UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

JESSICA MARIA DA SILVA QUEIROZ

THE GRAMMATICAL COMPLEXITY OF ENGLISH NOUN PHRASES IN BRAZILIAN LEARNER'S ACADEMIC WRITING: A CORPUS-BASED STUDY

BELO HORIZONTE
2019

JESSICA MARIA DA SILVA QUEIROZ

THE GRAMMATICAL COMPLEXITY OF ENGLISH NOUN PHRASES IN BRAZILIAN LEARNER'S ACADEMIC WRITING: A CORPUS-BASED STUDY

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 do título de MESTRE em Linguística Teórica e Descritiva

Área de Concentração: Linguística Teórica e Descritiva
Linha de Pesquisa: Estudos Linguísticos baseados em Corpora

Orientadora: Profa. Dra. Deise Prina Dutra

Belo Horizonte
Faculdade de Letras da UFMG
2019

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

Q3g

Queiroz, Jessica Maria da Silva.
The grammatical complexity of English noun phrases in Brazilian learner's academic writing [manuscrito] : a corpus-based study / Jessica Maria da Silva Queiroz. – 2019.
138 f., enc. : il., grafs., color., tabs., p&b.

Orientadora: Deise Prina Dutra.

Área de concentração: Linguística Teórica e Descritiva.

Linha de pesquisa: Estudos Linguísticos Baseados em Corpora.

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

Bibliografia: f. 113-119.

Anexos: f. 120-137.

1. Língua inglesa – Sintagma nominal – Teses. 2. Língua inglesa – Gramática – Teses 3. Linguística de corpus – Teses. 4. Redação acadêmica – Teses. I. Dutra, Deise Prina. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD : 425

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

PosLin

FOLHA DE APROVAÇÃO

THE GRAMMATICAL COMPLEXITY OF ENGLISH NOUN PHRASES IN BRAZILIAN LEARNER'S ACADEMIC WRITING: a corpus-based study

JESSICA MARIA DA SILVA QUEIROZ

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA TEÓRICA E DESCRITIVA, linha de pesquisa Estudos Linguísticos Baseados em Corpora.

Aprovada em 25 de fevereiro de 2019, pela banca constituída pelos membros:

Deise Prina Dutra

Prof(a). Deise Prina Dutra - Orientador
UFMG

Prof(a). Heliana Ribeiro de Mello
UFMG

Prof(a). Patrícia Pereira Bértoli
UERJ

Belo Horizonte, 25 de fevereiro de 2019.

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to:

my supervisor, Dr. Deise Dutra, for her teachings, trust, and understanding.

FAPEMIG, for the financial support given for the past two years.

all UFMG faculty members, for their dedication and influence on my career and research.

our research group, for the constructive discussions.

my dear graduate friends, Clarice, Isabelle, João, Flávia, Amália, Marcus, Matheus, Cecília, Filipe, for their true friendship in the good and not so good days.

my friends Clarice and Euller, for their friendship and great help with this research.

my adolescence friends, Ana, Marianna, Carolina, for their existence in my life.

my family, for everything. Always.

AGRADECIMENTOS

Eu gostaria de expressar a minha imensa gratidão:

à minha orientadora, profa. Deise Dutra, pelos ensinamentos, pela confiança e compreensão.

à FAPEMIG, pelo apoio financeiro dado nos últimos dois anos.

aos professores da UFMG, pela dedicação e influência sobre minha carreira e pesquisa.

ao nosso grupo de pesquisa, pelas discussões construtivas.

aos amigos do POSLIN, Clarice, Isabelle, João, Flávia, Amália, Marcus, Matheus, Cecília, Filipe, pela amizade sincera nos dias bons e nos não tão bons assim.

aos amigos Clarice e Euller, pela amizade e pela grande ajuda com o presente trabalho.

às minhas amigas de adolescência, Ana, Marianna, Carolina, pela existência em minha vida.

à minha família, por tudo. Sempre.

“The more you learn, the more you recognize that while things are not as blindingly complex as you first thought, neither are they as simple as you then hoped they would be.”
– Wayne Booth, Gregory Colomb, Joseph Williams, *The craft of research*

“We were scared, but our fear was not as strong as our courage.”
– Malala Yousafzai, *I am Malala*

ABSTRACT

There is an increasing number of research which is interested in investigating the grammatical complexity of academic writing in English. Some examine first language (L1) English writers' texts (e.g. GRAY, 2015; BIBER; GRAY, 2016; STAPLES *et al.*, 2016) while others analyze texts written by second language (L2) English learners (e.g. PARKINSON; MUSGRAVE, 2014; NITSCH, 2017; ANSARIFAR *et al.*, 2018). These studies investigate either or both pre- and post- noun modification arguing about how such devices may help in the elaboration or compression of information (GRAY, 2015). Noun premodification, in particular, has shown to serve the function of adding new information to a noun head in a way that makes the phrase more economical and faster to read (BIBER; GRAY, 2010, 2016). This thesis, then, is a corpus-based descriptive study which explores the grammatical complexity of the English noun phrase (NP) in two subcorpora (general topic essays: 46 texts and 18678 words – specific topic essays: 68 texts and 32509 words) of the *Corpus of English for Academic Purposes* (CorIFA), a corpus of Brazilian university students' writings. The study examines the NPs found in students' argumentative essays written in an upper intermediate English for Academic Purpose course and categorizes them according to their constituency into simple and complex NPs. To do so, we automatically parsed the texts and extracted the word groups parsed as NPs. We manually categorized them into simple and complex NPs, identifying their constituents and adding them to different subcategories, such as simple NP with a determiner and head noun or complex NP with prepositional phrases (PPs) as postmodifiers. The investigation reveals that Brazilian writers use more complex NPs than simple ones and, particularly, NPs with premodifying adjectives and NPs with postmodifying PPs. All in all, our corpus-based research shows that upper intermediate Brazilian university students are capable of producing structurally complex and compressed NPs, but we argue for more research on the grammatical complexity of NPs with a larger learner corpus and across learners' various disciplines as well as across registers.

Keywords: English noun phrase, grammatical complexity, academic writing, learner corpus

RESUMO

Há um número crescente de pesquisas interessadas em investigar a complexidade gramatical da escrita acadêmica em inglês. Algumas examinam os textos de escritores nativos em inglês (e.g. GRAY, 2015; BIBER; GRAY, 2016; STAPLES *et al.*, 2016) enquanto outras analisam textos escritos por aprendizes de inglês (e.g. PARKINSON; MUSGRAVE, 2014; NITSCH, 2017; ANSARIFAR *et al.*, 2018). Estes estudos investigam uma ou ambas pré- e pós-modificação de substantivos discutindo como estes elementos sintagmáticos podem ajudar na elaboração ou compressão de informação (GRAY, 2015). A pré-modificação, particularmente, mostrou servir a função de adicionar informações novas ao núcleo nominal de forma a tornar o sintagma mais econômico e rápido para leitura (BIBER; GRAY, 2010, 2016). A presente dissertação, então, é um estudo descritivo baseado em *corpus* que explora a complexidade gramatical do sintagma nominal (SN) em inglês em dois subcorpora (redações de tópicos gerais: 46 textos e 18678 palavras – redações de tópicos específicos: 68 textos e 32509 palavras) do *Corpus do Inglês para Fins Acadêmicos* (CorIFA), um *corpus* de escrita de alunos universitários brasileiros. O estudo examina os SNs encontrados nas redações argumentativas escritas pelos alunos que participam de uma turma intermediária superior de Inglês para Fins Acadêmicos e os categoriza a partir de seus constituintes em SNs simples ou complexos. Para isso, nós utilizamos um método automático de análise sintática e extraímos os grupos de palavras analisados como SNs. Nós categorizamos estes SNs em simples e complexos, identificando seus constituintes e adicionado os sintagmas em subcategorias diferentes, por exemplo, SN simples com determinante e núcleo ou SN complexo com sintagma preposicional como pós-modificador. O estudo revela que os escritores brasileiros usam mais SNs complexos do que SNs simples e, em especial, SNs com adjetivos como pré-modificadores e SNs com sintagmas preposicionados como pós-modificadores. Portanto, a nossa pesquisa baseada em *corpus* mostra que os alunos universitários brasileiros com proficiência intermediária superior são capazes de produzir SNs estruturalmente complexos e comprimidos, mas sugerimos mais pesquisas sobre a complexidade gramatical de SNs que tenham um corpus de aprendiz maior e que comparem diferentes níveis acadêmicos e registros.

Palavras-chave: sintagma nominal em inglês, complexidade gramatical, escrita acadêmica, corpus de aprendiz

LIST OF FIGURES

LIST OF GRAPHS

LIST OF TABLES

LIST OF ABBREVIATIONS

AdjP	Adjective phrase
BAWE	Corpus of British Academic Written English
Br-ICLE	Brazilian Portuguese subcorpus of the International Corpus of Learner English
CorIFA	Corpus of English for Academic Purposes
EAP	English for Academic Purposes
L1	First language
L2	Second language
LOCNESS	Louvain Corpus of Native English Essays
NP	Noun phrase
POS	Part-of-speech
PP	Prepositional phrase
SLA	Second Language Acquisition
UFMG	Federal University of Minas Gerais

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Writing constitutes a major part of academic activities. This may be because researchers can best communicate, access, share, and construct scientific knowledge through written papers. However, writing successfully to a scientific community means that we need to know how to use our community's scientific discourse and that requires training it over time as "one is not born with it and does not acquire it automatically or with ease" (MONTGOMERY, 2013, p. 108).

In our globalized world, researchers wanting to participate in their chosen disciplinary communities should not only become competent writers but also competent writers in English. By the end of the 20th century, English became the global language as well as the global language of science. This is undeniable when we learn that "over 90% of international scientific communication in every form, throughout the entire globe" is done in English (MONTGOMERY, 2013, p. 168) and that "there looks to be little chance of this changing anytime soon" (MONTGOMERY, 2013, p. 169). As a consequence, more and more studies will be published in English, requiring native and non-native speakers to master their community's scientific discourse.

Mastering scientific written discourse in English mostly means gaining fluency in the writing conventions of that particular disciplinary community (HYLAND, 2014). This is a matter of using technical vocabulary appropriately and structuring arguments in a certain way, but, more than that, it is a matter of using language according to certain recurrent grammatical patterns shared by writers. It involves being technical and clear while following the ways in which information is codified in each academic research field. Such codification is oftentimes grammatically complex as present-day academic discourse has its own structuring of information, not often accessible to the lay reader.

Recently, the grammatical structure in academic written texts has been shown to be predominantly structurally compressed (BIBER; GRAY, 2010, 2016). That means that grammatical complexity occurs more frequently in a phrasal level and that the stereotypical assumption that academic writing is complex in a clausal level is actually not true for all disciplines indistinctly. Biber and Gray's (2016) research on professional academic writing has actually demonstrated a historical shift from clausal to phrasal complexity as information started being packed most commonly in complex noun phrases (NPs) rather than in dependent clauses, making academic English complex by means of a different set of syntactic forms that are more economical and faster to read by expert readers (BIBER; GRAY, 2010, 2016).

This recent descriptive work is of paramount importance to several research fields like Biber and Gray (2016) themselves suggest, and not only to Second Language Acquisition (SLA), which has been the applied linguistics area most concerned with grammatical complexity[1] (PALLOTTI, 2015). Considering the tradition of English for academic purposes (EAP) programs for learners of English as a second language (L2) in Brazil and in other countries (SALAGER-MEYER *et al.*, 2016), it seems reasonable that SLA researchers study that. L2 English learners should know the structures most commonly used in academic written texts as well as the importance and purposes of their uses since they need to produce them in order to engage and succeed in their academic communities. Nonetheless, EAP instructors must have access to relevant research in order to create course syllabi and materials containing the appropriate linguistic features that are part of the grammatical style used by researchers in specific disciplines.

To provide students with the necessary input concerning grammatical complexity, it is then necessary to have more descriptive studies on the topic. First and foremost, it seems crucial to have research on published expert writing (e.g. BIBER *et al.*, 1999; BIBER; GRAY, 2016; GRAY, 2015), but it is equally important to have studies that investigate learner writing (e.g. PARKINSON; MUSGRAVE, 2014; NITSCH, 2017; ANSARIFAR *et al.*, 2018). Corpus-based studies on expert academic English, particularly on disciplinary and genre variations, could provide the necessary and salient data to be taught to varied EAP groups, but studies on learner language could also shed light on specific features that should be targeted immediately. Such seems to be case of simple and complex NPs, which are grammatical units not usually discussed with EAP students yet extremely important to the grammatical complexity of academic writing.

From our personal experience, Brazilian students are not aware of all the patterns and complexities of NPs as we linguists are. To confirm (or not) this observation and willing to fill

[1] Grammatical complexity has different definitions for different researchers. In this study, we will focus on the grammatical complexity of NPs. It involves the syntactic constituency of NPs, in which NPs containing at least one modifier are considered complex. A detailed discussion of grammatical complexity is given in Chapter 2.

the gap in research linking learner corpora and grammatical complexity, we propose a corpus-based investigation of the grammatical complexity of NPs in Brazilian learners' academic written production, in the light of the work developed by Biber *et al.* (1999) and Biber and Gray (2016). We expect to examine all noun-headed NPs produced by L2 learners, either simple or complex, and in all the configurations presented in previous corpus-based research (cf. BIBER *et al.*, 1999), for instance, adjective + noun or noun + prepositional phrase (PP). The corpus to be analyzed is a subcorpus of the *Corpus of English for Academic Purposes* (CorIFA), a learner corpus composed of a wide range of academic texts written by Brazilian university students at the Federal University of Minas Gerais (UFMG), which will be carefully described in Chapter 3.[2]

1.1 Research objectives

The general research objective of this thesis is to understand Brazilian university students' writing in English in terms of the grammatical complexity of NPs produced by these writers. More specifically, then, upper intermediate argumentative essays will be examined through the analysis of NP structures, as they are presented in reference grammar books (e.g. BIBER *et al.*, 1999). For that purpose, the specific objectives of this research are as follow:

- Quantify the simple and complex NPs used by Brazilian university students in their written essays;
- Identify the types of simple and complex NPs produced;
- Determine the types of NPs most commonly used;
- Detect the pre- and postmodifiers most frequently used in complex NPs;
- Discuss the NP structures found in the research corpus.

[2] Despite translating the names of institutions, programs, and corpora into English, their acronyms and initialisms, such as CorIFA and UFMG, will be kept as in Portuguese to facilitate their recognition throughout the text.

Having these objectives in mind, the results reported in previous studies , and our intuition as English instructors (see Chapters 2 and 3 for more details), this thesis will test the following hypotheses:

1) Brazilian upper intermediate learners use more simple NPs than complex NPs in their English written production;
2) Brazilian upper intermediate learners produce more NPs with postmodifier(s) than NPs with premodifier(s) or NPs with both pre- and postmodifiers;
3) Brazilian upper intermediate learners use more adjectives as NP premodifiers;
4) Brazilian upper intermediate learners use more prepositional phrases (PPs) as NP postmodifiers.

1.2 Outline of the chapters

This thesis is structured in five chapters. The present chapter has provided a background to this study along with the justification for the research, its objectives, and its hypotheses.

Chapter 2 provides a theoretical framework for the study by reviewing current literature on the NP as a linguistic construct. A historical overview and definitions of concepts necessary to this study, such as definitions of the NP and its constituents, are given in detail. It should be clear in that chapter what types of NPs are being investigated and what our operational definition of grammatical complexity is.

Chapter 3 explains the theoretical framework behind the methodological choices made in this study. It presents the research corpus organized for the study and describes the methods for automatic data retrieval and manual data analysis.

Chapter 4 reports the results of the data analysis and examines results quantitatively and qualitatively. It describes and discusses the production of NPs by Brazilian learners, testing the hypotheses proposed previously.

Chapter 5 summarizes the research findings and indicates some of their implications. Limitations to the study and suggestions for future research in this field are considered.

The following chapters should provide a reliable descriptive and corpus-based study to be consulted and replicated by colleagues who use learner corpora in their research. It is also expected that it can inform future research that aims at teaching the complexity of NPs in academic writing and helping Brazilian learners be successful in their written production in English so as they can become members of their scientific communities.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces the key constructs and concepts that guide the present thesis. As this research is a corpus-based description of the use of English NPs by upper intermediate Brazilian university learners, some linguistic constructs previously defined by linguistic theory should help us find the patterns of NPs used in their essays. Similarly, literature that could shed light on the discussion of grammatical complexity would also be useful as it is fundamental to clarify our perspective on the topic under analysis.

Section 2.1 gives a brief historical overview of the NP as a theoretical construct in the attempt to explain the relevance it has in linguistic studies and in this thesis. Section 2.2 explores in detail, but not exhaustively, the syntactic forms and functions of the NP based on the description given in English corpus-based grammars, which tend to have a functional perspective of language. The formal aspects of the NP will lead to the differentiation between simple and complex NPs which guides and defines our research analysis. Finally, section 2.3 summarizes some studies that propose the analysis of grammatical complexity of NPs in academic and learner writing and that could be generally compared to the results found in this research.

2.1 Historical overview of the noun phrase

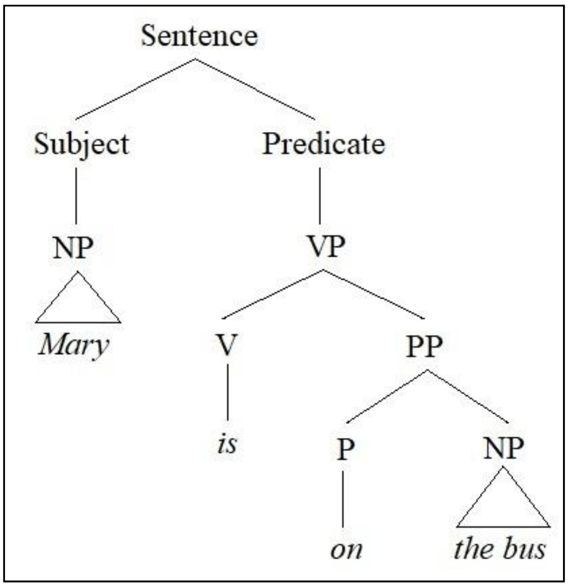
Historically, syntax has been the discipline concerned with analyzing “the way in which words are arranged to show relationships of meaning within (...) sentences” (CRYSTAL, 2005, p. 247). Early on, those who studied human languages knew that these word arrangements that constituted a sentence were not random and attempted to define the possible patterns of arrangement. Nevertheless, it took linguists a long time to propose anything similar to the syntactic methods and studies we have today, which rely on theoretical constructs such as the phrase.

During the 18th and 19th centuries, for instance, the studies that compose what is known as traditional grammar analyzed a sentence as a combination of words rather than as a combination of phrases (PERCIVAL, 1976, p. 230). They defined a sentence as the expression of a thought, so most of their focus when doing syntactic analyses was on identifying the subject and the predicate of sentences, i.e. the entity spoken about and what is being said about it, respectively, which are categories derived from logic (GRAFFI, 2001, p. 113; CRYSTAL, 2005). Consequently, a sentence like *Mary is on the bus* would be segmented into *Mary* as its subject and *is on the bus* as its predicate.

With the advent of structuralism in the 20th century, linguists started avoiding this logic/psychological definition of the sentence and, even though they continued segmenting sentences in subject and predicate, there was an attempt to define syntactic units and relationships following a grammatical and formal perspective (GRAFFI, 2001, p. 167). The American structuralist Leonard Bloomfield was the first one to sketch some notions that could distinguish sentential and non-sentential word groups, affirming that sentences were made of word groups or constituting elements which were organized in a structural order, i.e. a notion similar to that of the phrase (GRAFFI, 2001, p. 196). Although Bloomfield insisted on a formal approach to syntactic analysis, he asserted that the intuition of native speakers was sufficient to the segmentation of sentences into its constituents (GRAFFI, 2001, p. 282).

Bloomfield had a tremendous impact on American linguistic research from 1940 to 1960 and his followers emphasized the necessity of establishing systematic methods to divide the constituents of sentences and word groups (e.g. Pike and Wells cf. GRAFFI, 2001, p. 283-284). They, then, proposed varied graphic representations of the internal structure of sentences, assuming its constituents were binary (GRAFFI, 2001, p. 292-293). One type of representation was the tree diagrams, such as in Figure 2.1, which represents the immediate constituent analysis of the sentence *Mary is on the bus* into subject, predicate, and phrases. The objective was that once a systematic model for analysis was established, it would be possible to replicate it and analyze languages more consistently.

