,061	women	135	2,213
25	caffeine	134	1,944	drink	128	2,098
26	americans	131	1,901	see	127	2,082
27	consume	131	1,901	consumed	124	2,032
28	good	129	1,872	dont	121	1,983
29	cups	128	1,857	caffeine	118	1,934
30	way	124	1,799	cups	118	1,934
31	women	122	1,770	good	115	1,885
32	lot	121	1,756	reality	115	1,885
33	also	119	1,727	impression	112	1,836
34	billion	119	1,727	men	112	1,836
35	consumed	117	1,698	know	110	1,803
36	dont	117	1,698	blood	109	1,786
37	impression	116	1,683	divorce	107	1,754
38	know	115	1,669	billion	106	1,737
39	worlds	114	1,654	coffe	105	1,721
40	men	112	1,625	age	100	1,639
41	see	111	1,611	consume	100	1,639
42	coffe	110	1,596	cup	99	1,622
43	example	109	1,582	lot	99	1,622
44	just	105	1,524	get	98	1,606
45	production	103	1,494	production	98	1,606

Table 19 - Most frequent words for c-ind and c-int

	c-ind			c-int		
	word	**freq**	**per1000**	**word**	**freq**	**per1000**
1	media	803	11,662	coffee	1344	31,292
2	people	799	11,604	people	457	10,640
3	can	670	9,731	day	379	8,824
4	information	396	5,751	per	344	8,009
5	water	349	5,068	us	294	6,845
6	religion	333	4,836	year	294	6,845
7	one	318	4,618	americans	284	6,612
8	think	310	4,502	caffeine	255	5,937
9	news	309	4,487	women	252	5,867
10	important	269	3,906	cups	250	5,820
11	first	267	3,877	consumed	238	5,541
12	world	267	3,877	billion	230	5,355
13	reality	266	3,863	drink	217	5,052
14	many	258	3,747	coffe	216	5,029
15	person	231	3,355	men	214	4,982
16	impression	228	3,311	divorce	207	4,819
17	like	216	3,137	consume	202	4,703
18	will	202	2,933	can	200	4,656
19	need	188	2,730	worlds	198	4,610
20	know	187	2,715	divorced	196	4,563
21	time	185	2,686	cup	195	4,540
22	good	182	2,643	production	194	4,516
23	dont	181	2,628	world	176	4,097
24	see	169	2,454	american	172	4,004
25	believe	164	2,381	popular	165	3,841
26	way	164	2,381	age	160	3,725
27	bad	161	2,338	drug	154	3,585
28	talk	160	2,323	13	138	3,213
29	example	154	2,236	imports	134	3,119
30	things	153	2,222	years	132	3,073
31	opinion	145	2,105	united	126	2,933
32	us	135	1,960	marriage	121	2,817
33	just	129	1,873	450	120	2,793
34	change	128	1,859	need	120	2,793
35	lot	128	1,859	get	117	2,724
36	creates	119	1,728	states	116	2,700
37	life	115	1,670	breakfast	113	2,630
38	today	115	1,670	person	111	2,584
39	others	112	1,626	spend	111	2,584
40	religions	111	1,612	million	107	2,491
41	sometimes	110	1,597	rico	104	2,421
42	going	109	1,583	puerto	103	2,398
43	new	109	1,583	start	103	2,398
44	something	107	1,554	number	102	2,374
45	fact	104	1,510	lot	98	2,281

Table 20 - Most frequent words for c-bawe and c-sum

	c-bawe			c-sum		
	word	**freq**	**per 1000**	**word**	**freq**	**per 1000**
1	species	306	3,764	blood	182	12,197
2	et	247	3,038	mice	116	7,774
3	al	244	3,001	young	114	7,640
4	also	225	2,768	old	79	5,294
5	can	213	2,620	people	74	4,959
6	research	171	2,103	gdf11	70	4,691
7	study	160	1,968	can	64	4,289
8	one	156	1,919	coffee	63	4,222
9	may	152	1,870	results	56	3,753
10	different	147	1,808	university	56	3,753
11	control	144	1,771	will	54	3,619
12	used	137	1,685	study	53	3,552
13	will	137	1,685	also	51	3,418
14	new	127	1,562	effects	50	3,350
15	two	120	1,476	one	48	3,216
16	within	112	1,377	group	47	3,149
17	time	110	1,353	humans	42	2,814
18	however	109	1,341	language	39	2,613
19	many	108	1,328	older	39	2,613
20	2004	105	1,291	used	39	2,613
21	organic	101	1,242	transfusion	38	2,546
22	use	100	1,230	experiment	37	2,479
23	work	99	1,217	experiments	35	2,345
24	pulse	98	1,205	human	34	2,278
25	anthocyanins	93	1,144	disease	32	2,144
26	studies	93	1,144	dna	32	2,144
27	found	89	1,094	parabiosis	31	2,077
28	important	86	1,058	protein	31	2,077
29	using	85	1,045	younger	31	2,077
30	data	84	1,033	effect	30	2,010
31	2001	81	0,996	style	30	2,010
32	food	80	0,984	growth	29	1,943
33	system	80	0,984	alzheimers	28	1,876
34	laser	79	0,971	research	28	1,876
35	evolutionary	78	0,959	scientists	28	1,876
36	results	78	0,959	animals	27	1,809
37	b	76	0,935	body	27	1,809
38	example	76	0,935	cells	27	1,809
39	2002	75	0,922	factor	27	1,809
40	genes	75	0,922	health	27	1,809
41	genetic	75	0,922	plasma	27	1,809
42	effects	74	0,910	heterochronic	26	1,742
43	university	74	0,910	stanford	26	1,742
44	2003	73	0,898	using	26	1,742
45	2005	70	0,861	volunteers	26	1,742

5.1.3 c-sum and c-bawe

As expected, this third group results also present words related to the essay topic. However, they differ considerably from the two previous groups since they are written after scientific paper. Analysing the list of the most frequent words, we observe a great number of words specifically related to the articles which are being referred to, such as 'blood' and 'species'.