Figure 2.1 – Tree diagram of immediate constituent analysis



Source: Designed by the author, 2019.

Such a model would only become more consistent and be used by several linguists after the 1960s, when linguists became interested in defining more clearly “the theoretical foundations of their discipline” and establishing it as a scientific enterprise (GRAFFI, 2001, p. 309). Noam Chomsky and his followers, drawing from several disciplines but especially from the American structuralism, developed the theory of generative grammar and, for the first time, a replicable framework that allowed a systematic analysis of languages was available, such as the X-bar theory which included the ideas of hierarchy, endocentricity, head, modifiers, binary constituency, and graphic representations in form of tree diagrams (GRAFFI, 2001, p. 366). At that time, studies also established the most common definition of NP used in many works until today, i.e. the NP is a phrase whose constituents are a head noun accompanied or not by determiner(s) and modifier(s) (GRAFFI, 2001, p. 293).

The NP has been since understood as a linguistic phenomena and a relevant construct to linguistic theory. A product of a theory that relies on the notion of an internal structure of the sentence, the NP is central to several syntactic studies as it is in ours. Our revision of its proposal was not intended to be thorough, but it should be possible to understand how the NP became a unit of analysis widely accepted in modern linguistics and whose internal structure is linguistically and functionally important for research on languages and on academic writing.

2.2 The English noun phrase

What follows is a brief description of the English NP in terms of its grammatical features. It is worth reminding the reader that the description here reviewed serves for the English language only and works best in the investigation of NPs when used in written texts. It is also necessary to clarify that I do not intend to present the NP constituents thoroughly, discussing all of its linguistic facets, but I expect to give a brief overview with examples[3] which will help us understand the analysis outlined in Chapter 3 and undertaken in Chapter 4 of this thesis. Let us then start with the formal, purely syntactic aspects of the NP.

2.2.1 The noun phrase constituents

The definition of the English NP regarding its form is practically the same in the three corpus-based grammar books consulted. They were: the *Longman grammar of spoken and written English* (BIBER *et al.*, 1999), the *Cambridge grammar of English* (CARTER; McCARTHY, 2006), and *A comprehensive grammar of the English language* (QUIRK *et al.*, 1985). In that sense, the NP is a widely accepted syntactic unit of analysis, therefore, appropriate to be used in descriptive studies.

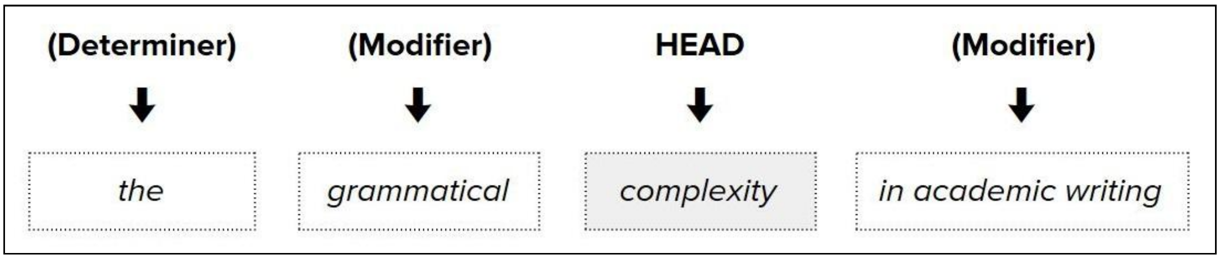
In grammar books, the NP is defined as a phrase that consists of a head noun or pronoun accompanied or not by determiners and/or modifiers. This is the canonical NP structure, but some authors denominate as nominal elements or expressions any other word group that occurs in the position where NPs frequently occur and gives a referential specification, such as adjectives and complement clauses (BIBER *et al.*, 1999; QUIRK *et al.*, 1985). In this study only noun-headed phrases will be taken under analysis because they are the only ones that may carry all simple and complex configurations of NPs.

The NP structure with a noun as head, then, could be formally represented by the sequence in bold in Figure 2.2, where the constituents known as determiner and modifier in between parentheses are optional, in the sense that they may or may not occur together with

[3] Whenever possible, academic or written register examples taken from Biber *et al.* (1999) and/or Biber and Gray (2016) are given.

head nouns, while the constituent head in capital letters is obligatory, meaning that this constituent always occurs in NP structures.

Figure 2.2 – Noun phrase constituents with example



Source: Designed by the author, 2019.

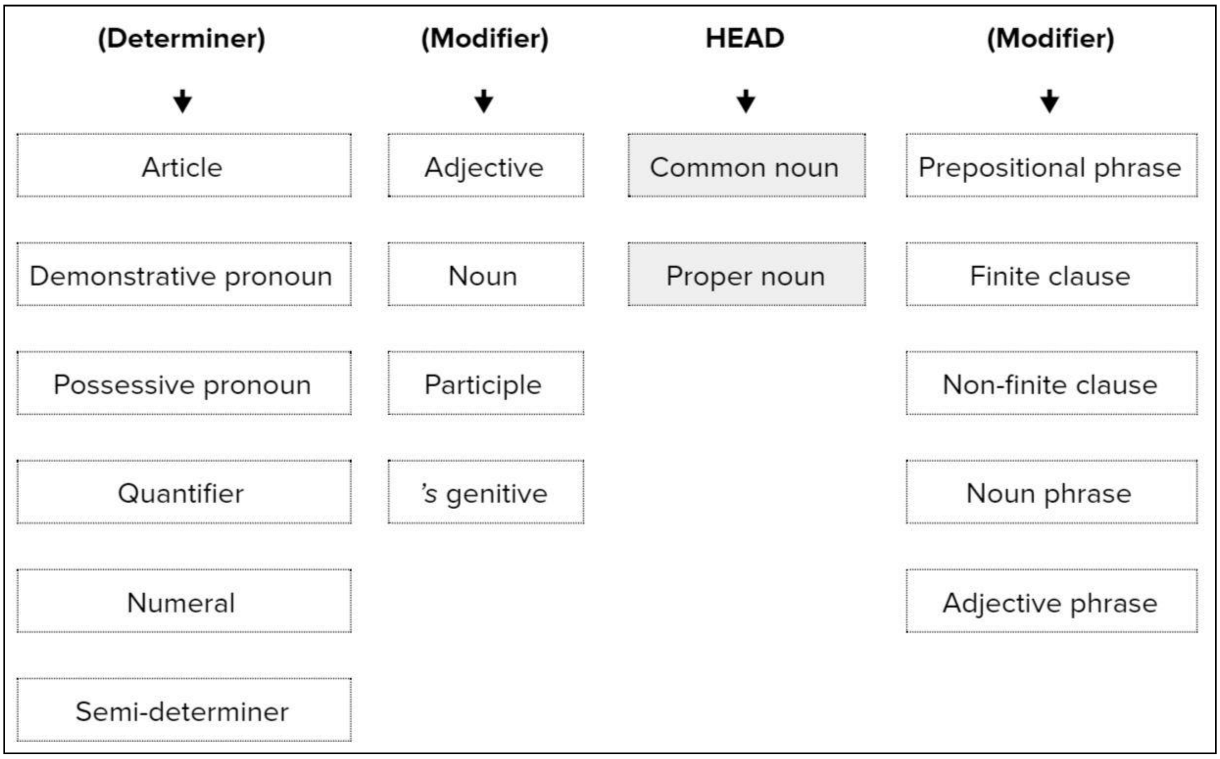
In the example given in Figure 2.2, the sequence of words *the grammatical complexity in academic writing* is a noun-headed NP that has all the constituents mentioned. It can be thus segmented as demonstrated, i.e. the singular noun *complexity* is the head noun, the article *the* is the determiner, and the adjective *grammatical* and PP *in academic writing* are the modifiers, pre- and postmodifiers, respectively. Likewise, the PP *in academic writing* also contains the NP *academic writing*, which could be further segmented into *academic* as premodifier and *writing* as head noun. As it will be seen, several patterns of NPs are possible, because "in principle, there is no limit to the complexity of noun phrases" (BIBER *et al.*, 1999, p. 576).

A quick note should be made regarding the functions of an NP. By that, we mean the various syntactic roles an NP might play in different sentential contexts. An English NP can occur as the subject, the direct, indirect, or prepositional objects, the complement of prepositions, an adverbial, among other uses as peripheral elements in clauses and sentences (BIBER *et al.*, 1999). Recognizing these roles, as they represent the relations of phrases to larger structures, is essential to the interpretation of the NPs in a sentence. Its wide range of roles also shows how significant the NP can be for the structure of the discourse in English, especially for informational registers[4] such as the academic.

[4] In this research, register is defined according to the discussion developed by Biber and Conrad (2009) as a set of specific texts that can be "associated with a particular situation of use (including particular communicative purposes)" (p. 6).

In the subsections that follow, more details are given about each one of the NP constituents: the head noun, its determiners, and its modifiers. Figure 2.3 gives an idea of the possible categories to be found in the position of each one of the NP constituents.

Figure 2.3 – Noun phrase constituents with their possible categories



(Determiner)	(Modifier)	HEAD	(Modifier)
Article	Adjective	Common noun	Prepositional phrase
Demonstrative pronoun	Noun	Proper noun	Finite clause
Possessive pronoun	Participle		Non-finite clause
Quantifier	's genitive		Noun phrase
Numeral			Adjective phrase
Semi-determiner			

Source: Designed by the author, 2019.

2.2.1.1 The noun phrase head

The headword in a phrase structure is the principal, central, or obligatory word of the phrase (BIBER; CONRAD; LEECH, 2002; CRYSTAL, 2005). That means that the headword is the element that defines and characterizes the phrase under analysis and is like an anchor around which all constituents of that word group, if these exist, revolve (QUIRK *et al.*, 1985). It is also "in a manner equivalent to the whole construction of which it is a part" (QUIRK *et al.*, 1985, p. 60). In NPs, it is a noun which prototypically defines and characterizes the word group as an NP.

Semantically, head nouns are words that point out to entities in the external or internal world which are being referred to by the phrase (BIBER *et al.*, 1999). For that reason, NPs are also defined as referring expressions by some authors (CARTER; McCARTHY, 2006). In noun-headed phrases, it is the head noun which explicitly determines the reference to an entity and this noun might have certain grammatical characteristics that reflect the way language users conceptualize that entity (BIBER *et al.*, 1999). These characteristics are used by grammarians in order to classify nouns, for instance, as countable and uncountable.

In this study, it is worth seeing some of the different types of NP heads and their respective examples in order to clearly define the phenomenon under analysis. There are divergences in classification depending on the grammarian's perspective, but we will see the ones that seem necessary for our analysis. Our parameter to decide the relevant types of head nouns will be the similarities found in the reference corpus-based grammar books selected and, to a certain extent, the part-of-speech (POS) tagset used in the Penn Treebank project (TAYLOR *et al.*, 2003; BIES *et al.*, 1995) (see APPENDIX A for the complete tagset and Chapter 3 for more details).

The noun, which is the word class that defines the NP and occurs prototypically as its headword, due to grammatical and semantic reasons, might be classified into different subclasses (QUIRK *et al.*, 1985), such as countable, uncountable, common, proper, collective, unit, quantifying, and species nouns (these are the classes proposed in BIBER *et al.*, 1999, p. 241-257). Considering the types of nouns commonly described in grammars and used in the Penn Treebank, we decided to consider for our analysis only the following two classes (followed by examples where the head nouns are marked in **bold**):

1) Common nouns
    a. *There is no way to tell how old a* ***rock*** *is merely by looking at its minerals.* (ACAD) (BIBER *et al.*, 1999, p. 243, our highlight)

2) Proper nouns
    a. *The traditional view is that* ***Parliament*** *has no power to bind its successors.* (ACAD) (BIBER *et al.*, 1999, p. 248, their highlight)

Common nouns could be categorized into countable and uncountable, meaning that certain nouns can have singular and plural forms as well as be accompanied by definite or indefinite articles, as in *a rock* shown in example (1a) (BIBER *et al.*, 1999). That is a clear distinction made in the Penn Treebank tagset, where common nouns receive the tags NN, when it is a singular or mass noun, or NNS, when it is a plural noun. Proper nouns usually lack the contrast regarding number and definiteness, as can be seen in *Parliament* in example (2a). However, there are particular occurrences where it is possible to use singular or plural forms of proper nouns and that fact is evident in the use of the tags NNP for singular proper nouns and NNPS for plural proper nouns. The other classes found in grammar books, i.e. collective, unit, quantifying, and species nouns, were considered common nouns during our analysis.

The position of NP head could be also filled with words from different word classes. In those cases, these words receive automatically the POS tag correspondent to its class but, in the phrase level, they are parsed as the headword of an NP. They are the pronouns (in most of its subclasses, i.e. personal, possessive, reflexive, demonstrative, reflexive, and indefinite pronouns), adjectives, determiners, numerals, and the existential *there*[5]. Those will not be taken under analysis because they are not as commonly used in the academic register as nouns are (BIBER *et al.*, 1999) and they might not occur as a headword in all the configurations of simple and complex NPs. For example, personal pronouns would not be preceded by the definite article *the*, and that is one possible configuration of a simple NP, i.e. determiner plus head.

A final remark should be made concerning noun-headed NPs. It is possible to coordinate NPs, as in *my brother and his friends* where *my brother* and *his friends* are linked with the aid of the coordinating conjunction *and* (BIBER *et al.*, 1999, p. 113). It is also possible to use coordinating conjunctions to coordinate NP heads, as in *red dresses and skirts* where the head nouns *dresses* and *skirts* are both modified by the adjective *red* (BIBER *et al.*, 1999, p. 113). In such cases, NPs could have the larger NP, the one which contains the two or more heads, analyzed as one NP that contains coordinated constituents but it is also possible to analyze the NPs inside the larger one, examining each individual head separately. Still, these NPs are not frequently analyzed in research. Biber *et al.* (1999), for instance, dedicate one small section to the analysis of coordinated binomial phrases, that is, phrases consisting of two words from the same word class, while Carter and McCarthy (2006) briefly consider these phrases as compounds. According to their analyses, phrases with two coordinated nouns

[5] The existential *there* is not usually described as an NP head, but the parser used in this study segment it as an NP.

are more common in academic prose (BIBER *et al.*, 1999, p. 1034). In this thesis, we decided to analyze these types of NPs and separate them into three different categories. We will distinguish simple coordinated NPs, which have no modifier associated to any of the heads, from complex coordinated NPs, which have modifiers associated to all the heads in the NP, from NPs that have simple and complex heads being coordinated into a single NP.

Having discussed the NP head, let us present next the determiners that may precede head nouns.

2.2.1.2 Determiners

Determiners[6] are NP constituents which always precede head nouns and premodifiers. They are a closed class of function words that occur in NPs so as to specify or narrow down in various ways the reference of the head noun (BIBER *et al.*, 1999). Usually, the reference indicated by a determiner is of definiteness, indefiniteness, possession, number, or quantity (CARTER; McCARTHY, 2006). In that sense, determiners differ from modifiers, as the former 'determine' the reference of a headword instead of modifying its meaning. Therefore, a determiner might be one of the following listed categories (from (3) to (9), followed by examples where the determiners are marked in **bold**).

3) Indefinite article (*a*, *an*)
    a. *He is **a** director of the Eastern Ravens Trust, which helps disabled people in the area.* (ACAD) (BIBER *et al.*, 1999, p. 261, our highlight)

The indefinite article can be used to "express an indefinite meaning" (CARTER; McCARTHY, 2006, p. 907) or indefinite reference not shared by writer and reader (QUIRK *et al.*, 1985), but it usually serves to introduce a new referent in discourse (BIBER *et al.*, 1999). It always occurs with head nouns in their singular form and its frequency is quite similar across the registers analyzed by Biber *et al.* (1999) because of its function to introduce new entities (p. 268).

[6] Determiners are also referred to as determinatives in Quirk *et al.* (1985).

4) Definite article (*the*)

   a. ***The** patterns of industrial development in the United States are too varied to be categorized easily.* (ACAD) (BIBER *et al.*, 1999, p. 264, our highlight)

The definite article is used to specify the entity expressed by the NP, assuming it is something known to the writer and reader (BIBER *et al.*, 1999; CARTER; McCARTHY, 2006; QUIRK *et al.*, 1985). Differently from the indefinite article, it can occur with singular and plural nouns. In academic prose, the definite article is 30-40% of the times used with a cataphoric reference, i.e. the definite NP refers to something that follows in the text (BIBER *et al.*, 1999, p. 264, 266). According to Biber *et al.* (1999), that reference pattern is probably recurrent due to the complexity of NPs in academic texts.

It should be remarked that articles in general are the most frequently used determiner in academic prose, but the definite article has a higher occurrence than the indefinite in written registers (BIBER *et al.*, 1999, p. 267). Moreover, in the Penn Treebank, there is not a specific POS tag for articles, so they are often tagged as determiner, i.e. DT.

5) Demonstrative pronoun (*this*, *that*, *these*, *those*)

   a. *The simplest form of chemical bond, in some ways, is the ionic bond. Bonds of **this** type are formed by electrostatic attractions between ions of opposite charge. **This** attraction is exactly of the same nature as the attraction that makes hair stand up when some synthetic fabrics are drawn over it.* (ACAD) (BIBER *et al.*, 1999, p. 273, their highlight)

Demonstrative pronouns might occur as determiners as a way to specify a known entity and, particularly in written texts, to refer back to the immediate preceding text in an anaphoric reference (BIBER *et al.*, 1999). Depending on the number of the referent, a different form will be used, that is, *this* and *that* are used with singular nouns and *these* and *those* are used with plural nouns (CARTER; McCARTHY, 2006). The pair *this* and *these* is considered proximate forms while *that* and *those* are distant forms. The former pair of demonstratives are more frequent than the latter in academic prose due to its anaphoric reference (BIBER *et al.*, 1999, p. 274). In general, demonstratives are less frequent than definite articles in the English language (BIBER *et al.*, 1999, p. 270) although both have a definite meaning (QUIRK *et al.*, 1985). Demonstrative pronouns are not categorized differently in the Penn Treebank and do not have its own distinctive POS tag, receiving then the determiner tag, i.e. DT.

6) Possessive pronoun (*my*, *our*, *your*, *her*, *his*, *its*, *their*)
   a. *We want industry to cut down on* ***its*** *own waste, and make better use of other people's.* (NEWS) (BIBER *et al.*, 1999, p. 271, our highlight)

Certain possessive pronouns are used as determiners so as to identify the reference of an NP in relation to the writer (*my*, *our*), the reader (*your*), or other referents presented in the text (*his*, *her*, *its*, *their*) (BIBER *et al.*, 1999).[7] Their frequency is low in academic prose, but *its* and *their*, which can be used for non-human reference, and *our*, which could refer to the author conjoined or not with the reader, is frequent in academic texts (BIBER *et al.*, 1999, p. 272). Nevertheless, possessive pronouns in general are not as frequently used as definite articles are in English (BIBER *et al.*, 1999, p. 270). Possessives have their own tag in the Penn Treebank, i.e. PRP$.

7) Quantifier
   a. ***Every*** *minute of* ***every*** *day, hundreds of millions of tonnes of coal are burned.* (ACAD) (BIBER *et al.*, 1999, p. 275, their highlight)

Quantifiers[8] are used to specify the general quantity of a head noun (BIBER *et al.*, 1999). There are quite a few quantifiers in English and they represent different quantity references, which could be said to be positive or negative (CARTER; McCARTHY, 2006). The quantifiers *all*, *both*, *each*, and *every* have an inclusive reference. *Any* and *either* specify arbitrary amounts while *no* and *neither* have negative references. *Many*, *much*, *more*, and *most* express large quantities. *Some*, *few*, *several*, e*nough*, *little*, and *less* refer to moderate or smaller quantities. Quantifiers that express generalization, such as *many* and *some*, or precision, such as *each* and *both*, are more frequently found in the academic register (BIBER *et al.*, 1999, p. 277). In the Penn Treebank, this type of determiners do not have a unique tag,

[7] Carter and McCarthy (2006) mention the reference done by other grammar books to possessive determiners as possessive adjectives. Quirk *et al.* (1985) also distinguish weak possessive pronouns, referring to their determiner forms such as *my*, from strong ones, referring to their nominal head form such as *mine*.

[8] Silero (2014) presents a study of the quantifiers *few* and *a few* in Brazilian learner corpora.

but they are most often tagged as determiners, i.e. DT, or predeterminer when occurring before an article, i.e. PDT.

8) Numeral
    a. ***Four*** *people were arrested.* (NEWS) (BIBER *et al.*, 1999, p. 89, their highlight)

Numerals are "a class of infinite membership" (QUIRK *et al.*, 1985) categorized as cardinal or ordinal. Cardinal numerals are the words *one*, *fourteen*, etc., and they serve to specify entities numerically (BIBER *et al.*, 1999). Ordinal numerals, on the other hand, are the words such as *first* and *fourteenth*, which are used when referring to entities in a sequential order (BIBER *et al.*, 1999; CARTER; McCARTHY, 2006). Biber *et al.* (1999) counted numerals as they were used in both determiner and nominal head positions together, which does not allow for a general comment about the use of numerals as determiners only. However, they found that cardinals are frequent in academic prose, a result that could be explained by this register's informational purposes (p. 279).

9) Semi-determiner
    a. ***Such*** *functions are not symmetrical.* (ACAD) (BIBER *et al.*, 1999, p. 281, their highlight)

Semi-determiners are often classified as adjectives or pronouns but they actually do not have a descriptive meaning. They mainly specify the reference of a head noun (BIBER *et al.*, 1999), often adding an indefinite meaning to it (QUIRK *et al.*, 1985). Examples of semi-determiners are the words *another*, *other*, *certain*, *former*, *latter*, *next*, *last*, *same*, *such*, that in written registers regularly refer to something mentioned previously in the text. Once more, Biber *et al.* (1999) analyzed semi-determiners being used as determiners and as pronouns altogether but they report that *same*, *other*, and *such* are more commonly used in academic texts, which could indicate the high precision required in those texts (p. 282).

In the NP structure, these determiners will occur in first position, before other constituents such as premodifiers. Interestingly, more than one determiner can occur in an NP and linguists have already identified possible patterns or fixed orders. More specifically, NPs could have a predeterminer (i.e. quantifier), a central determiner (i.e. an article, a possessive, or a demonstrative pronoun), and two postdeterminers (i.e. a ordinal numeral followed by a semi-determiner or a cardinal numeral followed by a quantifier) (BIBER *et al.*, 1999; QUIRK *et al.*, 1985). In example (10), the NP illustrates the occurrence of three determiners, i.e. a quantifier, a definite article, and a semi-determiner, in this order, before a head noun.

10) *all the other* ***books***

As it will be seen in detail in section 2.2.2, an NP head alone or accompanied by determiners can be classified as a simple NP. Before that, let us see the modifiers that may precede or follow an NP head.

2.2.1.3 Modifiers

Modifiers are another type of NP constituent used to "describe or classify the entity referred by the head" (BIBER *et al.*, 1999, p. 97). That means that the co-occurrence of modifier and head noun in an NP often restricts the reference of the entity and adds descriptive information, e.g. subjective qualities or physical attributes, to it (QUIRK *et al.*, 1985; CARTER; McCARTHY, 2006). Interestingly, "in academic prose, almost 60% of all noun phrases have some modifier" (BIBER *et al.*, 1999, p. 578), adding up to the importance of modifiers to the description of NPs. Some of these modifiers occur before the NP head while others occur after it. These are called premodifiers and postmodifiers, respectively.

2.2.1.3.1 Premodifiers

NP premodifiers are most common in written registers (BIBER *et al.*, 1999, p. 589) and, in academic prose, c. 25% of NPs have premodifiers (BIBER *et al.*, 1999, p. 578). Premodification usually conveys information in fewer words than postmodification and, in that sense, premodifiers are considered condensed or compressed forms of meaning (BIBER *et al.*, 1999). As a consequence, the meaning relationship established between premodifier and head noun is less explicit (BIBER *et al.*, 1999), depending on the shared knowledge between writer and reader. This type of modifier includes the categories (11), (12), and (13) listed, which are followed by examples where the premodifiers are marked in **bold**.

11) Adjectives
    a. ***preparative*** *treatment* (BIBER; GRAY, 2016, p. 246, our highlight)

Adjectives in the position of NP premodifier[9] are the most used type of premodifier in the English language given the wide range of meanings they might add, e.g. size, color, nationality, etc., to head nouns (BIBER *et al.*, 1999). They tend to modify most frequently common nouns, as in example (11a), but they can also modify proper nouns in certain contexts. These adjectives occur with high frequency in written registers, demonstrating "the heavy reliance on NPs to present information" (BIBER *et al.*, 1999, p. 506), and the ones categorized as classifiers, i.e. adjectives which are used to place the referent in a category, are remarkably frequent in academic prose (BIBER *et al.*, 1999, p. 511). In the Penn Treebank, adjectives have three specific tags that can be attached to them: JJ, for the base form of adjectives, JJR for comparative forms, and JJS for superlative forms.

12) Nouns
    a. ***peace treaties enforcement*** *action* (BIBER; GRAY, 2016, p. 180, our highlight)

Nouns used as NP premodifiers are not as common in academic prose as adjectives, but they represent approximately 30% of all premodifiers in this register (BIBER *et al.*, 1999, p. 589). Sequences of nouns in an NP not only of one but two, three, and four premodifying nouns as in example (12a) are extremely compressed in terms of packaging of information. This is perhaps due to limited space (CARTER; McCARTHY, 2006) or technical underlying meanings (BIBER; GRAY, 2016), and a multitude of semantic relations can be expressed with nouns in premodification, e.g. composition, purpose, identity, among others (cf. BIBER *et al.*, 1999, p. 590). Premodification by nouns in which the meaning relation is unpredictable might be unacceptable (QUIRK *et al.*, 1985) or cause ambiguities (CARTER; McCARTHY, 2006). Some NPs with premodifying nouns found in academic texts can be quite specialized, thus

[9] Adjectives when used as NP premodifiers are considered to have an attributive function, as opposed to the predicative function, in which the adjective follows a copular verb (QUIRK *et al.*, 1985).

unpredictable to a lay reader but still acceptable to the expert reader (BIBER; GRAY, 2016). Generally, a small number of nouns, particularly singular forms (CARTER; McCARTHY, 2006), combine with head nouns and produce different referents (BIBER *et al.*, 1999).

13) participial modifiers
    a. ***growing*** *problems* (BIBER *et al.*, 1999, p. 588, their highlight)

Participle forms used as NP premodifiers are those words, mostly derived from verbs, ending with the suffixes *-ed* or *-ing*, which "indicate a permanent or characteristic feature" of the referent (QUIRK *et al.*, 1985, p. 1325). In comparison to the other premodifiers presented, participial modifiers are relatively uncommon in English (BIBER *et al.*, 1999, p. 589). Despite being a verb form, participle forms premodifying head nouns can be grammatically analyzed as adjectives and be referred to as participial adjectives (BIBER *et al.*, 1999; QUIRK *et al.*, 1985). For this reason, during automatic parsing, participial modifiers might be tagged as adjectives through the tag JJ or as participle receiving the tags VBG for gerund forms or VBN for past participle forms. However, in this study, even if annotated as adjectives, participial forms will be manually categorized as participial modifiers so as to be consistent with the formal definition of participles.

14) *'s* genitive
    a. *To set the tone for our discussion and to put planning and evaluation into proper perspective, we present* ***Berg and Muscat's*** *definition of planning.* (ACAD) (BIBER *et al.*, 1999, p. 298, their highlight)

Another category of premodifiers that is not usually classified as one is that of the genitive forms of nouns.[10] In those cases, common and proper nouns are combined with the suffix -*'s* in order to specify the reference of the head noun, usually in terms of possession, or to classify it under a certain group or type (BIBER *et al.*, 1999). When the genitive specify another noun, it has a similar function to that of a determiner and, when classifying the noun, it functions as a premodifying adjective or noun (BIBER *et al.*, 1999). In academic prose, the

[10] Carter and McCarthy (2006) refer to the *'s* genitive as the possessive *'s* and treat it as a determiner, considering its similarity with possessive pronouns (p. 361). Quirk *et al.* (1985) opt for dealing with genitive as determiner or modifier depending on its functions of possession or attribution of particular characteristics (p. 326-327).

frequency of genitive forms is low (BIBER *et al.*, 1999, p. 300). In this study, all occurrences of genitive were recognized by its Penn Treebank tag, i.e. POS, and categorized as NP premodifier.

According to Biber *et al.* (1999), proportionately in all registers, the great majority of NPs with premodifiers, 70-80%, have only one premodifier whereas 20% have two premodifiers and 2% have three or four premodifying elements (p. 597). In these cases, we see a defining characteristic of NPs, that of allowing the compression of information and meaning relations through multiple premodifiers. There is no rule to the order of multiple premodifiers, but there is a tendency for language users to use premodifying nouns closest to the head noun (BIBER *et al.*, 1999, p. 599). An issue regarding the use of two or more premodifiers in an NP is that it is not always easy to identify if all the premodifiers are modifying the head noun as some words might be modifying premodifiers (BIBER *et al.*, 1999). Nonetheless, this ambiguity does not happen when multiple premodifiers are coordinated. In fact, coordination makes the meaning relations among premodifiers and the head noun direct and explicit (BIBER *et al.*, 1999) as it also avoids the repetition of the head (CARTER; McCARTHY, 2006). The most common structure of coordinated premodifiers found in academic prose is that of adjectives coordinated with the conjunction *and* (BIBER *et al.*, 1999, p. 601).

2.2.1.3.2 Postmodifiers

NP postmodifiers are also quite common in academic prose. More specifically, c. 20% of all NPs have postmodifiers (BIBER *et al.*, 1999, p. 578) and there are many types of postmodification available to writers, including the ones from (15) to (19), which are followed by examples where the postmodifiers are marked in **bold**. Some postmodified NPs, such as the ones with PPs and appositive NPs, might be considered more condensed or compressed than others, considering these NPs usually use fewer words than postmodified NPs with relative clauses.

15) Prepositional phrases
    a. *the search* ***for new solutions*** (ACAD) (BIBER *et al.*, 1999, p. 636, their highlight)

PPs used as NP postmodifiers serve to express several meanings established between the head noun and the NP that follows the head preposition of the PP (cf. BIBER *et al.*, 1999, p. 636). In example (15a), the preposition *for* expresses the meaning of purpose between the NPs *the search* and *new solutions*. As a consequence of the wide range of meanings made possible by prepositions, which may be more or less explicit in meaning (QUIRK *et al.*, 1985), PPs are "by far the most common type of postmodification in all registers" and "extremely common in academic prose" (BIBER *et al.*, 1999, p. 606). Interestingly, though, c. 90% of postmodifying PPs are headed by only six prepositions, i.e. *of*, *in*, *for*, *on*, *to*, and *with*, being *of*-phrases the most frequent ones across registers (BIBER *et al.*, 1999, p. 635). In the Penn Treebank tagset, PPs receive the phrasal-level tag PP and the head prepositions receive the word-level tag IN.

Postmodifying PPs usually represent a more dense packaging of information, especially because their prepositions are complemented by NPs and there is the possibility of PPs occurring in sequences, one embedded into the other, adding up layers of meaning (BIBER *et al.*, 1999). As a consequence, PPs used as postmodifiers are categorized as compressed features. However, it is possible to have PPs complemented by clauses, e.g. *wh*-clauses, and in such cases postmodifying PPs are considered less compressed.

16) Finite clauses
   a. *The lowest pressure ratio* ***which will give an acceptable performance*** *is always chosen.* (ACAD) (BIBER *et al.*, 1999, p. 611, their highlight)
   b. *ways* ***that could be construed as aggressive*** (ACAD) (BIBER *et al.*, 1999, p. 622, their highlight)
   c. *the way* ***we acquire knowledge*** (ACAD) (BIBER *et al.*, 1999, p. 621, their highlight)

A clause might follow a head noun in an NP and be used as a postmodifier to add new information about the head in a most explicit manner (QUIRK *et al.*, 1999). The finite clauses found in English texts are also known as relative clauses[11] and there are three types that could be used. First, there are the *wh-* clauses, illustrated in (16a), which make use of a relative pronoun (*which*, *who*, *whom*, or *whose*) or a relative adverb (*when*, *where*, or *why*) as a means

[11] The distinction between defining/restrictive and non-defining/non-restrictive proposed for relative clauses in grammar books will not be taken into consideration in this study.

of anaphorically referring to the head noun (BIBER *et al.*, 1999). Second, there are the *that* clauses, as in (16b), which also connect the head noun and a finite clause with the aid of the relative pronoun *that*. Third, there are the zero relativizer clauses, exemplified in (16c), which do not have a relative pronoun or adverb to make the relationship between head noun and clause explicitly but the postmodification exists and is apparent. In academic prose, the most common finite clauses used by writers are the *wh-* clause with the relative pronoun *which*, the *that* clause, the *wh-* clause with *who*, and the zero relativizer clause, in this order (BIBER *et al.*, 1999, p. 611).

The choice of finite clause varies according to certain conditions. One of these conditions refers to the head noun being human or non-animate. In the former case, *who* would be selected and, in the latter, *which* could be chosen. As it was mentioned in the last paragraph, in academic prose, *which* is more used than *that* and the zero relativizer and that happens because of stylistic reasons, in which the relative pronoun *which* is associated with a more literate and appropriate discourse while *that* and the zero relativizer are considered to be more colloquial (BIBER *et al.*, 1999). The Penn Treebank has a specific way to parse and tag finite clauses, i.e. in the clause level these clauses are parsed with the tag SBAR, which refers to a clause introduced by a subordinating conjunction, in the phrase level they receive the tags WHADJP, WHADVP, WHNP, which would represent the different classification of relative pronouns or adverbs, or no tag at all, and in the word level the tags WDT, WP, WP$, or WRB would be used, depending on each type of relative pronoun or adverb produced.

17) Non-finite clauses
    a. *products* ***required to support a huge and growing population*** (ACAD) (BIBER *et al.*, 1999, p. 604, their highlight)
    b. *a structure* ***consisting of independent tetrahedra*** (ACAD) (BIBER *et al.*, 1999, p. 604, their highlight)
    c. *Feynman offers us a simple way* ***to see that this happens****.* (ACAD) (BIBER *et al.*, 1999, p. 634, their highlight)

Non-finite clauses also add new information to the head noun when in a postmodifying position, but differently from the finite clauses, the verbs in these clauses are not inflected for tense. There are two types of them: the participle clauses, i.e. the *-ed* and the *-ing* clauses, and the infinitive clauses, i.e. the *to* clauses (BIBER *et al.*, 1999). In the Penn Treebank, these clauses can be parsed in the phrase level as a verb phrase, corresponding to the tag VP, and tagged in the word level with the tags VBN, VBG, and TO, which represent the two participle clauses and the infinitive clauses, respectively. Participle clauses are commonly used in academic texts, considering that they can be considered more economical than full relative clauses (BIBER *et al.*, 1999, p. 632), being *-ed* clauses more frequent than *-ing* clauses, while *to* clauses are rare in English (BIBER *et al.*, 1999, p. 606). Being more compressed, these structures are also considered less explicit in meaning than finite clauses (QUIRK *et al.*, 1985).

It should be remarked that there are noun complement clauses quite similar to *that* and *to* clauses; however, they present structural and semantic differences (BIBER *et al.*, 1999). Complement clauses are controlled by a set number of head nouns, such as *expectation*, *fact*, and *attempt*, which express the "stance towards the proposition in the complement clause" (BIBER *et al.*, 1999, p. 647). Due to time constraints, it was not possible to analyze noun complement clauses separately from the other types of finite or non-finite clauses and they were classified as either one or the other according to the verb tense used in the postmodifying clause.

18) NPs in apposition
    a. *Comparison of these scores to the studies in our meta-analysis reveals that they are all of moderate quality* ***(scores of 2 to 4 on a scale of 0 to 5)*** (BIBER; GRAY, 2016, p. 205, our highlight)

NPs can be used as NPs postmodifiers with the objective of providing extra descriptive information about the head noun (BIBER *et al.*, 1999). Appositive NPs refer to the same entity referred to in the NP modified and, for that reason, they could have their positions reversed without alterations in meaning (CARTER; McCARTHY, 2006). These appositive NPs are considered a maximally compressed form of postmodifier (BIBER *et al.*, 1999). Commonly used in academic prose, representing 15% of all NP postmodifiers in the register, they might follow proper nouns, technical terms, or introduce acronyms, labels for variables, formulas, and list of items that are part of a class (BIBER *et al.*, 1999, p. 639-640).

19) Adjective phrases
    a. *The extremely short duration varieties* ***common in India*** *were not used in West Africa.* (ACAD) (BIBER *et al.*, 1999, p. 605, their highlight)

Other phrases can also be used as NP postmodifiers such as adverb phrases (AdvP) and adjective phrases (AdjP). Nevertheless, they are less common in academic writing (BIBER *et al.*, 1999, p. 604). Still, we would like to analyze the use of AdjPs, which could be used properly by Brazilian learners or not, as in Brazilian Portuguese the use of adjectives is most common in noun postposition. In English, though, postmodification is restricted to some adjectives, such as *available*, and to AdjPs that have an adjectival complement (BIBER *et al.*, 1999), and to a few noun-adjective combinations, such as *president elect* (QUIRK *et al.*, 1985).

According to Biber *et al.* (1999), in academic prose, it is quite frequent to find multiple postmodifiers following a head noun (p. 642). In those cases, several postmodifiers might be adding new meaning to one single head noun or each postmodifier will modify the immediately preceding head noun, creating layers of embedding (BIBER *et al.*, 1999; QUIRK *et al.*, 1985). In NPs with multiple postmodifiers, PPs are more commonly used as the first postmodifier (BIBER *et al.*, 1999, p. 643).

As it will be further discussed in section 2.2.2, an NP accompanied by pre- and/or postmodifiers can be classified as a complex NP.

2.2.2 Simple and complex noun phrases

Having presented all the NP constituents, it is easier now to address a classification that distinguishes simple and complex NPs. Based on the formal description given in section 2.2.1, our classification should also use only formal and structural elements to define NPs as simple and complex. In other words, the different combinations of NP constituents will determine if an NP is simple or complex. It should be seen later that the structural combination can determine the structural complexity (HILLIER, 2004 *apud* AKINLOTAN; HOUSEN, 2017).

Knowing the NP constituents are the head noun, the determiner, and the pre- and/or postmodifier, the possible combinations of head noun and other constituents will define our classification, which is also given in Biber *et al.* (1999). On the one hand, simple NPs are the phrases that have a head noun by itself and phrases that have a head noun and a determiner(s) preceding it. In that configuration, the headword does not have its meaning modified; it simply has its reference specified, as explained previously. On the other hand, complex NPs are the phrases that have a head noun accompanied by at least one modifier. These NPs could have determiner(s) as their constituents, but it is the modifier(s) which distinguishes it from simple NPs as modifiers add new meaning to the head noun, as stated earlier.

More specifically, possible configurations of simple NPs involve the use of:

- Head noun alone;
- Determiner(s) + head noun.

Possible arrangements of complex NPs include:

- Premodifier(s) + head noun;
- Head noun + postmodifier(s);
- Premodifier(s) + head noun + postmodifier(s).

Complex NPs are very frequent in academic English and are strictly associated with the grammatical complexity in academic written texts. As it will be seen in section 2.3, grammatical complexity can be indicated by both clausal and phrasal features, but in academic texts, complex NPs serve the main purpose of packaging a good amount of information in fewer words, which represent an economy and efficiency of expression (BIBER *et al.*, 2009; BIBER; GRAY, 2016; CARTER; McCARTHY, 2006). An NP can have to a certain extent defined implicitly by the writers in an academic field several structures embedded in it and this can be useful to expert writers who often need to add new meanings to the same referent without creating new clauses or sentences. Such a strategy is made easier since authors can revise and edit their texts and the reader can (re)read them as many times as necessary (BIBER *et al.*, 2009; BIBER; GRAY, 2016).

Now that we understand what simple and complex NPs are, let us understand how researchers (are) used to understand grammatical complexity in academic English and how they started focusing on the study of the complexity of the NP in academic written language.

2.3 Overview of research on grammatical complexity in academic and learner writing

The idea of grammatical complexity has at large been part of the concern with human languages of modern linguistic studies in the late 20th century. The main concern was actually with the relative complexity of human languages, i.e. whether languages were equally complex or some were more complex than others (NEWMEYER; PRESTON, 2014). Different approaches of linguistics would try to understand and explain their viewpoints of linguistic complexity from varying foci, such as generative grammarians who would argue that languages had to be equally complex because of the demands of universal grammar (NEWMEYER; PRESTON, 2014). However, independent of the perspective adopted by linguists, all those interested have tried to propose measures, either grammar-based or user-based, that could indicate degrees of linguistic complexity (NEWMEYER; PRESTON, 2014).