When comparing these two subcorpora, we notice a considerably lower number of topic related words in the BAWE subcorpus. It is worth highlighting, though, that the topics in BAWE come from a much wider diversity. Considering the list with the most frequent words (table 20) for both subcorpus, c-bawe list reveals some function words that are not included in c-sum list. The modal 'may', for instance, is the 9th most frequent words in c-bawe, but it does not appear in c-sum list. The low frequency of the modal verb 'may' in c-sum, might indicate that learners use other expressions instead. When observing the list of most frequent words, the word 'possible' is listed as 66th in c-sum, but does not appear in c-bawe's one, once its frequency (raw: 29, normalised: 0.35678695) is below the cut-off point. A test for a significant difference in frequency between the two corpora was then performed[40], and it was confirmed that the higher frequency of 'possible' observed in c-sum is statistically significant, with a log-likelihood of 20.2 (table 21). The sentences shown below (e.g. 1 to 4) are examples from c-sum in which the word 'possible' could be, with appropriate adaptations, replaced by the word 'may'. Some examples of how 'may' (e.g. 5 to 7) is used in c-bawe is also shown.

---

[40] Calculated with Paul Rayson's Log-likelihood and effect size calculator (http://ucrel.lancs.ac.uk/llwizard.html). The log-likelihood (LL) is a significance statistic and a LL score above 3.84 means that the result can be treated as significant. The LogRatio is an "effect-size" statistic and indicates how big the difference is. A word is 2^n times more frequent in a corpus than in another, where n is the log ratio. So, if a a LogRation is 3, the word is 8 times more frequent in the first corpus.

1. *they believe some day will be **possible** to elaborate a new drug that could be injected in older people.* (c-sum)
2. *For example is **possible** to use gesture, body language or posture, facial expression or an object like clothing* (c-sum)
3. *Their discussion starts over the assumption that under sufficient exposure to a language it is **possible** to acquire complete knowledge about the constructions and structure of it.* (c-sum)
4. *From these conclusions **it is possible** to infer that even in an Iranian context people make assumptions about strangers based on their dress style and react to them differently.* (c-sum)
5. *The homologies **may** have arisen due to similarities in the pathways of intracellular survival.* (c-bawe)
6. *As shown, this **may** be a very naïve view in light of empirical evidence and cross-comparative study among primates.* (c-bawe)
7. *The results of this **may** show that two different approaches need to be adopted depending on the age of the pupil.* (c-bawe)

Table 21 - Log-likelihood results for the word 'possible'

word	corpus 1	frequency	corpus 2	frequency	LL	LogRatio
possible	c-sum	21	c-bawe	29	20.2	1.98

5.1.4 Word cloud

The word clouds were generated with the intent of assisting data visualization. For this analysis, however, the word clouds were not as effective as the word list. Pairing the lists of the two groups being compared proved to be a faster and simpler process. Nonetheless, the word clouds could be used to identify straight away the most frequent words, as it is the case with "people" and "can", on figures 3 and 4.

Figure 3 - Word cloud with c-high (left) and c-low (right) frequent words

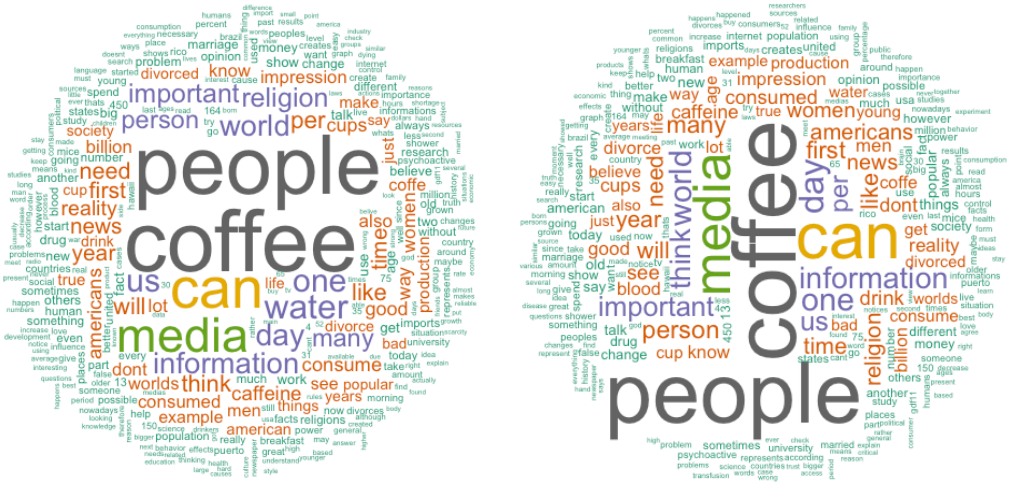


Figure 4 - Word cloud with c-ind (left) and c-int (right) frequent words

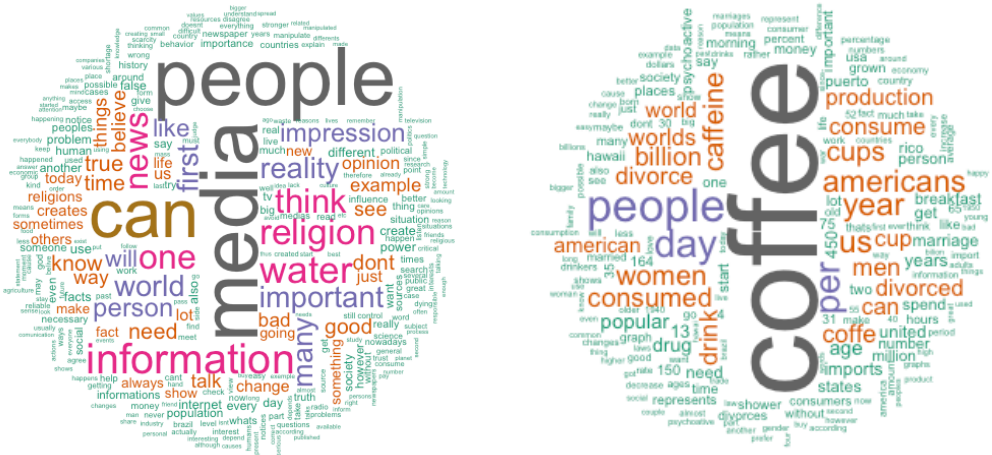


Figure 5 - Word cloud with c-sum (left) and c-bawe (right) frequent words

5.2 Parts-of-speech Usage

This section presents a detailed description of the distribution of the parts-of-speech usage for each of the six subcorpora. Tables with the POS are presented and some examples from the subcorpora are drawn. Item 5.2.4 presents an overview of all the groups and an analysis of the plots generated with the function created in the previous chapter.