Corpus-based research in the late 20th century would also concern itself with linguistic complexity but the language under study would not be taken as the same language in its spoken and written modes. That means that studies such as Biber (1988) would find that the English language has different grammars depending on its mode and, consequently, there would be different levels of complexity in each mode.

The *Longman grammar of spoken and written English* (BIBER *et al.*, 1999) would follow as one of the first studies to carry out research on large datasets of specific registers of English, i.e. conversation, fiction, newspapers, and academic prose, and demonstrate that the grammatical features considered complex were used differently in particular registers. As for the findings in academic prose, NPs were found to be a pervasive element even though not necessarily salient in the register as well as more syntactically complex than in other registers (BIBER; GRAY 2016).

Those findings confirmed the nominal style of academic writing that had already been noticed by Wells (1960 *apud* BIBER; GRAY 2010) and Halliday and Martin (1993). The latter, particularly, discuss the nominalization present in the history of scientific languages, in which experiences and processes become objects through the rewording of verbs and clauses into nouns, which could be extended through modification (HALLIDAY; MARTIN, 1993). Those shifts in the lexicogrammatical level of scientific language would give it a distinctive quality and make the scientific discourse possible (HALLIDAY; MARTIN, 1993).

This distinctive quality of academic language has been continually investigated during the earliest 21st century (e.g. BIBER, 2006; BIBER; GRAY, 2010; BIBER *et al.*, 2011) and culminated with the work by Biber and Gray (2016), *Grammatical complexity in academic English*. The historical-oriented investigation therein conducted advances the research on complex grammatical features, providing an overview of their use and importance to the singularity of academic written grammar. Nominal grammatical structures are not often found in earlier historical periods, for back in the 18th and 19th centuries the academic discourse was organized most frequently around clausal features and, as a result, was more structurally elaborated. However, in need of efficient and concise expressions of information, the use of a more compressed style resulted in the use of phrasal features in the 20th and 21st centuries. The structures found to be more frequently used and considered representative of the grammatical complexity in present day academic writing in general are (BIBER; GRAY, 2016, p. 167-217):

a) Noun(s) + head noun;
b) Attributive adjective + head noun;
c) Noun-participle + head noun;
d) Head noun + PP;
e) Head-noun + appositive NP.

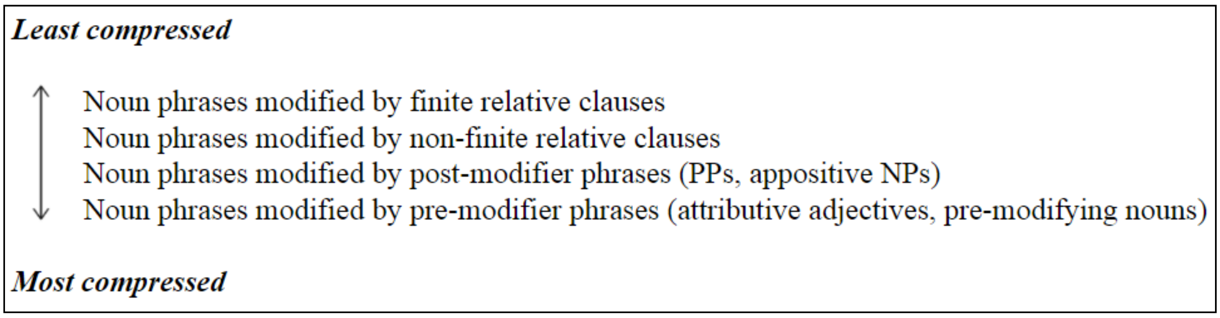
These grammatical complexity devices, along with others described in section 2.2 and clausal features not analyzed in this thesis, are generally seen as structural variants that could express the same information in phrasal or clausal packagings. For example, the complex NP *the Communist Party chief* is a sequence of nouns that could be used as an alternative to the *'s* genitive, *the Communist Party's chief*, or the *of*-phrase, *the chief of the Communist Party*

(BIBER; GRAY, 2016, p. 171). Nonetheless, the preference for one variant rather than another depends on contextual and discourse factors that change over time and can be documented through corpus analysis (BIBER; GRAY, 2016).

These preferences in the use of complex features also vary across the different academic sub-registers. The corpus of contemporary academic English used in Biber and Gray (2016) was composed of texts from the humanities, popular science, social science, and specialist science, and their analysis showed that these disciplines tend to prefer certain features over the others. More specifically, humanities writing relies more heavily on clausal features while specialist science and social science most frequently use phrasal structures modifying nouns. Specialist science research writing is actually considered the most representative academic sub-register in the use of grammatical complexity, as it uses more compressed structures.

Taking into account these different structural realizations of complex NPs, Biber and Gray (2016) offer the possibility of organizing them in a cline of structural compression (see FIGURE 2.4), which could be also interpreted as half of the cline of grammatical complexity (see FIGURE 2.5). The cline of compression is a representation of the possibilities of compression of information at a phrasal level, as well as in fewer words, in opposition to its elaboration at a clausal level. In this cline, NPs postmodified by clauses are considered the least compressed structures while NPs premodified by phrases are the most compressed, concise, and complex in academic written texts. This same cline also demonstrates the systematic changes over time in the use of grammatical complex features that shows the system has undergone “a ‘drift’ towards greater structural compression” (BIBER; GRAY, 2016, p. 208).

Figure 2.4 – Cline of structural compression



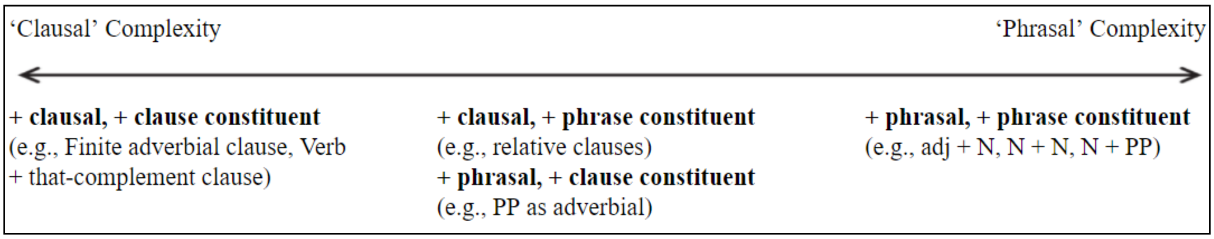
Least compressed

↕ Noun phrases modified by finite relative clauses
Noun phrases modified by non-finite relative clauses
Noun phrases modified by post-modifier phrases (PPs, appositive NPs)
Noun phrases modified by pre-modifier phrases (attributive adjectives, pre-modifying nouns)

Most compressed

Source: BIBER; GRAY, 2016, p. 207.

Figure 2.5 – Cline of grammatical complexity



'Clausal' Complexity

'Phrasal' Complexity

+ clausal, + clause constituent
(e.g., Finite adverbial clause, Verb + that-complement clause)

+ clausal, + phrase constituent
(e.g., relative clauses)
+ phrasal, + clause constituent
(e.g., PP as adverbial)

+ phrasal, + phrase constituent
(e.g., adj + N, N + N, N + PP)

Source: BIBER; GRAY, 2016, p. 62.

Taking Biber and Gray (2016) as a reference, some recent studies have analyzed the use of grammatical complexity features in academic texts in order to verify whether there are differences across academic level, disciplines, genres, and learners. Staples *et al.* (2016), for instance, bring forth an investigation of the development of grammatical complexity in the texts written by L1 English university students, from first-year undergraduate to graduate level, from different disciplines, and in varying genres. The texts analyzed are part of the *British Academic Written English* (BAWE) corpus, and an automatic tagger (not specified) is used to identify both clausal and phrasal features in them. After statistical analysis, the findings confirmed the authors' hypothesis that L1 writers develop the academic and disciplinary style later in their university lives than previously assumed, "actually us[ing] more compressed phrasal structures and more simple clausal structures" (STAPLES *et al.*, 2016, p. 179). For example, there was a robust increase of premodifying nouns and a decrease of postmodifying finite clauses in the texts across levels of study (STAPLES *et al.*, 2016, p. 163-164). It was also observed a variation in the use of clausal and phrasal features depending on the disciplines, which have certain preferences, and genres, which have particular functions (STAPLES *et al.*, 2016).

The research just presented examined grammatical complexity in L1 English texts, but other studies investigated it in L2 English writing. Parkinson and Musgrave (2014), in a research article, analyzed the use of complex NPs after manually coding essays written by two groups of L2 writers who are at graduate level: one of EAP learners and another of MA applied linguistics students. The corpus compiled had more than 26000 words and more than 7000 NPs were analyzed. Despite the corpus being small and composed of essays that were untimed and the participants being from various Asian countries, the findings showed expected aspects of the development of complexity in students' writing. In comparison to MA students, who used to a greater extent more nouns as premodifiers and more PPs as postmodifiers, EAP learners used significantly more adjectives as premodifiers and less PPs as postmodifiers. Those results lead the authors to suggest that EAP classes with less proficient L2 writers should have "a focus on nouns as premodifiers and prepositional phrases as postmodifiers" (PARKINSON; MUSGRAVE, 2014, p. 58).

Ansarifar *et al.* (2018) also report in a research article their analysis of grammatical complexity in graduate level abstracts written by L1 Persian writers in comparison to abstracts written by published writers in journals of applied linguistics. The corpus collected was composed of applied linguistics abstracts and contained more than 75000 words. The abstracts written by L2 MA and PhD writers were compared to published abstracts in terms of 16 NP features of grammatical complexity, which were coded manually by two linguists. Results showed that L1 Persian writers only differed from expert writers in producing four features, i.e. nouns and adjective/noun sequences as premodifiers as well as *-ed* clauses and multiple PPs as postmodifiers (ANSARIFAR *et al.*, 2018, p. 67). Still, the most common premodifiers used in the whole corpus were adjectives and nouns while the most common postmodifiers were PPs. Most interestingly, PhD writers used more complex features, with exception to multiple PPs, making their abstracts grammatically similar to expert writers' abstracts (ANSARIFAR *et al.*, 2018, p. 68).

Most relevant for us perhaps is Nitsch's (2017) PhD thesis, which analyzes the complexity of NPs produced by Brazilian learners. In that research, three corpora are used: the *Brazilian Portuguese subcorpus of the International Corpus of Learner English* (Br-ICLE), a learner corpus of general topic essays written by Brazilian L2 English writers (c. 160 thousand words), the *Louvain Corpus of Native English Essays* (LOCNESS), a corpus of essays written by L1 English writers, therein used as a comparative corpus, and BAWE, used as a reference corpus to create a keyword list. Based on the keyword list created with *WordSmith Tools* and taking only the words used in both Br-ICLE and LOCNESS, Nitsch (2017) analyzed a sample of ten head nouns (i.e. *people*, *money*, *problems*, *things*, *students*, *life*, *job*, *person*, *children*, *television*), several participle *-ing* forms used as NP heads in NPs with determiners (e.g. *beginning* and *feeling*), and personal pronouns. In total, more than 8000 NPs were manually checked and categorized according to four levels of complexity (see TABLE 2.1). From those NPs, more than 2000 NPs produced by learners were identified as complex, which corresponded to 25.8% of all NPs analyzed, while the other 74.2% of NPs analyzed were found to be simple (NITSCH, 2017, p. 98). In contrast to the results found in the native corpus, the learner corpus sample presented the use of more simple NPs and less complex NPs by learners. Concerning complex NPs in particular, Brazilian students were found to use more premodifiers than postmodifiers, in particular adjectives (80%) and participle forms (12%) than nouns (8%) as premodifiers (NITSCH, 2017, p. 106-107). As for postmodifiers, finite clauses (52%) and PPs (40%) were more commonly used by Brazilian writers than non-finite clauses (8%). Despite the careful methodology applied, the higher use of simple NPs in the corpora analyzed could be a result of the generic list of head nouns and/or of the texts analyzed being general in topic and written by students as opposed to specialized texts produced by experts.

Table 2.1 – Levels of NP complexity

NP type	**Complexity**
(det) **head**	0
(det) pre-mod **head**	1
(det) **head** post-mod	2
(det) pre-mod **head** post-mod	3

Source: NITSCH, 2017, p. 89.

Note: Table was translated from Portuguese to English by the author.

Bearing all those studies, findings, and considerations in mind, it is necessary to define grammatical complexity. Biber and Gray (2016) do not explicitly define this construct; they essentially describe the forms, functions, and meaning relations of the features responsible for the complexity in professional academic writing as well as report on the frequencies of complex structures. Nevertheless, it is possible to summarize their main ideas and define grammatical complexity as the normalized rates of occurrence of complex features, both phrasal and clausal, in the corpus. In other words, having the rates of occurrences, it is possible to compare the use of phrasal and clausal features in texts and determine whether their grammatical complexity is more connected to one or the other.

The grammatical complexity of NPs, in particular, can be defined in terms of the frequency of certain combinations of modifiers and the head noun as well as regarding the implicit meaning relation among the noun head and its constituents. The higher use of complex NPs demonstrate that the grammatical complexity in academic writing is economical in expression, in the sense that there is an attempt “to convey the maximum amount of information in the fewest words possible” (BIBER; GRAY, 2016, p. 207). This makes sentences and texts easier and faster to read, especially for experts, and demonstrates how “writing is shaped by individuals making language choices in social contexts” (HYLAND, 2014, p. 109).

In this thesis, our definition of grammatical complexity, then, is grammar-based and centered in the NP structure, meaning that the structural combination of NP constituents determines whether the NP is simple or complex (see section 2.2.2). Complex NPs, particularly, can be more or less complex depending on the use of multiple pre- and/or postmodifiers. Moreover, if more complex NPs are used in a text, the degree of grammatical complexity in that text is higher than if more simple NPs were used. It is still important to remark that grammatical complexity, as it appears in recent studies, involves both clausal and phrasal levels of complexity, but herein we are only analyzing the phrasal level through the description of NPs as used by Brazilian learners.

Thinking about the structural complexity of texts, Gray (2015) gives a clear notion of how a set of grammatical features can create a compressed or an elaborated style. In other words, the higher use of phrasal features means information is being compressed whereas the higher use of clausal features represents information is being elaborated. In Gray’s (2015) investigation of the structural complexity in journal registers, she proved that “all disciplines and registers maintain the nominal style of academic writing, relying on phrasal features of compression to much greater extents than clausal embedding” (p. 128). More specifically, adjectives are used more frequently as premodifiers (60-75 times per 1000 words) and PPs as postmodifiers (30-40 times per 1000 words) by expert writers in general, while quantitative research tends to use more premodifying nouns than qualitative research (GRAY, 2015, p. 123).

It is still indispensable to make reference to the interest SLA researchers have in grammatical complexity. Together with accuracy and fluency, complexity is usually approached holistically, a perspective that examines several grammatical features in order to identify if learners have control over the wide range of the L2 linguistic resources available (ORTEGA, 2015). In many SLA studies, though, complexity has been vaguely defined

(BULTÉ; HOUSEN, 2012) and taken to be represented by the use of varied and sophisticated syntactic forms (e.g. LU, 2011; CROSSLEY; McNAMARA, 2014; LU; AI, 2015), focusing on phrasal but mostly on clausal features such as subordinate clauses.

This leads to the use of several quantitative methods to identify and quantify complexity in L2 learner writing which are not as fine-grained as they should be (BULTÉ; HOUSEN, 2012). Bulté and Housen's (2012) analysis of 40 studies on grammatical and/or lexical complexity published between 1995 and 2008, for instance, showed that most studies used measures targeting complexity at the sentential level, e.g. mean length of clauses, clauses per AS-unit, c-unit, or T-unit, and only a few measures were calculated in each study despite the great number of measures at disposal. Seven of the latest corpus-based articles from L2 writing research (LU, 2011; CROSSLEY; McNAMARA, 2014; TAGUCHI *et al.*, 2014; BULTÉ; HOUSEN, 2015; LU; AI, 2015; MAZGUTOVA; KORMOS, 2015; STAPLES; REPPEN, 2016), all concerned to a certain extent with finding ways to automatically measure the syntactic complexity of learners' production and determine learners' progress, used a range of 11 to 15 measures of sentence variety, syntactic transformations and embeddings, phrase types and length. Most studies found that measures of NP complexity are more representative of advanced learners and their development in writing and suggested further research on those (LU, 2011; CROSSLEY; McNAMARA, 2014; TAGUCHI *et al.*, 2014; LU; AI, 2015; MAZGUTOVA; KORMOS, 2015).

In this thesis, we do not use any specific measure to represent the complexity of NPs, except for normalized rates of occurrence per 1000 words. Our main goal is to quantify the NPs as produced by learners and categorize as well as analyze them according to their constituency. However, if we were to suggest it, complexity should perhaps be defined and measured in terms of: 1) the number of NP constituents, i.e. words per phrase (linear parameter); and 2) the number of relationships between the constituents and other elements dependent on the same head noun (hierarchical parameter) (cf. BULTÉ; HOUSEN, 2012, p. 22, 40; PALLOTTI, 2015, p. 123).

In the next chapter, the methodology involved in extracting and analyzing the NPs produced by learners is discussed in detail.

CHAPTER 3

METHODOLOGY

This chapter focuses on describing the methodological decisions behind the design of the corpus used in this research and the choice of procedures for data retrieval and analysis. Section 3.1 details the theoretical framework adopted for our research methodology, which is circumscribed by L2 writing research. Section 3.2 describes the learner corpus selected for this study, giving a brief account of its creation, purpose, design criteria, and accessibility. It also provides essential information about the design of the subcorpus of CorIFA which represents our research corpus and gives details about its participants and type of text written by them. Section 3.3 deals with the annotation of the subcorpus, which refers to its automatic constituency parsing. Also within this section, the process of automated data extraction and manual data analysis is specified.

3.1 Theoretical framework

The research here proposed is a descriptive and corpus-based L2 writing study. Historically, L2 writing research is a branch of applied linguistics since several of its domains contributed to the development of frameworks and methods for the analysis of L2 written texts, with the primary goal of "creat[ing] pedagogical models for teaching L2 writing" (HINKEL, 2011, p. 523). However, herein, we do not concern ourselves with suggesting pedagogical applications based on our analysis but with describing the formal aspects of the English NP as produced in L2 written texts. Nevertheless, this could certainly lead to future studies on applied linguistics and the development of pedagogical materials specially created for Brazilian learners of English (cf. DUTRA; SILERO, 2010; DUTRA; BERBER SARDINHA, 2013; ALMEIDA, 2014; SILERO, 2014; DUTRA; GOMIDE, 2015; OLIVEIRA, 2015 for research on Brazilian students' writing).

In its own right, L2 writing research has been concerned over the last decades with various topics related to writing instruction and research and has taken pedagogical, linguistic, and cognitive approaches (RANSDELL; BARBIER, 2002). As for investigations of lexical, morphological, and syntactic phenomena, L2 writing research has relied on empirical quantitative methods, very often using corpus linguistics methodologies and tools, which made the analysis of a large quantity of data produced by learners feasible.

What should be kept in mind is that both L2 writing research and corpus linguistics can offer the necessary principles, tools, and methods for a research concerned with the grammatical aspects of learner writing such as ours. Considering that this thesis aims at investigating how a syntactic construct (NPs) is used in written texts (argumentative essays) produced by a particular population (Brazilian L2 writers), it seems most appropriate to propose a descriptive, empirical, and deductive approach to this research. In other words, this study describes and analyzes authentic language in use and, at the same time, it tests the following hypotheses (see Chapter 1 for details):

1) Brazilian upper intermediate learners use more simple NPs than complex NPs in their English written production;
2) Brazilian upper intermediate learners produce more NPs with postmodifier(s) than NPs with premodifier(s) or NPs with both pre- and postmodifiers;
3) Brazilian upper intermediate learners use more adjectives as NP premodifiers;
4) Brazilian upper intermediate learners use more PPs as NP postmodifiers.

It should be clear that the first hypothesis formulated comes from the results found in Nitsch (2017), which showed the use of more simple NPs by Brazilian learners in an intermediate or upper linguistic proficiency level (see Chapter 2). The second hypothesis was proposed based on our perception as English teachers, imagining learners are not as used as expert writers to use more complex NPs, especially the ones with multiple premodifying elements. The last two hypotheses are based on results from studies on EAP learners' writing such as Parkinson and Musgrave (2014) which found a higher use of premodifying adjectives and postmodifying PPs by learners (see Chapter 2).