5.2.1 c-high and c-low

When the frequency of each POS in both corpora was compared, no considerable differences were identified. The greatest difference observed was with determiners (DT) usage, occurring 8786 in c-high and 7497 in c-low. However, when the log-likelihood test is performed, the difference is rather subtle, with a log-likelihood of 5.51 and a LogRatio of 0.05 (table 22). This result may suggest that POS usage is more closely related to the type of task assigned than to difference in learner background and proficiency, as it will be demonstrated in the next section. However, it is worth emphasizing that the limited size of the data used in this study might have influenced this analysis.

Table 22 - Log-likelihood results for the usage of 'determiners' in c-high and c-low

POS	corpus 1	frequency	corpus 2	frequency	LL	LogRatio
determiner (DT)	c-high	8,786	c-low	7,497	5.51	0.05

5.2.2 c-ind and c-int

When contrasting the POS usage in c-ind and c-int some considerable differences were noticed, especially with the following POS: cardinal numbers, adverbs,

coordinating conjunctions, comparative adjective and personal pronoun. Table 23 shows the distribution of POS in both corpora.

Table 23 - Most frequent POS for c-ind and c-int and its normalised frequency

	C-IND			C-INT		
	POS	**freq**	**percentage**	**POS**	**freq**	**percentage**
1	NN	11649	16,827	NN	8338	19,277
2	IN	9087	13,126	IN	6819	15,765
3	DT	8667	12,520	DT	5641	13,041
4	NNS	5657	8,172	NNS	4022	9,298
5	JJ	5434	7,849	JJ	2408	5,567
6	VB	3768	5,443	CD	1973	4,561
7	RB	3332	4,813	VBZ	1696	3,921
8	CC	3180	4,593	RB	1471	3,400
9	VBP	2949	4,260	CC	1434	3,315
10	PRP	2836	4,096	VBP	1426	3,296
11	VBZ	2672	3,859	VB	1357	3,137
12	TO	1811	2,616	PRP	1203	2,781

The normalised frequency of cardinal numbers (CD) and comparative adjectives (JJR) are eight times and two times higher in c-int than c-ind, respectively. This difference is highly expected since for all the integrated tasks the prompt requires students to compare two graphs. On the other hand, the frequency of adverbs (RB), coordinating conjunctions (CC), and personal pronoun (PRP) is considerably higher in the independent tasks, as demonstrated in table 23. The underuse of these POS in c-int may also be explained by the nature of its task. When comparing graphs, the information is often punctual, making the use of conjunctions and adverbs less necessary, as demonstrated in example 8. Furthermore, the tone of this descriptive essay is even less personal than the one of argumentative essays, reducing the usage of personal pronouns.

8. *According to the graphs, women presents a bigger index of divorce that men. This can be explained mainly by the fact that women have been more independent em relation to their husbands. For example, women who are 30 years old presents a index of divorce of 15%, while men at the same age presents 10%. The divorce's indexes of men and women at age 55 and 50 were almost similar, probably because older people tend to stay married for a longer time, or because in the past years, to get marry was more serious, rigid than today.* (c-int)

Although the observations made here were expected given the different purpose of both task types, they were important in emphasizing how specific tasks demand very different skills from students. While integrated tasks deal with numbers and comparisons, independent tasks are useful resources when argument construction is to be worked.

5.2.3 c-bawe and c-sum

Both subcorpora presented a similar distribution, with some minor differences in its frequency (table 24). The high frequency of cardinal numbers in c-bawe was expected, since various texts are written after academic articles from the hard sciences. The greater number of adjectives was also expected, given the fact the c-bawe is composed of literature survey, with an evaluative opinion from the writers[41], while for c-sum students were required to write a short summary[42] of one of the three articles suggested.

---

[41] BAWE directions for the writing a literature survey: "includes summary of literature relevant to the focus of study and varying degrees of critical evaluation" (p. 44, BAWE documentation)

[42] IsF directions for the summary activity: "Choose one of the articles below and write a one-page summary." (extracted from the *Academic Writing Course Package* prepared for the IsF course)

Table 24 - Most frequent POS for c-bawe and c-sum and its normalised frequency

	C-BAWE			C-SUM		
	POS	**freq**	**percentage**	**POS**	**freq**	**percentage**
1	NN	17891	21,982	NN	3093	20,686
2	IN	11443	14,059	IN	2192	14,660
3	JJ	8717	10,710	DT	1740	11,637
4	DT	8578	10,539	NNS	1456	9,737
5	NNS	7378	9,065	JJ	1405	9,396
6	CC	3192	3,922	VB	569	3,805
7	RB	2849	3,500	VBN	503	3,364
8	VBN	2832	3,479	RB	450	3,009
9	VBZ	2403	2,952	VBZ	435	2,909
10	CD	2337	2,871	TO	433	2,895
11	VB	2302	2,828	CC	422	2,822
12	TO	1904	2,339	VBD	398	2,661

There is also a slightly lower frequency of coordinating conjunctions (CC) in c-sum when compared to c-bawe, which may suggest that Brazilians learners adopt different cohesive techniques when compared to native production. One assumption would be that the production in c-sum is more descriptive and less argumentative. When analysing the use of the coordinating conjunctions in context, we observe that the token 'yet' appears only two times in c-sum and in both of them as an adverb (example 9). C-bawe, on the other hand, counts with 16 occurrences, being 13 of them used as a coordinating conjunction (example 10).

9. *Wyss-Coray alerts that they don't know* ***yet*** *what can be the results.* (c-sum)
10. *they propagate themselves in much the same way as alleles,* ***yet*** *instead of leaping from body to body via eggs and sperm, memes jump from brain to brain* (c-bawe)

Table 25 - Log-likelihood results when the CC use in c-bawe and c-sum is compared

POS	corpus 1	freq	corpus 2	freq	LL	LogRatio
coordinating conjunction	c-bawe	3,192	c-sum	422	43.66	0.47

A final and interesting point observed is that c-sum production presented considerably more modal verbs and verbs in its base form than c-bawe (a normalised frequency 1.5 times higher for each case). After closely investigating both subcorpora, two inferences were made. The first one is that it seems that learners usually rely on modal verbs to express modality, while native speakers rely on other modal expressions. For instance, the expressions "to be likely" and "to be bound" (examples 11 and 12) are found in c-bawe, but not in c-sum. Another explanation for this divergence in POS frequency could be the overuse of "will" and "can" (examples 13 and 14) by learners, which present a frequency 2.15 and 1.6 times higher, respectively.