According to Hyland (2016), "texts can be approached in different ways and for different purposes: looking at systems of choices, institutional ideologies, L1 and L2 practices, what they say about communities of users and how they link to other texts" (p. 120). Our choice of approach is supported by corpus linguistics and text analysis concerned with texts as “systems of forms” (HYLAND, 2016, p. 120). First, a corpus can provide a representative sample of the language used by the population analyzed. Second, corpus linguistics methods can be applied as a means of analyzing machine-readable data whose analysis is not feasible by hand and eye alone (McENERY; HARDIE, 2012). Third, written texts can be objects for the analysis of grammatical patterns and regularities particular to an academic genre (HYLAND, 2016; POLIO, 2012).

Regarding our belief about writing, it is possible to see it according to six paradigms: 1) as expressive activity; 2) as cognitive activity; 3) as completed activity; 4) as situated activity; 5) as social activity; and 6) as ideology (cf. HYLAND, 2016, p. 122-123). In this thesis, writing is viewed as a completed activity as well as a social activity. More specifically, the former concept refers to a preference for describing the language rather than the writers or the writing process (HYLAND, 2016). The latter sees the linguistic regularities of texts as a result of social constraints that can influence the choices made by the writers in a given context (HYLAND, 2016). By associating both paradigms, it is expected that this research will be better aligned to the analysis proposed.

As for the subscription to a model of language, we consider language as a cognitive and sociocultural phenomenon particular to humans (WIDDOWSON, 1996) which can be observed and described by using scientific methodologies. This perspective is usually the basis for functional, sociolinguistic, and cognitive models, but taking a less strict approach, language can be said to be systematically organized and variable due to functional reasons and communicative purposes, and that leads to the generation of grammatical patterns that can be observed and analyzed (BIBER, 2010). Consequently, instead of just analyzing grammatical structures per se, the communicative context and purposes related to these structures are investigated.

Likewise, the language used by learners can be seen as systematic and variable in use depending on their communicative purposes. This is the hypothesis developed by Selinker (1972), and used loosely in this research, when proposing the existence of an interlanguage in non-native speakers’ minds.[12] It seems quite reasonable to believe that learner language is

[12] This is a strong hypothesis fundamental to SLA research, whose “main goal (…) is to characterize learners’ underlying knowledge of the L2” (ELLIS, 1994, p. 13). The notion of interlanguage can be defined as a non-native language “created and spoken whenever there is language contact,” when learners attempt “to express meaning in a second language” (SELINKER, 2014, p. 223). In that sense, studies on interlanguage, particularly corpus-driven ones, would not be investigating the target language (L1 English) but a system which is developed in L2 learners’ mind, in the attempt to uncover some or any of its new features (GRANGER, 2012).

"highly structured, containing new/novel forms" (SELINKER, 2014, p. 223) and that it can be studied for its own sake.

Lastly, learning is seen in the perspective of an EAP context and this could be considered to have a dual role. EAP is centered in general language learning, especially in contexts of English as a foreign language, but also in helping students master academic genres they might need to write during their academic lives. As a consequence, L2 writing is seemingly more than a cognitive activity (POLIO, 2012). It is an important medium for teaching students linguistic patterns pervasive in English academic discourse, allowing them to gain fluency in the writing conventions of their disciplinary communities and become part of them (HYLAND, 2014).

As a final word, it is necessary to remind the reader that this research, as a descriptive work, is "a systematic presentation of language facts – not the elaboration or validation of some specific language theory"[13] (PERINI, 2008, p. 8), even though every description presupposes a theory. That is done in the belief that L2 writing research can benefit from more detailed and accurate descriptions of relevant linguistic features with the aid of large datasets and reliable methods of scientific research. Having those in hand, it is possible to have useful descriptions of learner writing in order to inform future research and evaluate the relevance of theoretically predefined linguistic patterns and the effectiveness of automatic and manual methods of research.

3.2 Research corpus

As a corpus-based study of L2 writing, this research should make use of a learner corpus. This type of corpus can be defined as "computerized databases of foreign or second learner language" (GRANGER, 2012, p. 7) which usually consist of "apprentice texts as unpublished pieces of writing that have been written in educational or training settings, (often) for purposes of assessment" (SCOTT; TRIBBLE, 2006, p. 133).[14] In that sense,

[13] Our translation of: "a apresentação sistemática dos fatos da língua – não a elaboração ou validação de alguma teoria específica da linguagem".

[14] Scott and Tribble (2006) are contrasting apprentice texts with expert texts that have been published and accepted by members of a discourse community.

academic learner texts are quite different from texts written and published by experts and professionals in an academic field but equally valuable for analysis.

The corpus selected, which will provide authentic data for the description of the use of NPs by Brazilian EAP students, is a subcorpus of a Brazilian learner corpus that could be representative of this population. This corpus is CorIFA,[15] compiled at UFMG (cf. GUEDES, 2017; DUTRA; QUEIROZ; ALVES, 2017; DUTRA; ORFANÓ; ALMEIDA, 2019 for research which used CorIFA). What follows is a brief description of CorIFA.

3.2.1 CorIFA

As mentioned above, CorIFA is a learner corpus of academic texts written in English by university students enrolled in EAP classes offered in a Brazilian university context, where English is not the medium of instruction. Created in 2013, CorIFA was compiled following a specific sampling frame, in the expectation to construct a corpus that would allow linguists to explore learner language phenomena available in written texts of six different genres, i.e. statement of purpose, abstract, argumentative essay, literature review, research article, and summary. These texts are part of the course assignments and each student is expected to write at least two texts per semester, a first non-edited version and a final edited version of their texts. The corpus is still being compiled and expanded every semester until a total of 200 thousand words is obtained for each proficiency level available in the corpus (B1 to C1).[16] At the moment, CorIFA as a whole has more than 530 thousand word tokens.

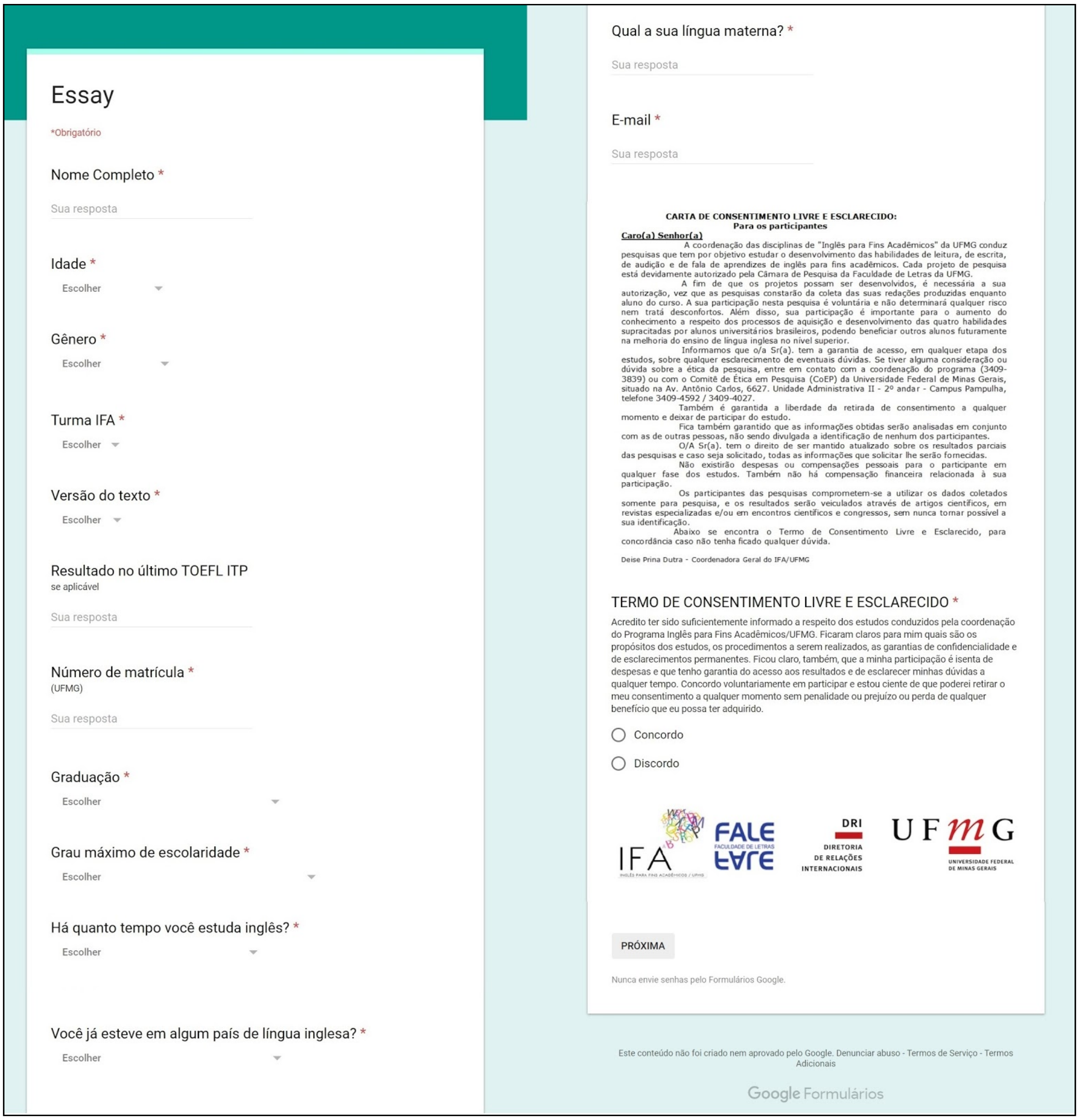
Each semester, at least one of the five EAP classes/levels is offered to both undergraduate and graduate students enrolled at UFMG. These students are of various ages, academic levels and disciplines, and have different first languages (L1s) (e.g. Spanish). Throughout the semester, these students are instructed about the one academic genre, learning its general structure and purposes. After instruction, students are supposed to write their own

[15] It should be remarked that CorIFA is a free learner corpus to be used for non-commercial purposes. Any researcher interested in using should contact Dr. Deise Prina Dutra, coordinator of the *Learner Corpus Research Group* at UFMG, who will authorize the use of the whole corpus or its subcorpora for research and share the CorIFA metadata and dataset requested. At the moment, there are research efforts to make the corpus available online. More information about CorIFA at https://sites.google.com/site/corpusifa/home .

[16] According to the *Common European Framework of Reference*, which is a guide to linguistic proficiency level of foreign language learners. The levels vary from A1 to C2, in which A1 is the most basic level and C2 is the most advanced.

texts in class or at home (depending on the instructors' choice), consulting dictionaries, grammar books, and other reference materials of their preference, without a time constraint, and send their first draft via Google Form (see FIGURE 3.1). On this online form, students should answer some questions that will serve as metadata for future research (see a list of these questions and some possible answers on APPENDIX B). Students are also given an informed consent term which they read and sign if they authorize or not the use of their metadata and text(s) in research (see APPENDIX C).

Figure 3.1 – CorIFA: Sample Google Form used



Essay

*Obrigatório

Nome Completo *

Sua resposta

Idade *

Escolher

Gênero *

Escolher

Turma IFA *

Escolher

Versão do texto *

Escolher

Resultado no último TOEFL ITP
se aplicável

Sua resposta

Número de matrícula *
(UFMG)

Sua resposta

Graduação *

Escolher

Grau máximo de escolaridade *

Escolher

Há quanto tempo você estuda inglês? *

Escolher

Você já esteve em algum país de língua inglesa? *

Escolher

Qual a sua língua materna? *

Sua resposta

E-mail *

Sua resposta

CARTA DE CONSENTIMENTO LIVRE E ESCLARECIDO:
Para os participantes

Caro(a) Senhor(a)

A coordenação das disciplinas de "Inglês para Fins Acadêmicos" da UFMG conduz pesquisas que tem por objetivo estudar o desenvolvimento das habilidades de leitura, de escrita, de audição e de fala de aprendizes de inglês 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 tratá 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) ou com 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 IFA/UFMG

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO *

Acredito ter sido suficientemente informado a respeito dos estudos conduzidos pela coordenação do Programa Inglês para Fins Acadêmicos/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

PRÓXIMA

Nunca envie senhas pelo Formulários Google.

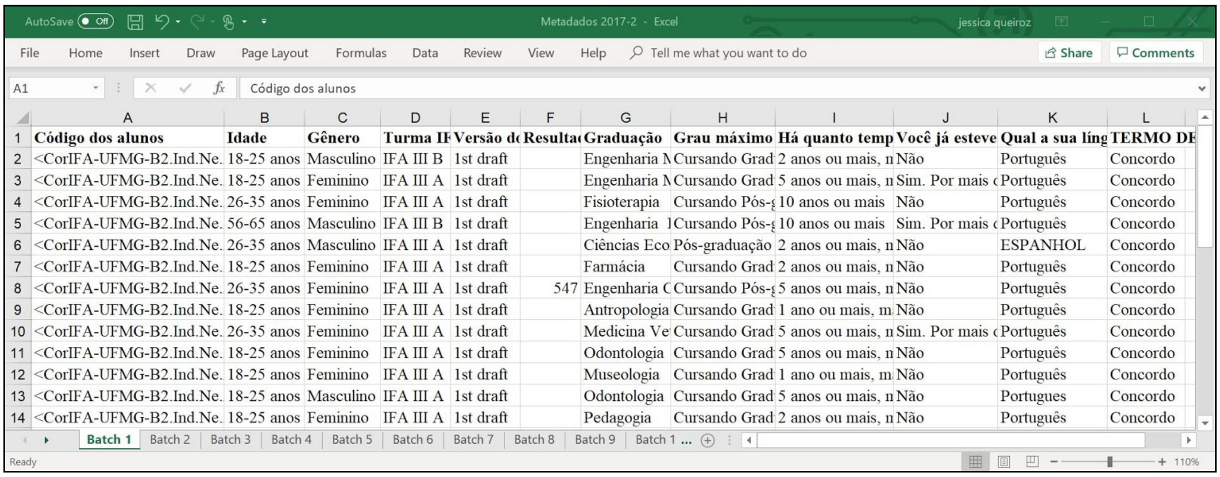
Este conteúdo não foi criado nem aprovado pelo Google. Denunciar abuso - Termos de Serviço - Termos Adicionais

Google Formulários

Source: Provided by the CorIFA team.

All the information sent via Google Form is automatically organized on a Google Spreadsheet. This spreadsheet is later manually edited, as any information that might indicate the identity of a learner has to be excluded or substituted with a code (see FIGURE 3.2 for a sample, where codes that identify students and their texts are represented on column A). Each learner text is also manually coded and organized on individual simple text files (see FIGURE 3.3 for a sample). These files are manually revised and any information that is not considered useful for linguistic analysis, such as titles, quotes, and reference lists, is put in between angle brackets (< >), as these allow that these textual features are ignored during automatic analysis (see section 3.3 for more details).

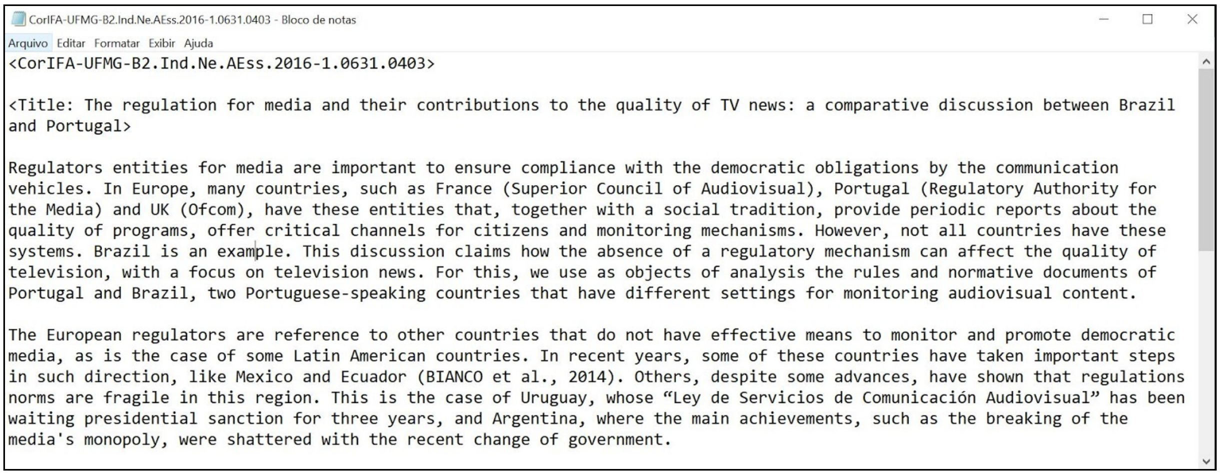
Figure 3.2 – CorIFA: Sample spreadsheet with students' metadata



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Código dos alunos**	**Idade**	**Gênero**	**Turma IF**	**Versão d**	**Resulta**	**Graduação**	**Grau máximo**	**Há quanto temp**	**Você já esteve**	**Qual a sua líng**	**TERMO DE**
2	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Masculino	IFA III B	1st draft		Engenharia M	Cursando Grad	2 anos ou mais, n	Não	Português	Concordo
3	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Engenharia M	Cursando Grad	5 anos ou mais, n	Sim. Por mais (	Português	Concordo
4	<CorIFA-UFMG-B2.Ind.Ne.	26-35 anos	Feminino	IFA III A	1st draft		Fisioterapia	Cursando Pós-g	10 anos ou mais	Não	Português	Concordo
5	<CorIFA-UFMG-B2.Ind.Ne.	56-65 anos	Masculino	IFA III B	1st draft		Engenharia I	Cursando Pós-g	10 anos ou mais	Sim. Por mais (	Português	Concordo
6	<CorIFA-UFMG-B2.Ind.Ne.	26-35 anos	Masculino	IFA III A	1st draft		Ciências Eco	Pós-graduação	2 anos ou mais, n	Não	ESPANHOL	Concordo
7	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Farmácia	Cursando Grad	2 anos ou mais, n	Não	Português	Concordo
8	<CorIFA-UFMG-B2.Ind.Ne.	26-35 anos	Feminino	IFA III A	1st draft	547	Engenharia C	Cursando Pós-g	5 anos ou mais, n	Não	Português	Concordo
9	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Antropologia	Cursando Grad	1 ano ou mais, m	Não	Português	Concordo
10	<CorIFA-UFMG-B2.Ind.Ne.	26-35 anos	Feminino	IFA III A	1st draft		Medicina Ve	Cursando Grad	5 anos ou mais, n	Sim. Por mais (	Português	Concordo
11	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Odontologia	Cursando Grad	5 anos ou mais, n	Não	Português	Concordo
12	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Museologia	Cursando Grad	1 ano ou mais, m	Não	Português	Concordo
13	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Masculino	IFA III A	1st draft		Odontologia	Cursando Grad	5 anos ou mais, n	Não	Portugues	Concordo
14	<CorIFA-UFMG-B2.Ind.Ne.	18-25 anos	Feminino	IFA III A	1st draft		Pedagogia	Cursando Grad	2 anos ou mais, n	Não	Português	Concordo

Source: Provided by the CorIFA team.

Figure 3.3 – CorIFA: Sample of a simple text file


CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0631.0403 - Bloco de notas
Arquivo
Editar
Formatar
Exibir
Ajuda

<CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0631.0403>

<Title: The regulation for media and their contributions to the quality of TV news: a comparative discussion between Brazil and Portugal>

Regulators entities for media are important to ensure compliance with the democratic obligations by the communication vehicles. In Europe, many countries, such as France (Superior Council of Audiovisual), Portugal (Regulatory Authority for the Media) and UK (Ofcom), have these entities that, together with a social tradition, provide periodic reports about the quality of programs, offer critical channels for citizens and monitoring mechanisms. However, not all countries have these systems. Brazil is an example. This discussion claims how the absence of a regulatory mechanism can affect the quality of television, with a focus on television news. For this, we use as objects of analysis the rules and normative documents of Portugal and Brazil, two Portuguese-speaking countries that have different settings for monitoring audiovisual content.