11. *They **are likely to** support all types of tourism development for the economic benefit of themselves and the community.* (c-bawe)
12. *Raising awareness about its presence **is bound to** have a positive effect on conservation efforts in the area.* (c-bawe)
13. *Wyss-Coray thinks that this treatment **will** have the same effects on humans.* (c-sum)
14. *Knowing this, we **can** infer that clothing is possibly one of the most important ways to communicate* (c-sum)

5.2.4 Plots and some considerations

The plots (figure 7) proved to be useful for a fast analysis since it was not necessary to verify individual values to identify the frequencies that are distinct. For c-high and c-low, for instance, it is clear that all the frequencies for the POS are approximately the same, whereas for c-int and c-ind, we can easily identify the POS that do not present similar frequency for both subcorpora.

5.3 n-grams

This section covers the analysis of the most frequent grams of 3, 4 and 5 words in all the subcorpora. The cut-off point adopted (20 times per million words) was the same as the commonly used in other studies (e.g. BIBER ET AL. 2004; CORTES, 2004; DUTRA; BERBER SARDINHA, 2013). Figure 8 presents the plot of the most frequent n-grams for each group.

5.3.1 C-high and c-low

In both groups, there was an extremely high number of topic-related chunks. Out of a list of the 124 most frequent chunks of 3, 4 and 5 n-grams, only seven of them are not connected to the task topics (table 26).

From the chunks that are not topic-related, the most frequent chunk by far for both subcorpora is 'a lot of'. The most used word with this expression is "people", in both groups. As previous research has shown, learners' written production bears more resemblance to spoken English than the language of native speakers (e.g. JUKNEVIČIENĖ, 2009), which is reinforced by the high frequency the 3-gram *a lot of (people).*

The second most frequently used chunk is "I think that", being its frequency slightly higher in c-low. Such difference should not be an indication of difference in

proficiency, but of the existence of the similar expression “I think the”. This chunk is the 78th most frequent n-gram in c-low but is absent from the list of the most frequent n-grams in c-high.

Table 26 - Most frequent non-topic related chunks of 3-5-grams in c-high and c-low

	C-HIGH		**C-LOW**	
chunk	**freq**	**per thousand**	**freq**	**per thousand**
a lot of	100	0,492	79	0,440
i think that	33	0,162	39	0,217
we can see	28	0,137	24	0,133
we need to	27	0,132	25	0,139
is very important	24	0,118	25	0,139
it is a	24	0,118	12	0,069
one of the	24	0,118	13	0,075
we have to	16	0,078	26	0,144
i think the	15	0,073	25	0,139

Other expressions that present a distinguished frequency is “we have to” and “one of the”. The former occurs 26 in c-low, but only 17 times in c-high, whereas the latter occurs 24 times in c-high and 11 in c-low. Although this quantitative analysis may not yield much insight, it leads to the qualitative analysis of sentences in which these n-grams are used. For instance, the sentences shown below suggest a more advanced use of English with the 3-gram ‘one of the’ than with ‘we have to. While the first sentence uses a subordinated clause and a more elaborated vocabulary, the second sentence is built with the personal pronoun ‘we’.

15. *“Although the reasons remain unknown, GDF11 falls with age in both humans and mice, which appears to be* ***one of the*** *causes of age-related deficiencies.”* (c-high)
16. *“We can't forgot that* ***we have to*** *has a critical opinion with all the things that we see.”* (c-low)

Nevertheless, these minor differences should be seen with cautious, once this is a very small sample size and the variation in frequency is also not salient. In any case, when noticing these variances in n-gram frequency and by analysing its use in context, inferences can be made.

5.3.2 C-int and C-ind

As it was expected, it was observed a wider diversity of n-grams in c-int than in c-ind. Not only that, it was observed that the grams were not only higher in variety, but in its individual frequency as well. Such difference may be attributed to the reliance of students on the task prompt, as it was observed with the analysis of the word frequency list.

A corresponding behavior to the word frequency list was also noted when topic-related grams are considered. We observe a stronger tendency of topic-related words and n-grams in the c-int subcorpus than in the c-ind. From a list of the fifty most frequent n-grams, 16 of c-ind are not topic-related, while for c-int this number falls to only one (table 27). Since there is only one non-topic related chunks for c-int with the adopted cut-off frequency, the analysis of for this section will be focused on some frequent non-topic related c-ind n-grams and the usage of topic-related n-grams in c-int.

The n-grams "(I) think (that) the", "in my opinion" and "i believe that" are commonly used in c-ind to introduce or to conclude the topic, as we note that the most frequent collocate with the expressions are the words "media" and "first impression" and "religion" (figure 5), being all of them related to the essay theme. When accessing the full text, these expressions are either placed at the beginning or at the end of the essay, revealing its introductory and concluding functions.