The European regulators are reference to other countries that do not have effective means to monitor and promote democratic media, as is the case of some Latin American countries. In recent years, some of these countries have taken important steps in such direction, like Mexico and Ecuador (BIANCO et al., 2014). Others, despite some advances, have shown that regulations norms are fragile in this region. This is the case of Uruguay, whose "Ley de Servicios de Comunicación Audiovisual" has been waiting presidential sanction for three years, and Argentina, where the main achievements, such as the breaking of the media's monopoly, were shattered with the recent change of government.

Source: Provided by the CorIFA team.

3.2.2 CorIFA subcorpus

For the purposes of this research, a subcorpus of CorIFA of argumentative essays was organized to be used as our research corpus. This subcorpus was sampled based on the design criteria shown in Table 3.1, which guided the collection of first drafts of argumentative essays written by upper intermediate students (B2) who were taking the upper intermediate EAP class at UFMG and whose L1 was Portuguese.

Table 3.1 – CorIFA subcorpus: Design criteria

Genre	Argumentative essay
Medium	Written
Text version	Non-edited
L1	Brazilian Portuguese
English proficiency level	Upper intermediate (B2)
Learning context	EAP

Source: Designed by the author, 2019.

Following these criteria, 114 texts were selected for analysis (see APPENDIX D for a list of the codes of the essays selected). The essays selected were written by 114 learners over the second semester of 2015 and the second semester of 2017. This genre was chosen for being widely required in undergraduate programs (HYLAND, 2009), representative of student academic writing in general, and often used in SLA research (GRANGER, 2012). Essays written only by upper intermediate students were chosen as these students are enrolled in the EAP class (level 3) that concentrates on teaching the general structure and purposes of this genre. More details about the argumentative essay such as its definition, structure, and purposes will be given in section 3.2.3. Students should have Portuguese as their L1, because of the focus on Brazilian learners. Their academic level, either undergraduate or graduate, was not a factor of selection even though it adds a layer of variability to the corpus and could be analyzed in future research.

In total, the subcorpus used in this research contains 51187 word tokens (see TABLE 3.2).[17] Its size might seem small if compared to other learner corpora available (e.g. ICLE, c. 3 million word tokens[18]), but it should be kept in mind that "the optimal size of a learner corpus depends on the targeted linguistic phenomenon" (GRANGER, 2012, p. 9). Knowing that nouns are "by far the most frequent lexical word" (BIBER *et al.*, 1999, p. 65), each argumentative essay should have as many NPs as it is necessary for the present linguistic description since each text in our subcorpus has more than 300 words. The mean number of words per text is 449.0 (see TABLE 3.2). Moreover, all noun-headed NPs extracted from this subcorpus will be manually analyzed for our data analysis (see section 3.3).

Table 3.2 – CorIFA subcorpus: Size

Number of texts	114
Mean number of words per text	449.0 (SD=108.0)
Number of students	114
Word types	6078
Word tokens	51187

Source: Designed by the author, 2019.

It seems worth mentioning that 121 essays that were part of the corpus shared by the CorIFA team had to be removed from our research subcorpus, following the design criteria mentioned above and because of the following reasons. First, 107 texts written in 2013 were excluded as they do not have enough metadata, such as information about students' L1 and academic level. Then, three texts were excluded since they were written by students whose L1 was Spanish, seven had to be left out as theirs writers had written another text on a previous semester (in this case, texts with more words were preferred), and one was a duplicate of another text in the subcorpus. Other three texts had to be excluded after the NPs extraction (presented in section 3.3), because they could not be automatically parsed (perhaps because they had too many long sentences, which seems to be problematic for the automatic parser used). A list of the codes of the excluded essays is organized in Appendix E.

[17] All numbers of word tokens and types given in this thesis were obtained with *AntConc 3.5.7* (ANTHONY, 2018).

[18] Centre for English Corpus Linguistics (August 9, 2018): Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. More information at https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html .

The 114 argumentative essays in our research corpus were further organized and classified in terms of their topics (see APPENDIX F). Based on the tasks given to students, texts were divided into a subcorpus of general topic texts and a subcorpus of specific topic texts. General topic texts were those ones in which the EAP instructor presented a topic or question, e.g. *Does technology makes us more alone?*, to all students to write an argumentative essay about. Many of these topics were similar to the ones used in English proficiency tests. On the other hand, specific topic texts were those in which students were allowed to choose a topic of their preference to write about. Many wrote essays about their graduate studies, such as one dentistry student who wrote about periodontal disease and premature delivery. All in all, our corpus is composed of 46 texts of general topics, which correspond to 18678 word tokens, and 68 texts of specific topics, which correspond to 32509 word tokens (see TABLE 3.3). The difference in the number of texts and words tokens between the two subcorpora should not affect significantly our analysis since the frequencies of the data analyzed will be normalized according to the size of each subcorpus, which should enable the comparison of results.

Table 3.3 – Research corpus: Subcorpora size

Subcorpus		
General topic	Number of texts	46
	Mean number of words per text	406.0 (SD=73.1)
	Word types	2704
	Word tokens	18678
Specific topic	Number of texts	68
	Mean number of words per text	478.1 (SD=118.1)
	Word types	5064
	Word tokens	32509

Source: Designed by the author, 2019.

3.2.2.1 CorIFA subcorpus participants

Some more information about the learners whose essays are part of CorIFA subcorpus could be given based on the metadata available. This information can be essential to writer-oriented analyses, but herein it only serves the purpose of characterizing these students. None of this information will be directly used in this research analysis, because there is no intention to investigate each text or each learner's language use separately. Our main concern is to analyze a collection of texts, having only a general idea about our L2 writers.

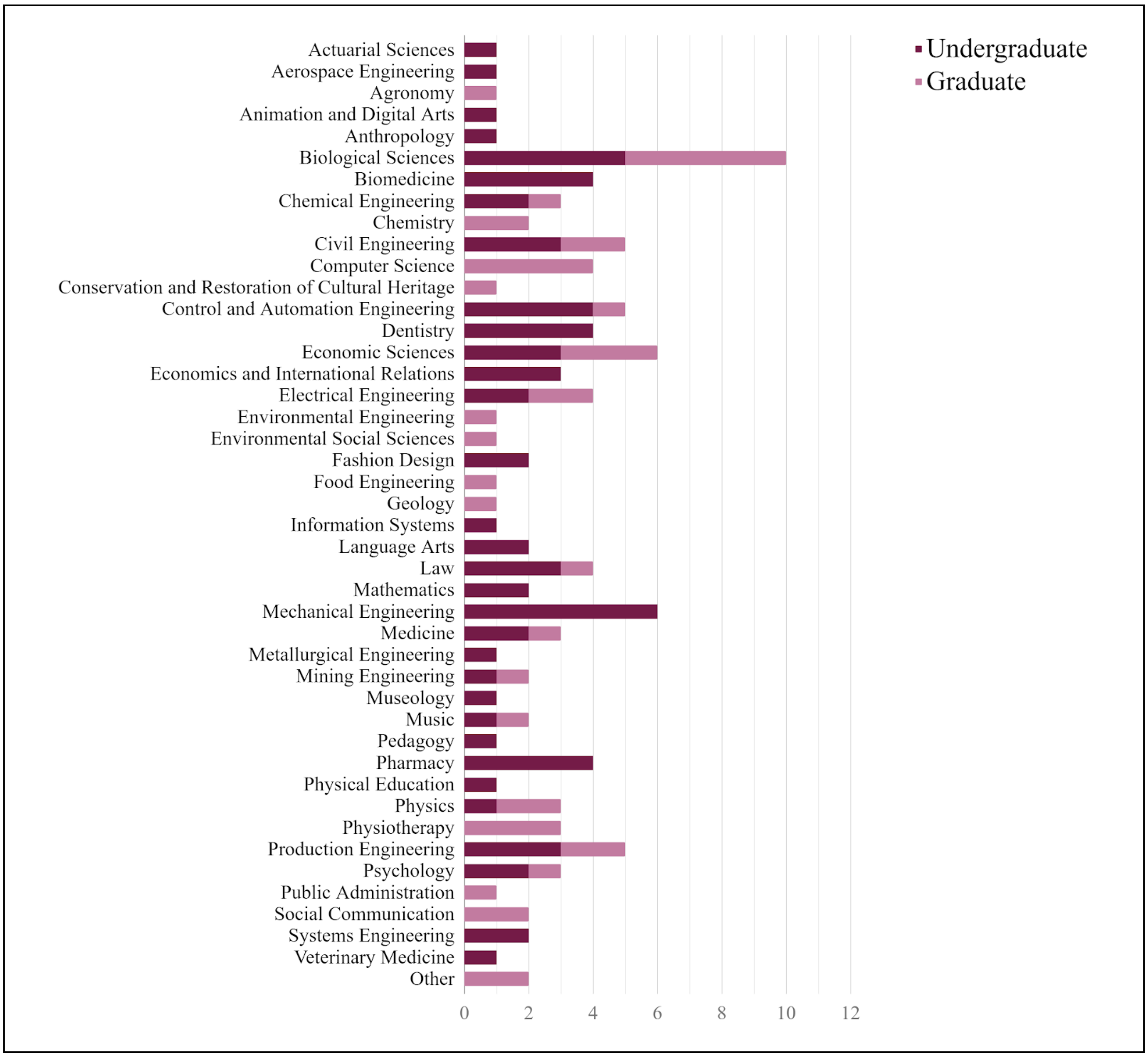
Table 3.4 – CorIFA subcorpus: Participants

		raw	%
Gender	Female	75	65.8
	Male	39	34.2
	Others	-	-
	Prefer to not answer	-	-
	TOTAL	114	100
Age	Less than 18 years old	-	-
	18-25 years old	73	64.0
	26-35 years old	35	30.7
	36-45 years old	3	2.6
	46-55 years old	1	0.9
	56-65 years old	2	1.8
	TOTAL	114	100
Academic level	Undergraduate	71	62.3
	Graduate (Master's)	19	16.7
	Graduate (Doctorate)	24	21.0
	TOTAL	114	100

Source: Designed by the author, 2019.

Table 3.4 presents some details about students' gender, age, and academic level. From 114 students, 75 declared themselves as female and 39 as male; 73 are 18 to 25 years old, 35 are 26 to 35 years old, and 6 are older than 36 years old; 71 are undergraduate students and 43 are members of a graduate program. A more detailed graph that relates students per major and academic level was prepared (see GRAPH 3.1, where c. 43 different majors were represented in alphabetical order).

Graph 3.1 – CorIFA subcorpus: Number of students per major and academic level



Source: Designed by the author, 2019.

It is also interesting to see that 55.2% of participants in our research corpus said that they have studied English for at least five years (see TABLE 3.5). That shows that a good number of our students have had contact with English for a reasonable amount of time even though we do not know under which conditions that contact took place.

Table 3.5 – CorIFA subcorpus: Number of students per time learning English

	raw	%
Never studied	4	3.5
Less than one year	9	7.9
One year or more but less than two years	6	5.3
Two years or more but less than five years	32	28.1
Five years or more but less than ten years	47	41.2
Ten years or more	16	14.0
TOTAL	114	100

Source: Designed by the author, 2019.

Having detailed the research corpus design and characterized its participants, some information about the argumentative essay as a genre is given in the next subsection.

3.2.3 The argumentative essay

The selection of the written genre,[19] argumentative essay, for analysis was not random. This genre is part of EAP programs with genre-based curricula, in which specific academic genres are thoroughly taught to students. In this context, the argumentative essay is seen as a genre that students need to master and, as EAP learners, they learn more about them to support their general L2 English learning (POLIO, 2012, p 139). Moreover, having only one genre under analysis can help us have more control over the contextual variables that influence language variation.

Students should master this genre because the essay "is perhaps the most common undergraduate genre (…) found across the disciplinary spectrum" (HYLAND, 2009, p. 130). Moreover, the argumentative essay has as its fundamental purpose, found throughout academic writing in general, "the presentation of a written argument to defend or explain a position, typically drawing on library sources" (HYLAND, 2009, p. 130). Therefore, working with it in class can be a first step to introduce EAP students to the perspective of writing as a

[19] In this research, genre is defined according to the discussion developed by Biber and Conrad (2009) as a set of specific texts that can be grouped together because of certain cultural textual and linguistic patterns shared among them.

process in the construction of effective arguments (PARKINSON; MUSGRAVE, 2014). Furthermore, L2 learners, who want to be part of their disciplinary communities, are expected to learn how "texts are organized and the lexico-grammatical patterns that are typically used to express meanings in the genre[s]" specific to their research field (HYLAND, 2004, p. 12).

Students, then, learn that in this genre they must defend a stance and provide evidence to support it, i.e. writing skills and purposes that are inherent in most academic written genres. It then allows students to work on recurrent structures and lexico-grammatical patterns found in the academic community (NUNAN, 2008). Argumentative essays are academic, even though some researchers might see an issue of authenticity when it comes to writing a text for assessment, and could be expected to display the use of the grammatical complexity features presented earlier (PARKINSON; MUSGRAVE, 2014). All of that can help students participate effectively in the world outside the ESL classroom (HYLAND, 2004).

Considering that grammatical complexity is a characteristic of professional academic written texts, EAP programs should add the study of this feature to their materials. However, it is equally necessary that L2 learners' texts are analyzed so as to have an overview on their knowledge about simple and complex NPs. A corpus-based investigation as it is proposed here offers the possibility to put students' written production and L2 learners' language in use under scrutiny, which could show their L2 competence and their awareness to the preferred grammatical features in their disciplinary written discourses. All academic writers have to be aware of the fact that information has been widely packed in phrasal structures, particularly in complex NPs, "to exploit them effectively" (BIBER *et al.*, 1999, p. 44).

Having presented all the information necessary about the research corpus, let us explore in the next section the procedures used in the annotation of our data.

3.3 Data retrieval and analysis

3.3.1 Corpus annotation

Corpus annotation is quite common in linguistics and it can be defined as "the process of providing – in a systematic and accessible form – those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they worked with" (McENERY; HARDIE, 2012, p. 13). Annotation, then, means storing the analysis done somehow as a means to share them with others. Such kind of analyses can be syntactic, semantic, discursive, pragmatic, etc., but, in all cases, this process is of paramount importance for replicability of research as any researcher with access to the analysis done in a corpus is able to scrutinize it (McENERY; HARDIE, 2012).

Several researchers, especially in corpus linguistics, have been using automated annotation programs as part of their analyses. However, despite the great advances in linguistics and programming, our understanding of language still does not provide automatic analyses without certain inconsistencies. For that reason, annotation is often followed by manual correction (McENERY; HARDIE, 2012).

In this thesis, considering the wish to extract and analyze all, if possible, NPs used in learner written texts, it would be necessary to segment the sentences of the corpus into phrases and, then, consistently and reliably extract the NPs identified. As can be imagined, such work of segmentation and extraction would be quite complicated if done manually, and as will be shown below, much can be done with the help of automated methods which are freely available.

The segmentation of sentences, in linguistics, is known as parsing. Since we want to syntactically segment written sentences, we need to use a constituency parser, in which the program, based on predefined rules of tagging and parsing, identify the syntactic units or phrases that constitute each sentence in the corpus texts (MEYER, 2002). The parser chosen for that job was part of the Stanford Core Natural Language Processing – Stanford CoreNLP, henceforth – toolkit (MANNING *et al.*, 2014), which gave us better results and seemed easier to use than the Natural Language Toolkit (NLTK) parser or the Python libraries TextBlob[20] and spaCy[21]. In order to use the Stanford CoreNLP parser, a Python script was written by a professional programmer. The extraction of the NPs identified after parsing was done with another Python script.[22] More details are given below.

The Stanford CoreNLP was developed in 2006 and four years later released to be used as a free open source software (MANNING *et al.*, 2014). Since then, it is one of the most used natural language analysis toolkits in research, as it is useful for several types of automated linguistic analyses, such as morphological, syntactic, sentiment analysis, and others, and it is

[20] More information at https://textblob.readthedocs.io/en/dev/ .

[21] More information at https://spacy.io/ .

[22] The Python scripts used in this study were written by the programmer Euller Borges.

regularly updated for not only English but also Arabic, Chinese, French, just to name a few languages (cf. MANNING *et al.*, 2014, p. 55).

Our interest was in syntactically parsing our data. That encompasses some of the tools provided by the Stanford CoreNLP, i.e. tokenizing, sentence splitting, tagging, parsing the data.[23] Each one of these is defined in Manning *et al.* (2014) and it is understood that the constituency parsing of a text includes that it is first tokenized into a sequence of tokens and these tokens are then split into sentences. Having each sentence identified and each token organized individually, POS labels are given to each token and groups of POS-labeled tokens are probabilistic organized as phrases and labeled.[24] Below is an example of a sentence from CorIFA which was parsed with the Stanford CoreNLP toolkit.

```
(ROOT
  (S
    (NP
      (NP (NNS Regulators) (NNS entities))
      (PP (IN for)
        (NP (NNS media))))
    (VP (VBP are)
      (ADJP (JJ important)
        (S
          (VP (TO to)
            (VP (VB ensure)
              (NP (NN compliance))
              (PP (IN with)
                (NP (DT the) (JJ democratic) (NNS obligations)))
              (PP (IN by)
                (NP (DT the) (NN communication) (NNS vehicles))))))))
    (. .)))
```

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0631.0403

This parsed example is helpful in understanding the arrangement of a sentence in terms of phrases and its constituents. Each line, after the first left parenthesis, starts with its respective clause-level, e.g. S or SBAR, or phrase-level, e.g. NP, VP, or PP, tag. Similarly, on the left side of each word token comes its POS tag, such as NNS or IN (see APPENDIX A for the tagset). Moreover, the indentation provided in each line shows the phrases which are directly connected to the main clause and the phrases inside other phrases. For instance, the

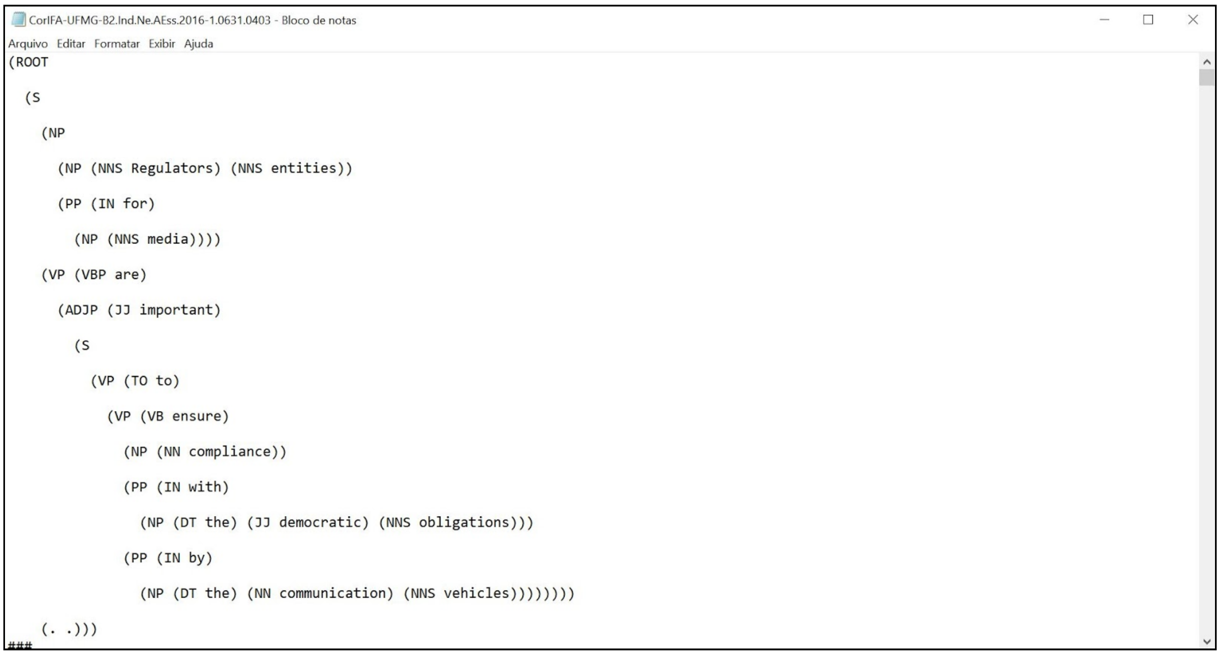
[23] It is possible to see a demonstration of the parser results at http://nlp.stanford.edu:8080/parser/index.jsp .