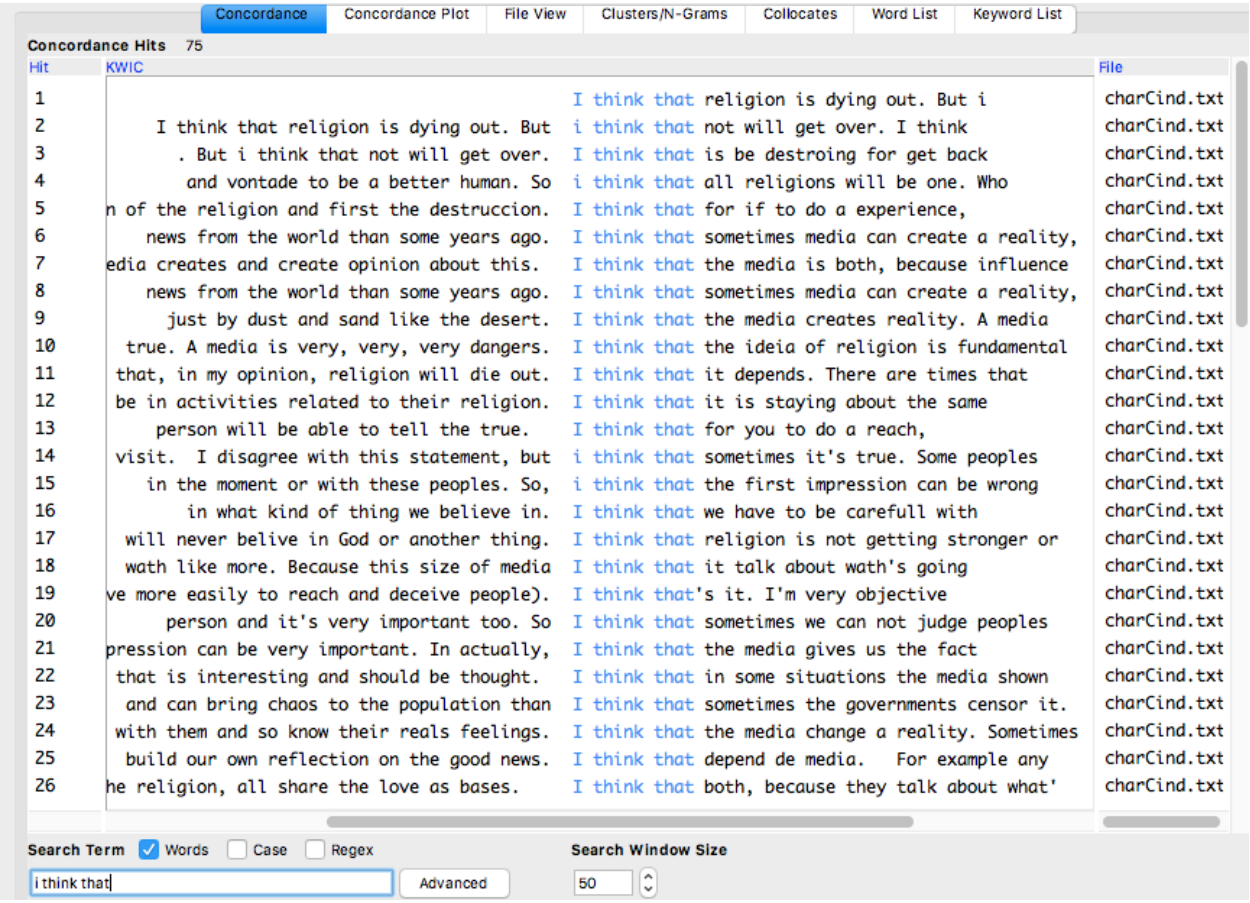
Table 27 - Distribution of the c-ind most frequent n-grams that are not topic-related and their frequency in c-int

	c-ind		**c-int**	
a lot of	104	0,512	77	0,611
i think that	64	0,315	9	0,071
we need to	50	0,246	4	0,031
i think the	46	0,226	0	0
in my opinion	41	0,202	0	0
we have to	38	0,187	4	0,031
is very important	37	0,182	12	0,095
in the past	27	0,133	14	0,111
think that the	27	0,133	1	0,007
we can see	25	0,123	25	0,198
i think that the	24	0,118	0	0
it is not	24	0,118	3	0,023
and it is	23	0,113	12	0,095
i believe that	23	0,113	1	0,007
there is a	23	0,113	8	0,063
in order to	22	0,108	3	0,023

As for the topic-related grams adopted by learners in the integrated tasks, we observe that most expressions are obtained from the infograph and used with no alterations from the prompt, as it is the case with "(of the) world's coffee production" (example 17).

1. *The U.S. imports 1/3 of the **world's coffee production**, that´s more than $4 billion in coffee a year.* (c-int)
2. *The caffeine is the **psychoactive drug** most popular of the U.S and your consume in coffee represents 75%.* (c-int)
3. *The caffeine is the **most popular psychoactive drug** in the world* (c-int)
4. *That is show caffeine is **psychoactive drug** most popular of the world.* (c-int)
5. *The coffee is the drink more consumed in the world.* (c-int)

Figure 6 - Screenshot of the concordance lines in c-ind with the 3-gram “I think that”



However, other n-grams are derived from the essay prompt and graphs with some alterations. The expression “the world’s most popular psychoactive drug” given in the task infograph (Appendix B), for instance, is paraphrased by learners as examples 18, 19, 20 and 21 illustrate. The examples shed some light on not only how learner of A2 level produce their own paraphrases, but also on some frequent language misuse, as of the use of the determiner “the”. The expression “the coffee is” and “the caffeine is”, for instance, are among the most frequent n-grams in c-int, occurring 93 and 45 times respectively.

5.3.3 C-sum and C-bawe

C-sum is a very small corpus, composed of only 14,921 words, which demands the analysis to be taken cautiously. Nonetheless, some assumptions could be made based

on the most frequent non-topic related n-grams, as well as on on a comparison made with c-bawe's list.

A first observed point is related to the n-gram diversity in both corpora, which is far wider in c-bawe. If we take the frequencies of the n-grams with the greatest occurrence in each both subcorpora, for instance, we observe that the one in c-sum ("dna strand breaks") is 3.46 times higher than in the one in c-bawe ("the use of"). Such result suggests that there is a need to increase learners' lexical diversity, who, in this study, demonstrated a strong reliance on the text to be summarised.

Regarding the use of non-topic related grams by learners, a correspondent behaviour was observed. The following grams - "as well as", "the number of", "in order to", "one of the" and "the presence of" - were observed in both subcorpora list. When the use in context is investigated (example 22), we notice the grams convey the same meaning in both corpora, despite some structure inaccuracies found in the learners' production.

6. *This compounds can capture free radicals and act as metal chelating reagents, as well as induce the production of antioxidant enzymes.* (c-sum)

One n-gram that has proven frequent in c-bawe, but with only two occurrences in c-sum was "due to the". Once the gram "due to the" is on the second place on c-bawe's list and almost inexistent in c-sum, it is suggested that learners are not familiarised with the expression commonly adopted in the academic discourse.

This chapter has presented some possible analyses made with the method developed for this study. The chapter that follows moves on to consider how the research questions were addressed and to present an evaluation of these study objectives.

Figure 7 - Plot of POS frequency in all three subcorpora

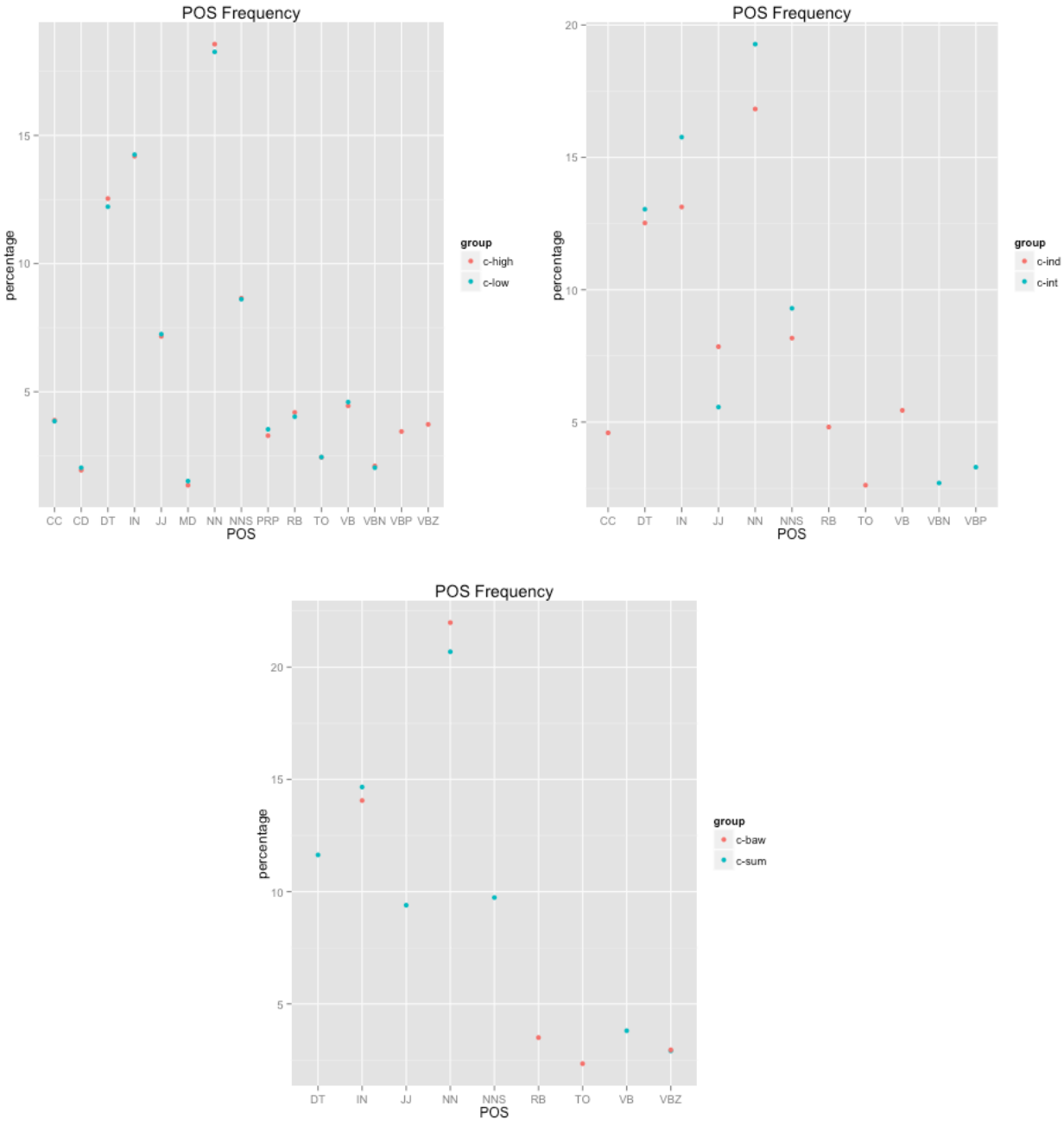
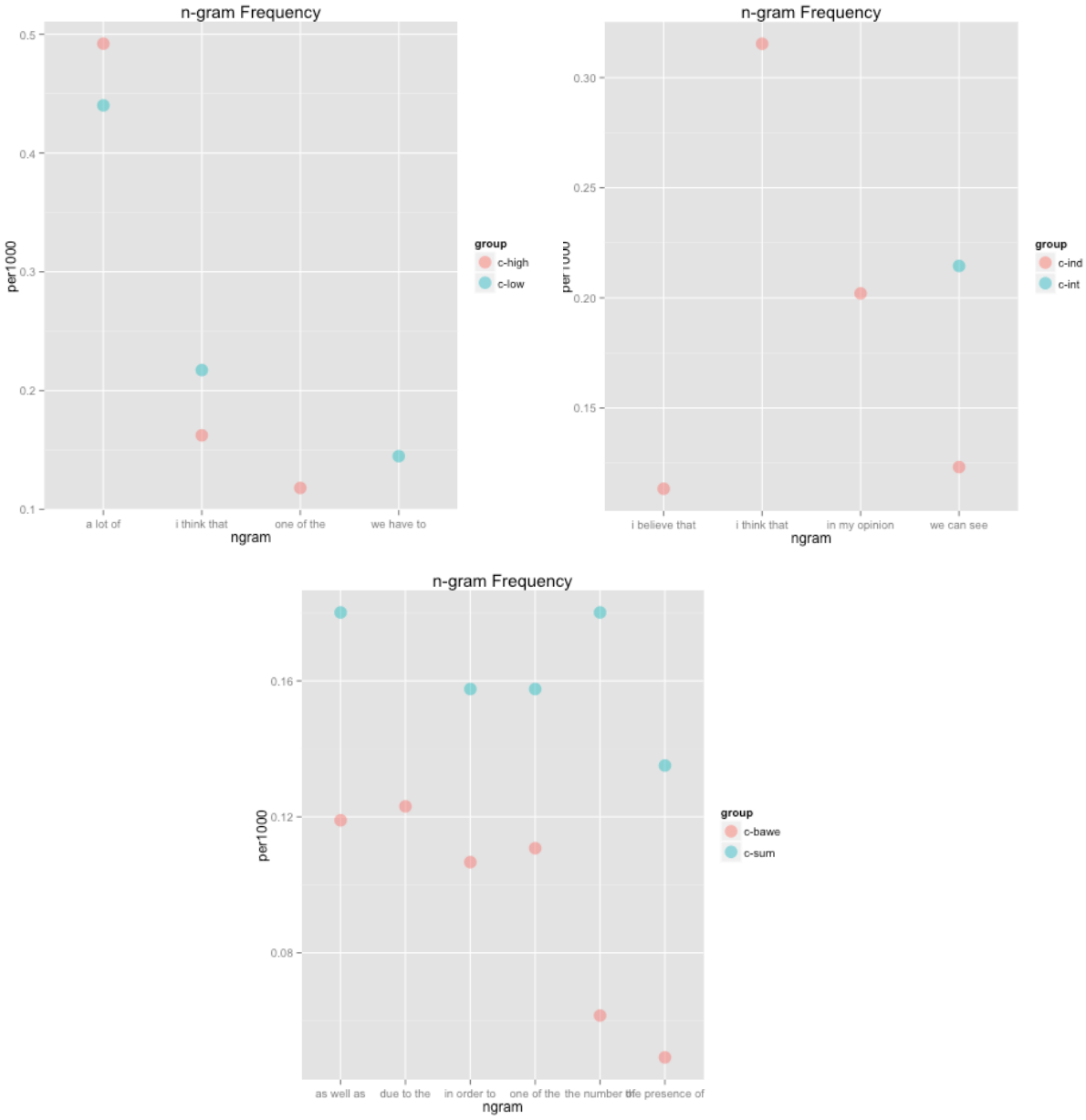


Figure 8 - Plot of some of the most frequent n-grams for each group



6 CONCLUSION

In this thesis, an outline of the procedure required to compile, clean and process the CorIsF dataset was presented. Establishing well-defined criteria for data collection has demonstrated to be useful for the further analysis, once it considerably reduces cleaning time, which would otherwise be necessary. Furthermore, integrating the learners' profile and their textual production made the retrieval of specific batches, or subcorpora, easier and faster.