[24] The POS and phrasal labels used in the Stanford CoreNLP come primarily from the work developed by the Penn Treebank, which is a project developed from 1989 to 1996 responsible for designing three annotation schemes: POS tagging, syntactic bracketing, and disfluency annotation (TAYLOR *et al.*, 2003).

sequence of words *Regulators entities for media* is parsed as an NP, whose constituents are the NP *regulators entities* and the PP *for media*, which contains the NP *media*.

In order to have all the processes working, it was necessary to use a programming language to access the Stanford CoreNLP toolkit and have it parse several individual simple text files stored in two directories, one with text files from the general topic subcorpus and another with the text files from the specific topic subcorpus. For that purpose, we used Python 3.7.0[25] and created one first script that would consistently go over the directories and files and parse the written texts in there. As we were using Python and the Stanford CoreNLP, which is primarily designed for Java, it was important to use a Python interface of the toolkit. The one used was the version 3.9.1.1 developed by Lynten Guo.[26] It was also necessary to download a set of files at the Stanford CoreNLP website in order to have all its tools, version 3.9.1.1 also, working.[27]

Figure 3.4 – Python script 1: Sample of a simple text file with parsed data


CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0631.0403 - Bloco de notas
Arquivo Editar Formatar Exibir Ajuda

```
(ROOT
  (S
    (NP
      (NP (NNS Regulators) (NNS entities))
      (PP (IN for)
        (NP (NNS media))))
    (VP (VBP are)
      (ADJP (JJ important)
        (S
          (VP (TO to)
            (VP (VB ensure)
              (NP (NN compliance))
              (PP (IN with)
                (NP (DT the) (JJ democratic) (NNS obligations)))
              (PP (IN by)
                (NP (DT the) (NN communication) (NNS vehicles))))))))
    (. .)))
###
```

Source: Designed by the author, 2019.

The first Python script (see APPENDIX G) uses several functions to do the step-by-step process of having each learner text that composed the research corpus parsed. In

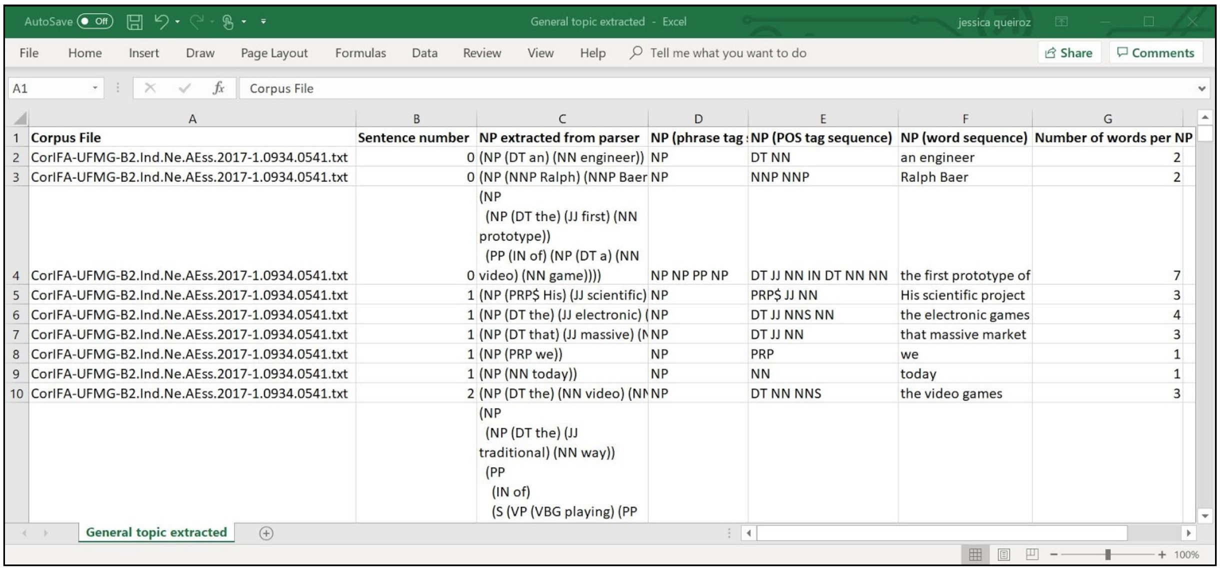
[25] More information and download at https://www.python.org/downloads/ .
[26] More information at https://github.com/Lynten/stanford-corenlp .
[27] More information and dowload available at https://stanfordnlp.github.io/CoreNLP/index.html#download .

a few lines, the source directory where the texts are stored is first opened and each simple text file in it is opened and read. As each file is read, the Stanford CoreNLP tools necessary for constituency parsing annotate each text. As said previously, any information in between angle brackets is ignored during this procedure and any file that cannot be parsed sends back a warning message but does not stop the script from running. After the annotation, each parsed text is saved in a new file with the same title as the original file and stored into a new output directory. See Figure 3.4 for an example of text file saved with the parsed data and compare it to Figure 3.3 presented in section 3.2 of this chapter.

3.3.2 Data extraction

Once all data is parsed and saved in a new directory, it is time to extract the data which will be later analyzed in this study. It is essential to create a systematic, consistent, and reliable manner to retrieve the data. Therefore, to have all NPs extracted from the parsed files, another Python script was written (see APPENDIX H). For this script though, Python 3.6.6 was used as one of the tools used during the process, the NLTK 3.4[28] developed by Steven Bird, would not work with Python 3.7.0.

Figure 3.5 – Python script 2: Sample of a spreadsheet with the NPs extracted



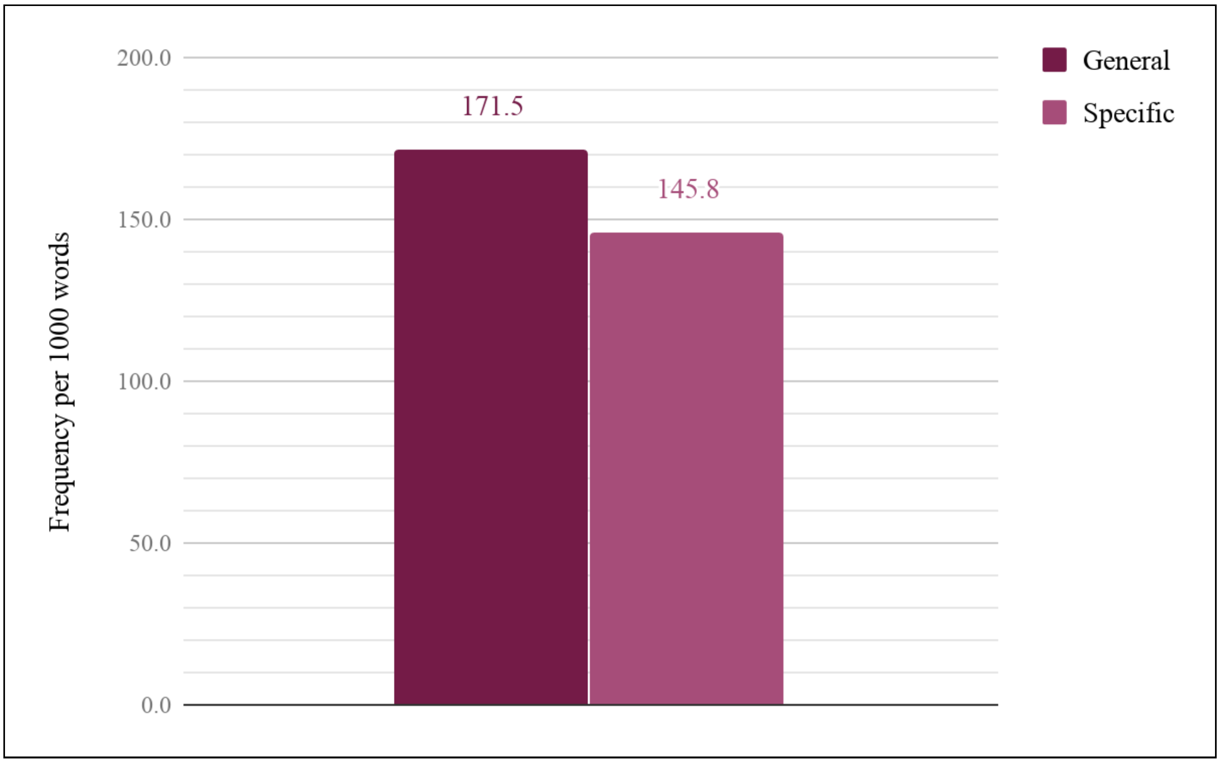
	A	B	C	D	E	F	G
1	Corpus File	Sentence number	NP extracted from parser	NP (phrase tag	NP (POS tag sequence)	NP (word sequence)	Number of words per NP
2	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	0	(NP (DT an) (NN engineer))	NP	DT NN	an engineer	2
3	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	0	(NP (NNP Ralph) (NNP Baer	NP	NNP NNP	Ralph Baer	2
4	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	0	(NP (NP (DT the) (JJ first) (NN prototype)) (PP (IN of) (NP (DT a) (NN video) (NN game))))	NP NP PP NP	DT JJ NN IN DT NN NN	the first prototype of	7
5	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	1	(NP (PRP$ His) (JJ scientific)	NP	PRP$ JJ NN	His scientific project	3
6	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	1	(NP (DT the) (JJ electronic) (	NP	DT JJ NNS NN	the electronic games	4
7	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	1	(NP (DT that) (JJ massive) (N	NP	DT JJ NN	that massive market	3
8	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	1	(NP (PRP we))	NP	PRP	we	1
9	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	1	(NP (NN today))	NP	NN	today	1
10	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	2	(NP (DT the) (NN video) (NN	NP	DT NN NNS	the video games	3
			(NP (NP (DT the) (JJ traditional) (NN way)) (PP (IN of) (S (VP (VBG playing) (PP				

Source: Designed by the author, 2019.

[28] More information at https://www.nltk.org/ .

This second script is similar to the first one presented in the last section, for it opens the output directories where the parsed files are stored. Then, it opens and reads the text files. From each text file, the groups parsed as NP are identified and extracted to a spreadsheet, which is opened separately and saved as a .csv file at the final stage of the process. It is crucial to report that NPs inside other NPs were not extracted separately, as these would inflate the size of data in our analysis and generate incorrect results.

Graph 3.2 – NPs extracted by the parser (per subcorpus)

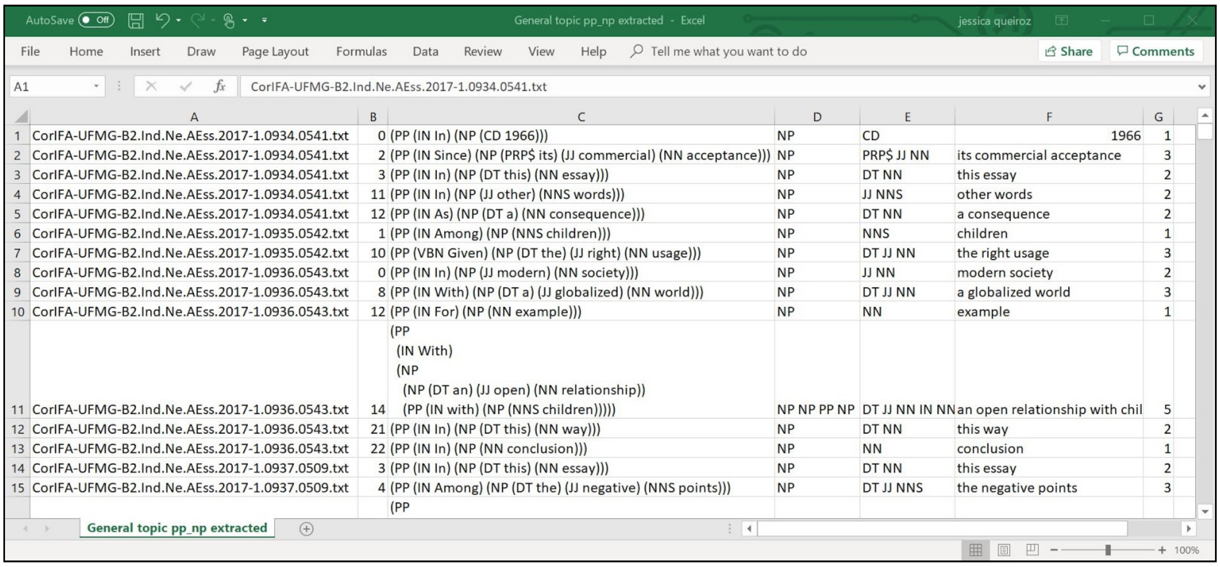


Source: Designed by the author, 2019.

Two other pieces of information are also elicited from each text file, i.e. the title of the file from which each NP was taken and the number of the sentence where each NP can be found. These are organized in the spreadsheet in columns A, B, and C (see FIGURE 3.5). Together with those, other information is added to the spreadsheet so as to help in the manual organization of the data. As can be seen in Figure 3.5, these are: the sequence of phrase tags in each NP (column D), the sequence of POS tags that each word token received (column E), the sequence of word tokens without any tags (column F), and the number of word tokens per phrase (column G). Once all that is organized in the spreadsheets, they are saved in the output directory and ready to be manually checked. As a result, two spreadsheets were created, one with the NPs extracted from the general topic subcorpus and another with the NPs from the specific topic subcorpus. This is further explored in Chapter 4, but 7944 groups parsed as NPs were extracted from both subcorpora (see GRAPH 3.2 for more details).

At the same time that the process described above was happening, NPs that were inside PPs were differently treated by the script. This was a decision made after considering that some NPs that were part of PPs that function as adverbials could not be analyzed together with NPs that have other functions and occur by themselves in sentences. The condition added to identify these PPs was that they should be directly subordinated to the label ROOT of the sentences. These PPs, which were 415 in total, were saved into separate spreadsheets (see FIGURE 3.6). That way, it was possible to keep the analysis of PPs that were constituents of NPs only and not the other way around. For this thesis, no analysis was proposed for these PPs or the NPs that complement the prepositions.

Figure 3.6 – Python script 2: Sample of a spreadsheet with the adverbial PPs extracted



CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt

	A	B	C	D	E	F	G
1	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	0	(PP (IN In) (NP (CD 1966)))	NP	CD	1966	1
2	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	2	(PP (IN Since) (NP (PRP$ its) (JJ commercial) (NN acceptance)))	NP	PRP$ JJ NN	its commercial acceptance	3
3	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	3	(PP (IN In) (NP (DT this) (NN essay)))	NP	DT NN	this essay	2
4	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	11	(PP (IN In) (NP (JJ other) (NNS words)))	NP	JJ NNS	other words	2
5	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541.txt	12	(PP (IN As) (NP (DT a) (NN consequence)))	NP	DT NN	a consequence	2
6	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0935.0542.txt	1	(PP (IN Among) (NP (NNS children)))	NP	NNS	children	1
7	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0935.0542.txt	10	(PP (VBN Given) (NP (DT the) (JJ right) (NN usage)))	NP	DT JJ NN	the right usage	3
8	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	0	(PP (IN In) (NP (JJ modern) (NN society)))	NP	JJ NN	modern society	2
9	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	8	(PP (IN With) (NP (DT a) (JJ globalized) (NN world)))	NP	DT JJ NN	a globalized world	3
10	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	12	(PP (IN For) (NP (NN example)))	NP	NN	example	1
11	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	14	(PP (IN With) (NP (NP (DT an) (JJ open) (NN relationship)) (PP (IN with) (NP (NNS children)))))	NP NP PP NP	DT JJ NN IN NN	an open relationship with chil	5
12	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	21	(PP (IN In) (NP (DT this) (NN way)))	NP	DT NN	this way	2
13	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543.txt	22	(PP (IN In) (NP (NN conclusion)))	NP	NN	conclusion	1
14	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509.txt	3	(PP (IN In) (NP (DT this) (NN essay)))	NP	DT NN	this essay	2
15	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509.txt	4	(PP (IN Among) (NP (DT the) (JJ negative) (NNS points)))	NP	DT JJ NNS	the negative points	3
			(PP				

Source: Designed by the author, 2019.

3.3.3 Data categorization

Having the NPs from our research corpus extracted, it was then possible to categorize them according to the structural combinations of NP constituents (see Chapter 2) and to create the graphs and tables shown in Chapter 4 (for that purpose, Excel tools were used). Some decisions had to be made in order to reorganize the 7944 rows of NPs into the following categories:

1. **Simple NP**
   a. Noun head alone
   b. Determiner(s) + noun head
2. **Complex NP**
   a. Premodifier(s) + noun head
   b. Noun head + postmodifier(s)
   c. Premodifier(s) + noun head + postmodifier(s)
3. **Coordinated head nouns**
   a. Simple heads
   b. Complex heads
   c. Simple head(s) + Complex head(s)
4. **NPs excluded from analysis**
   a. Groups headed by a word that is not a noun
   b. Groups misparsed as NPs
   c. Other types of NPs not used

Having those categories in mind, it seemed logical to separate the NPs based on their sequences of phrase tags. The spreadsheet rows were thus reorganized alphabetically after column D and the rows containing the sequences of phrase tags constituted by the tag NP by itself were cut and pasted into a new sheet. The rows with sequences of phrase tags constituted by the tag NP followed by other phrase tags were cut and pasted into another sheet. That allowed the identification, among the sequences with the NP tag by itself, of most NPs categorized as: simple NP with head noun alone (1a), simple NP with determiner(s) and head noun (1b), and complex NP with premodifier(s) (2a). See examples (1) to (3) representing each one of these categories, respectively. Among the sequences with the NP tag followed by other phrase tags, it was possible to identify the NPs to be categorized as: complex NP with postmodifier(s) (2b), complex NP with pre- and postmodifiers (2c), NPs with coordinated head nouns (3a, 3b, and 3c). For all the sequences identified, particularly for the ones corresponding to the latter categories, a thorough manual checking had to be done.

1) [**Loneliness**] affects all of us at some point in our lives and (…)
(NP (NN Loneliness))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0741.0471 (General)

2) With talent a person can get anything that she wants which in accordance with [her **expertise**].
(NP (PRP$ her) (NN expertise))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0732.0463 (General)

3) Many games (…) induces kids to [aggressive **behavior**].
(NP (JJ aggressive) (NN behavior))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509 (General)

The manual checking consisted of varied steps, but the most important one of them involved the search, using the shortcut Ctrl+F, of the POS tags and certain sequences of POS tags that were more representative of each category, e.g. the sequences JJ NN or NN NN for NPs containing premodifiers or the tag CC for NPs with coordinated noun heads. That process took a couple of weeks to get done and another couple of weeks for revision. During that time, we have seen some minor mistakes done by the parser. These mistakes, such as a few cases similar to the one shown in example (4), where adverbs which were not part of the NP would come at the end of the word group, could be corrected and still be considered for analysis.

4) In conclusion, games do have influence on [**children**], however not as it is (…)
(NP (NP (NNS children)) (, ,) (ADVP (RB however)))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0944.0549 (General)

Some other word groups parsed as NPs had to be put into the category of NPs excluded from analysis (4a, 4b, and 4c), meaning they were not part of our operational definition of NP or they represented other mistakes done by the parser. Consequently, NPs headed by other words, such as pronouns, adjectives, determiners, and numerals, should not be taken into consideration during analysis and had to be ignored in our analysis (4a). That process involved searching for the POS tags for pronouns, adjectives, determiners, and numerals that happened to be alone, in final position of short word groups, or in initial position of very long word groups, as in examples (5) and (6).

5) Even that [**most** of the greek ancient stories] had some kind of magic or natural gift (…)
(NP (NP (JJS most)) (PP (IN of) (NP (DT the) (JJ greek) (JJ ancient) (NNS stories))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0728.0460 (General)

6) Also, [**someone** that speaks a second language] is, many times, a requirement for a job.
(NP (NP (NN someone)) (SBAR (WHNP (WDT that)) (S (VP (VBZ speaks) (NP (DT a) (JJ second) (NN language))))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0599.0373 (General)

Some of those mistakes usually included the parsing of word groups that were not NPs, such as full clauses similar to example (7), groups composed of Portuguese words as in example (8), of references to authors and years of publication such as example (9), and of copies of the general topics given by the EAP instructor like example (10). In those cases, all these word groups could not be taken under analysis and were added to the categories of groups misparsed as NP (4b) and types of NPs that could not be used for analysis (4c).