To deal with the data present in CorIsF, it was first necessary to develop a cleaning process. In this process, it was created functions to delete information from learners who do not wish to participate in the research, to anonymize the collaborators, and to delete irrelevant information from the dataset. These functions made the cleaning process automatic, so that the data can be continuously cleaned as it grows.

A main contribution of this work was to set a framework to collect and keep learner data in a tidy format. In this way, once the data goes through the cleaning process, it can easily be subset according to the research needs. Making the extraction of subcorpora from the dataset before applying the investigation techniques has proved to reduce processing time. Once the subcorpora are set, the investigation functions here developed can then be applied. Additionally, these batches can also be extracted as .txt so that they can be analysed with more user-friendly interfaces such as AntConc and WordSmiths Tool.

To what language analysis is concerned, differences among the five subcorpora derived from CorIsF (c-high, c-low, c-int, c-ind and c-sum) were observed and described in chapter five. The analyses were restricted due to the small size of the subcorpora. However, as the data grows, new data analytics can be implemented in order to assist further investigations.

Another major contribution of this work is the creation of the CorIsF, which is available for research purposes[43]. The corpus, which is composed of 145,043 words from four different institutions, as well as the scripts used for its processing can be used to address different studies on the written production of Brazilian learners of English.

[43] The scripts and the data used in this study are available at: https://github.com/andressarg/thesis

REFERENCES

ANTHONY, L. A critical look at software tools in corpus linguistics. **Linguistc Research**, v. 30, n. 2, p. 141–161, 2013.

ANTHONY, L. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/ 2014.

BAAYEN, R. H. **Analyzing linguistic data**. New York: Cambridge University Press, 2008.

BERNARDINI, S. Exploring New Directions for Discovery Learning. Kettemann, 2002.

BIBER, D. Experimental evidence concerning the acquisition of a Somali discourse rule. In Proceedings of the First International Congress of Somali Studies, ed. by H. M. Adam and C.L. Geshekter, 398-423. Chico, CA: Scholars Press. 1992.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at: Lexical bundles in university teaching and textbooks. **Applied Linguistics**, p. 371–405, 2004.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S., FINEGAN, E. Longman Grammar of Spoken and Written English. Essex: Longman, 1999.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus Linguistics: Investigating Language Structure and Use**. Cambridge: Cambridge University Press, 1998.

BRAND, C.; GÖTZ, S. Fluency versus accuracy in advanced spoken learner language: A multi-method approach. **International Journal of Corpus Linguistics**, v. 16, n. 2, p. 255–275, 2011.

CALLIES, M.; PAQUOT, M. An interview with Yukio Tono. **International Journal of Learner Corpus Research**, v. 1, n. 1, p. 160–171, 2015.

CHANDLER, B. Longman Mini-Concordancer [Computer Software]. Harlow, UK: Longman Press. 1989.

COBB, T. Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. Canadian Modern Language Review/ La Revue canadienne des langues vivantes, v. 59, n. 3, p. 393–424, 2006.

CORTES, V. Lexical bundles in published and student disciplinary writing: Examples from history and biology, English for specific purposes 23, 4, 397-423, 2004.

COTOS, E. Enhancing writing pedagogy with learner corpus data. **ReCALL**, v. 26, n. Special Issue 02, p. 202–224, 2014.

COXHEAD, A. A new academic word list. TESOL Quarterly, 34, 213–238. 2000.

CUMMING, A., KANTOR, R., BABA, K., ERDOSY, U., EOUANZOUI, K., & JAMES, M. Analysis of discourse features and verification of scoring levels for independent and discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL. Princeton, NJ: Educational Testing Service. integrated prototype written tasks for the new TOEFL. Princeton, NJ: Educational Testing Service. 2006.

DAVIES, M. The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/. 2008-

DASU, T., JOHNSON, T. Data Quality, in Exploratory Data Mining and Data Cleaning, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

DUTRA, D. P.; BERBER, T. Referential expressions in English learner argumentative writing. p. 117–127, 2013.

ELLIS, N. C.. Implicit and Explicit Learning of Languages. San Diego, CA: Academic Press. 19994.

EVISON, J. What are the basics of analysing a corpus? In: O'KEEFFE, A.; MCCARTHY, (Ed.) M. New York: Routledge, 2010.

FEINERER, I.; HORNIK, K.; MEYER, D. Text Mining Infrastructure in R. **Journal of Statistical Software**, v. 25, n. 5, 2008.

FLOWERDEW, L. Corpora and Language Education. Basingstoke: Palgrave Macmillan. International Journal of Corpus Linguistics 19(1), 2012.

GARDNER, D., DAVIES, M. A new academic vocabulary list. Applied Linguistics, 35: 1- 24. 2013.

GARSIDE, R., SMITH, N. A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121, 1997.

GILQUIN, G, DE COCK, S. Errors and disfluencies in spoken corpora: Setting the scene. **International Journal of Corpus Linguistics** 16(2): 141-172, 2011.

GILQUIN, G.; GRANGER, S.; PAQUOT, M. Learner corpora: The missing link in EAP pedagogy. **Journal of English for Academic Purposes**, Vol. 6, no. 4, p. 319-335 2007.

GRANGER, S. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In: AIJMER K.; ALTENBERG B.; JOHANSSON M., Languages in Contrast. Text-based cross-linguistic studies (Lund Studies in English; 88), Lund University Press: Lund, p. 37-51, 1996.

GRANGER, S. The computer learner corpus: a versatile new source of data for SLA research. **Learner English on Computer**, p. 3–18, 1998.

GRANGER, S. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. **TESOL Quarterly**, Vol. 37, no. 3, p. 538-546, 2003.