7) (…) due to [the **training**] is just a way to improve the gift.
(NP (NP (DT the) (NN training)) (SBAR (S (VP (VBZ is) (ADVP (RB just)) (NP (DT a) (NN way) (S (VP (TO to) (VP (VB improve) (NP (DT the) (NN gift))))))))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0718.0450 (General)

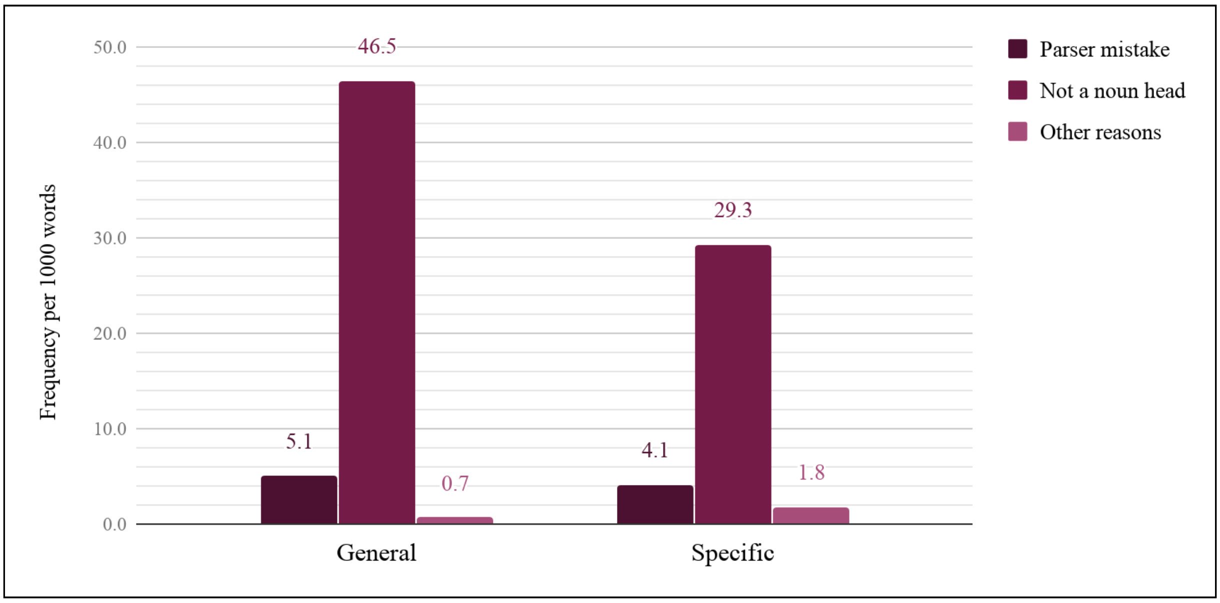
8) Prado actully is so influent that his ideas founded an “school of thinking” named [“**Sentido da Colonização**”].
(NP (NNP Sentido) (NNP da) (NNP Colonização))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0644.0415 (Specific)

9) (…) although two versions are currently highlighted by researches like [**Pujol (1930)**].
(NP (NP (NNP Pujol)) (PRN (-LRB- -LRB-) (NP (CD 1930)) (-RRB- -RRB-)))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0647.0417 (Specific)

10) 'Achieved or able to' expert level is [**question** that return for discussion: "Which is more important: talent or hard work?"].
(NP (NP (NN question)) (PP (IN that) (NP (NP (NN return)) (PP (IN for (NP (NP (NN discussion)) (: :) (`` ``) (SBARQ (WHNP (WP Which)) (SQ (VP (VBZ is) (ADJP (RBR more) (JJ important)))) (: :) (NP (NP (NN talent)) (CC or) (NP (JJ hard) (NN work))) (. ?)))))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0725.0457 (General)

Graph 3.3 shows the number of NPs excluded from analysis based on the reasons mentioned above (more details regarding raw counts are given in Chapter 4). Clearly, most of these NPs were the ones headed by words which are not classified as common or proper nouns.

Graph 3.3 – NPs excluded from analysis (per subcorpus)



Source: Designed by the author, 2019.

In the chapter that follows, the results obtained after the treatment of this research data will be presented and discussed.

CHAPTER 4

RESULTS AND DISCUSSION

Chapter 3 specified the methods that were selected and adopted to empirically investigate the research proposition of describing the grammatical complexity of NPs in learner writing. This chapter examines quantitatively and qualitatively the outcomes after the data categorization in the attempt to learn more about Brazilian students' use of NPs in academic texts. Section 4.1 reports the general results obtained in this study. Sections 4.2 and 4.3 detail the use of simple and complex NPs, respectively, exploring certain configurations of the NP as used by learners.

4.1 Overview of results

As shown in Chapter 3, the research corpus used in this thesis was a subcorpus of CorIFA composed of 114 argumentative essays divided in two subcorpora, one of general topic texts and another of specific topic texts. From this corpus, it was possible to automatically extract all the NP groups parsed by the Stanford CoreNLP toolkit. The total raw number of NPs extracted in the process was 7944, being 3204 NPs (40.3%)[29] from the general topic subcorpus and 4740 NPs (59.7%) from the specific topic subcorpus.

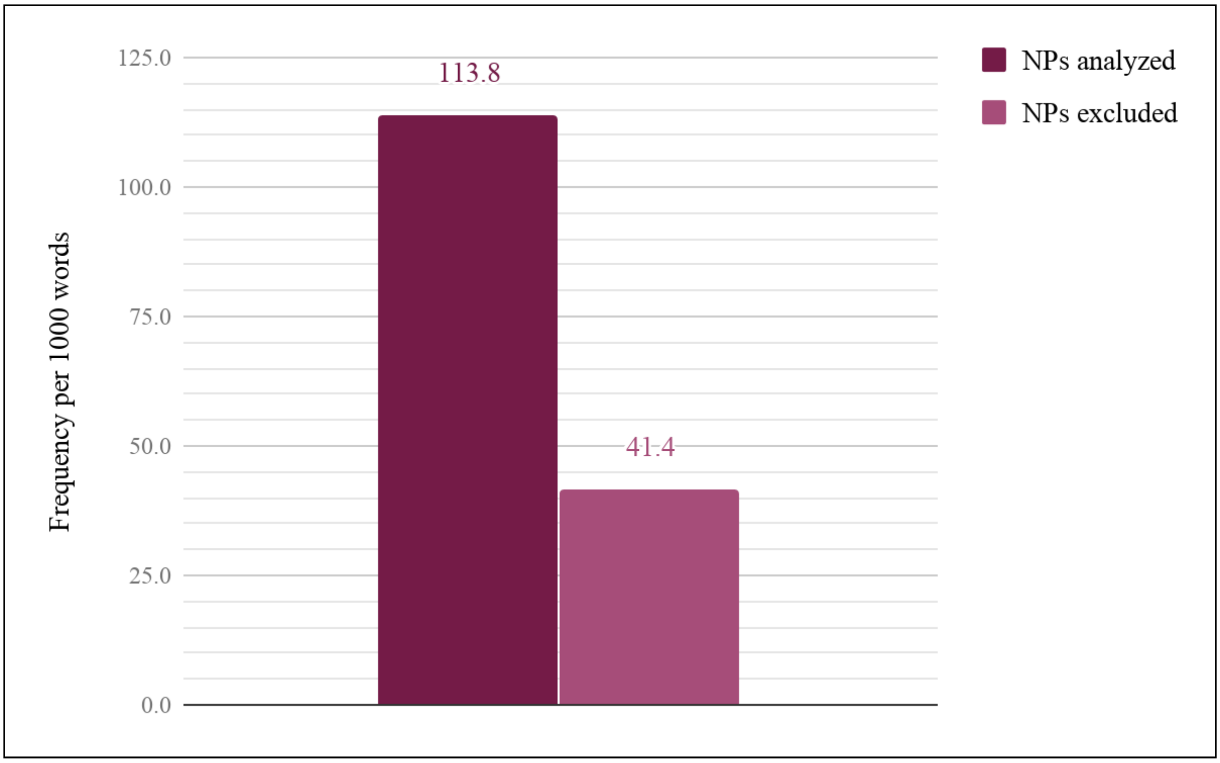
Nonetheless, it would not be appropriate to analyze these groups indiscriminately as there were a few NP heads that should not be taken into consideration in this research, e.g. pronoun heads, illustrated in (a), in which the pronoun *someone* is postmodified by a *that* clause, making it a complex NP. There were also some mistakes done by the parser, e.g. full clauses parsed as NPs, exemplified in (b) (see details in Chapter 3). For that reason, the NPs extracted were scrutinized and the ones that were apt for analysis, basically those NPs that had a noun as their headword, were separated from the ones that should be excluded from the analysis. As a consequence, 5823 NPs (73.3%) extracted from the research corpus were analyzed while 2121 (26.7%) had to be removed from the data analysis (see GRAPH 4.1).

[29] When necessary, percentage rates or raw frequencies are given in between parentheses.

a) Also, [**someone** that speaks a second language] is, many times, a requirement for a job.
(NP (NP (NN someone)) (SBAR (WHNP (WDT that)) (S (VP (VBZ speaks) (NP (DT a) (JJ second) (NN language))))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0599.0373 (General)

b) (…) due to [the **training** is just a way to improve the gift].
(NP (NP (DT the) (NN training)) (SBAR (S (VP (VBZ is) (ADVP (RB just)) (NP (DT a) (NN way) (S (VP (TO to) (VP (VB improve) (NP (DT the) (NN gift))))))))))
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0718.0450 (General)

Graph 4.1 – NPs extracted: NPs analyzed vs. NPs excluded (whole corpus)

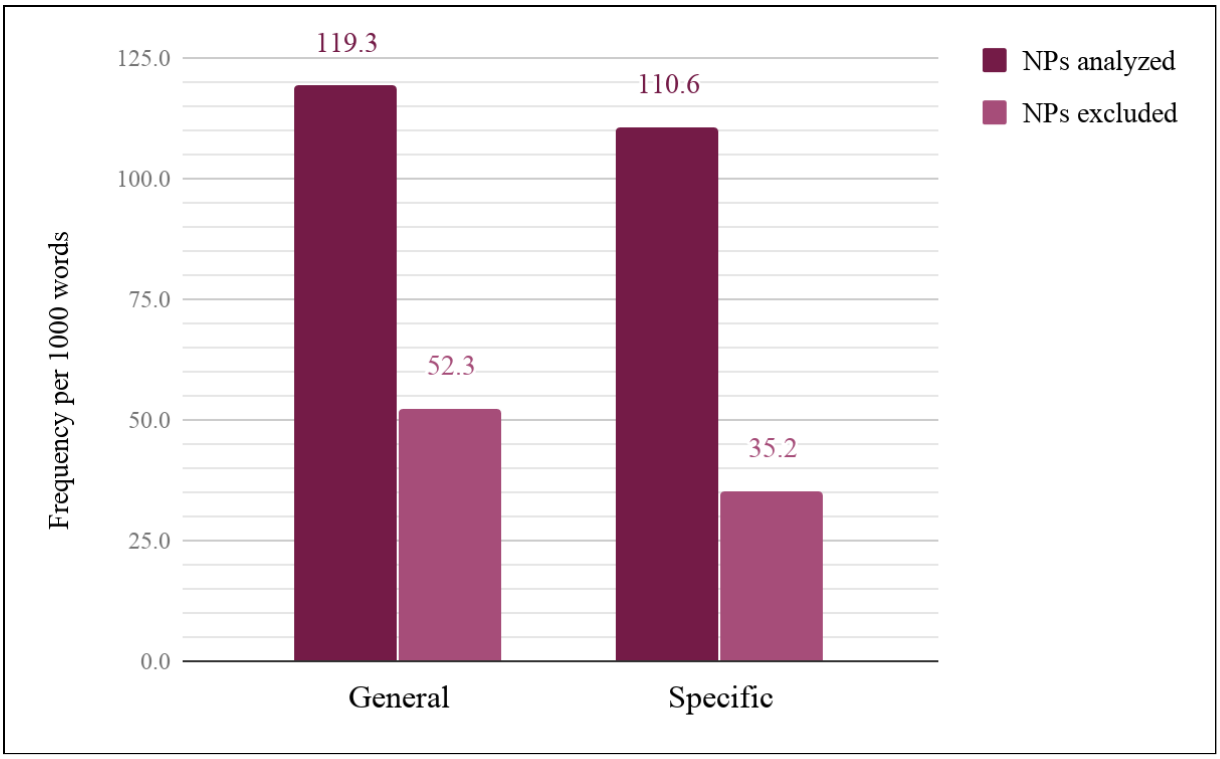


Source: Designed by the author, 2019.

Considering the difference in the number of words in each subcorpus of this study (see details Chapter 3), it would be relevant to see the frequency of analyzed NPs compared to the frequency of excluded NPs in both subcorpora (see GRAPH 4.2). In the general topic subcorpus, 69.5% (2228) of the NPs extracted were analyzed while 30.5% (976) of them were excluded. In the specific topic subcorpus, 75.8% (3595) of the NPs extracted were analyzed while 24.2% (1145) of them were ignored. There is a small difference in the proportion of

NPs analyzed and excluded between the two subcorpora, but we will remedy this difference by counting each category under analysis per subcorpus and normalizing frequencies per 1000 words in relation to the total number of word tokens in each subcorpus, as it has been done in Graph 4.2.

Graph 4.2 – NPs extracted: NPs analyzed vs. NPs excluded (per subcorpus)



Source: Designed by the author, 2019.

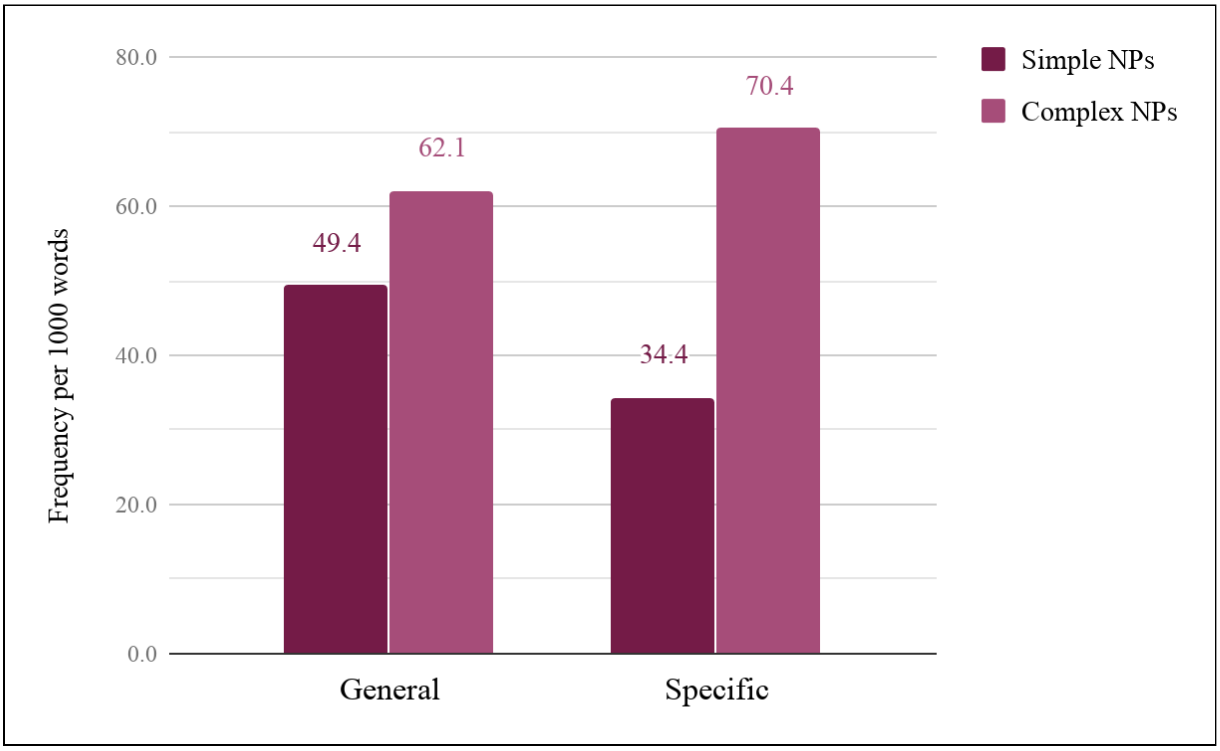
Having seen the frequencies of NPs to be used for analysis, it is possible to look at the number of simple and complex NPs found in each subcorpus (see GRAPH 4.3). The general topic subcorpus has 922 simple NPs and 1160 complex NPs, while the specific topic subcorpus has 1117 simple NPs and 2290 complex NPs. Proportionately, our whole research corpus is composed of 35% of simple NPs and 59.3% of complex NPs. The other 5.7% of NPs analyzed correspond to coordinated head nouns to be discussed in section 4.4 of this chapter.

It is beyond doubt that in both subcorpora complex NPs are more frequent than simple NPs. That evidence refutes our first hypothesis that Brazilian learners use more simple NPs than complex NPs (see Chapter 1) and contradicts Nitsch's (2017) results, in which 74% of NPs produced by the Brazilian learners analyzed by her were simple and 26% were complex.

We assume that the higher use of simple NPs in Nitsch (2017) is a consequence of the analysis of general topic essays only.

The higher frequency of complex NPs in our corpus might suggest that students' English proficiency level (B2/upper intermediate) and the academic context of writing justify this result. This is a functional justification that conforms to our theoretical framework, in which functional reasons and communicative purposes are directly responsible for the grammatical patterns in a language and in a particular genre (BIBER, 2010; BIBER; CONRAD, 2009). Moreover, the larger difference in the use of simple and complex NPs in the specific topic subcorpus is quite reasonable because the specificity of the topics chosen by students means the texts are more specialized and particular to a discourse community and, for that reason, the writers will rely on more complex and compressed structures to express specialized knowledge. Still, both groups of students have produced more complex NPs independent of the essay topics being general or specific.

Graph 4.3 – Simple NPs vs. Complex NPs (per subcorpus)



Source: Designed by the author, 2019.

In view of the overall results of our study, it also seemed important to count the number of NPs produced in each one of the 114 argumentative essays analyzed. By having those numbers, we could calculate the mean numbers of NPs per text and their standard deviations (see TABLE 4.1).

Table 4.1 – Means and standard deviations of NPs per text

	General topic subcorpus		**Specific topic subcorpus**		**Whole corpus**	
Types of NP	Mean	SD	Mean	SD	Mean	SD
Simple NPs	**20.0**	**7.5**	**16.4**	**8.4**	**17.9**	**8.2**
noun head alone	8.6	4.6	6.7	5.0	7.5	4.9
determiner + noun head	11.5	4.9	9.7	4.9	10.4	5.0
Complex NPs	**25.2**	**6.5**	**33.7**	**9.7**	**30.3**	**9.5**
premodifier	10.4	4.7	12.5	5.1	11.6	5.0
postmodifier	10.9	4.2	14.0	5.3	12.7	5.1
both pre- and postmodifiers	3.9	2.3	7.3	3.9	5.9	3.7
Coordinated noun heads	**3.2**	**2.5**	**2.8**	**2.0**	**2.9**	**2.2**
simple heads	1.2	1.2	0.7	0.9	0.9	1.0
complex heads	1.0	1.1	1.3	1.3	1.2	1.2
simple head(s) + complex head(s)	1.0	1.4	0.8	0.9	0.9	1.1

Source: Designed by the author, 2019.

Based on the mean scores, it is possible to see the variation in the use of NPs across texts and subcorpora. For instance, the average of simple NPs is higher in the general topic subcorpus (20.0 vs. 16.4) whereas the average of complex NPs is higher in the specific topic subcorpus (33.7 vs. 25.2), as we have seen with the normalized frequencies shown in Graph 4.3. This suggests that each general topic text tends to have more simple NPs while each specific topic text tend to have more complex NPs.

Based on the standard deviations, it is possible to describe the dispersion of NPs across texts relative to the mean scores found. For example, the mean for complex NPs (25.2) found in the general topic subcorpus is about four times its standard deviation (6.5), meaning one or more texts could have 18.7 complex NPs in them while other texts could have 31.7 complex NPs. As for the specific topic subcorpus mean, it is 33.7, which is more than three times its standard deviation (9.7), showing that the number of complex NPs in the texts can substantially deviate from its mean number. Therefore, in both subcorpora there is a great deal of variation in the number of complex NPs per text. However, the mean score in the specific topic subcorpus is clearly higher (33.7). That can lead us to speculate that specific topic essays allow learners to use a good number of complex NPs which are very likely discipline specific, such as the NP *an anisotropy that depends of the crystalline direction* (shown in example (33) and explored in section 4.3.2).