GRANGER, S. Corpora and Language Teaching. In: AIJMER, K. (Ed.). . Amsterdam / Philadelphia: John Benjamins Publishing Company, 2009. p. 13.

GRANT, J., GINTHER, A. Using computer-tagged linguistic features to describe L2 writing differences. **Journal of Second Language Writing**, 9, pp.123-145, 2000.

GRIES, S. T. **Statistics for Linguistics with R: A Practical Introduction**. 2. ed. Berlin: De Gruyter Mouton, 2013.

GRIES, S. T. H. **Quantitative Corpus Linguistics With R: A Practical Introduction**. New York: Routledge, 2009.

HARDIE, A. CQPweb — combining power, flexibility and usability in a corpus analysis tool. **International Journal of Corpus Linguistics**, v. 17, n. 3, p. 380–409, 2012.

HASKO, V. Capturing the Dynamics of Second Language Development via Learner Corpus Research: A Very Long Engagement. **Modern Language Journal**, v. 97, n. SUPPL.1, p. 1–10, 2013.

HAWKINS, J. A.; BUTTERY, P. Criterial Features in Learner Corpora: Theory and Illustrations. **English Profile Journal**, v. 1, n. 01, p. e5, 24 set. 2010.

HYLAND, K. As can be seen: Lexical bundles and disciplinary variation. **English for Specific Purposes**, v. 27, n. 1, p. 4–21, jan. 2008.

JUKNEVIČIENĖ, R. Lexical bundles in learner language: Lithuanian learners vs. native speakers. **Kalbotyra**, v. 61, n. 3, p. 61–72, 2009.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition**. 2. ed. Upper Saddle River: Prentice Hall, 2008.

HORNIK, K. openNLP: Apache OpenNLP Tools Interface. R package version 0.2-3. http://CRAN.R-project.org/package=openNLP, 2014.

KIRK, A. **Data Visualisation: A Successful Design Process**. Birmingham: Packt Publishing, 2012.

MCENERY, T.; HARDIE, A. **Corpus linguistics: Method, theory and practice**. Cambridge: Cambridge University Press, 2012.

MEUNIER, F. Learner Corpora and English Language Teaching: Chekup Time. **Anglistik: International Journal of English Studies**, v. 21, n. 1, p. 209–220, 2010.

MUKHERJEE, J.; ROHRBACH, J.-M. Rethinking Applied Corpus Linguistics from a Language-pedagogical Perspective: New Departures in Learner Corpus Research. **Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop**, p. 205–232, 2006.

NESSELHAUF, N. How to Use Corpora in Language Teaching. In: SINCLAIR, J. (Ed.). . Amsterdam: John Benjamins Publishing Company, 2004. p. 125.

NESSELHAUF, N. **Collocations in a Learner Corpus**. [s.l: s.n.]. v. 29

PRAVEC, N. A. Survey of learner corpora. **ICAME Journal**, v. 26, n. 1, p. 81–114, 2002.

SIMPSON-VLACH, R.; ELLIS, N. C. An Academic Formulas List: New Methods in Phraseology Research. **Applied Linguistics**, v. 31, n. 4, p. 487–512, 12 jan. 2010.

SINCLAIR, J. **Corpus, Concordance, Collocation**. Oxford: Oxford University Press, 1991.

STAPLES, S. et al. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. **Journal of English for Academic Purposes**, v. 12, n. 3, p. 214–225, set. 2013.

VINCENT, B. Investigating academic phraseology through combinations of very frequent words: A methodological exploration. **Journal of English for Academic Purposes**, v. 12, n. 1, p. 44–56, mar. 2013.

WICKHAM, H. Journal of Statistical Software. **Journal of Statistical Software**, v. 59, n. 10, 2014.

APPENDIX A – Tasks used for CorIsF data collection

Test	CEFR	Task Type	Prompt
Coffee	A1	Integrated	The infograph below presents some information about coffee. Organise the information by selecting and reporting the main features, and make comparisons where relevant. https://docs.google.com/forms/d/1E4S1vIOKF0LOqDYfyusJFTBa-Fb0QeMKcVIxurklyF4/viewform
News and bias	A2	Independent	Do you think the media creates reality? Or does the media talk about what's going on? Or both? https://docs.google.com/forms/d/1GCU1p2wBWnd-raEvIQkitZ7h4TdtNmrYowxHf7yw4YU/viewform
Religion	B1	Independent	It appears that religion has been around in one form or another for most of human history. Do you think it is getting stronger, dying out, or staying about the same level of importance? Why? https://docs.google.com/forms/d/12_H2wgMJurt20qp-sshWAdctnWVGsLEDF3AkVMYPchk/viewform
Love, marriage and divorce	B1	Integrated	The graph below gives information about the percent of American adults ever divorced, by age and gender. Summarize the information by selecting and reporting the main features. Make comparisons where relevant. https://docs.google.com/forms/d/16QqFxmL-wplZUFBnvxHifQEUBqK0t6X04ZttAhmWuKs/viewform
Thought and Mind	B1	Independent	Do you agree or disagree with the following statement? The first impression is the most important one. Use specific reasons and examples to support your opinion. https://docs.google.com/forms/d/1pdbdIPukoBJ3F2KEnXNxFZTRFcZuTv5gAhOseIvE6RA/viewform

Fire	B2	Integrated	The graph below shows a summary of prescribed burn acres and the frequency of fires from 1985 until 2013. Organise the information by selecting and reporting the main features, and make comparisons where relevant. https://docs.google.com/forms/d/1RKAKLAfAdcaW38tRh9V_yFJe6pUZvYe-NflZz410AtQ/viewform
Languages	B2	Independent	Do you agree or disagree with the following statement? Children should begin learning a foreign language as soon as they start school. Use specific reasons and examples to support your position. https://docs.google.com/forms/d/1V6X-gELPU8d9EefWab3v9lnd5ODSzog9QXCVBNCH-UQ/viewform
Water	B2	Independent	In many countries all over the world there is a serious shortage of water. What are the causes of, and possible solutions to, the scarcity of water resources? https://docs.google.com/forms/d/19eAafsyBymsGssWLW_PAdabamAjQlK4N0yW7K5baubs/viewform

APPENDIX B – Integrated Task image (Coffee)

Source: http://www.designinfographics.com/food-infographics/a-coffee-crazed-america

APPENDIX C – Part-of-speech tagset

	TAG	DESCRIPTION
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb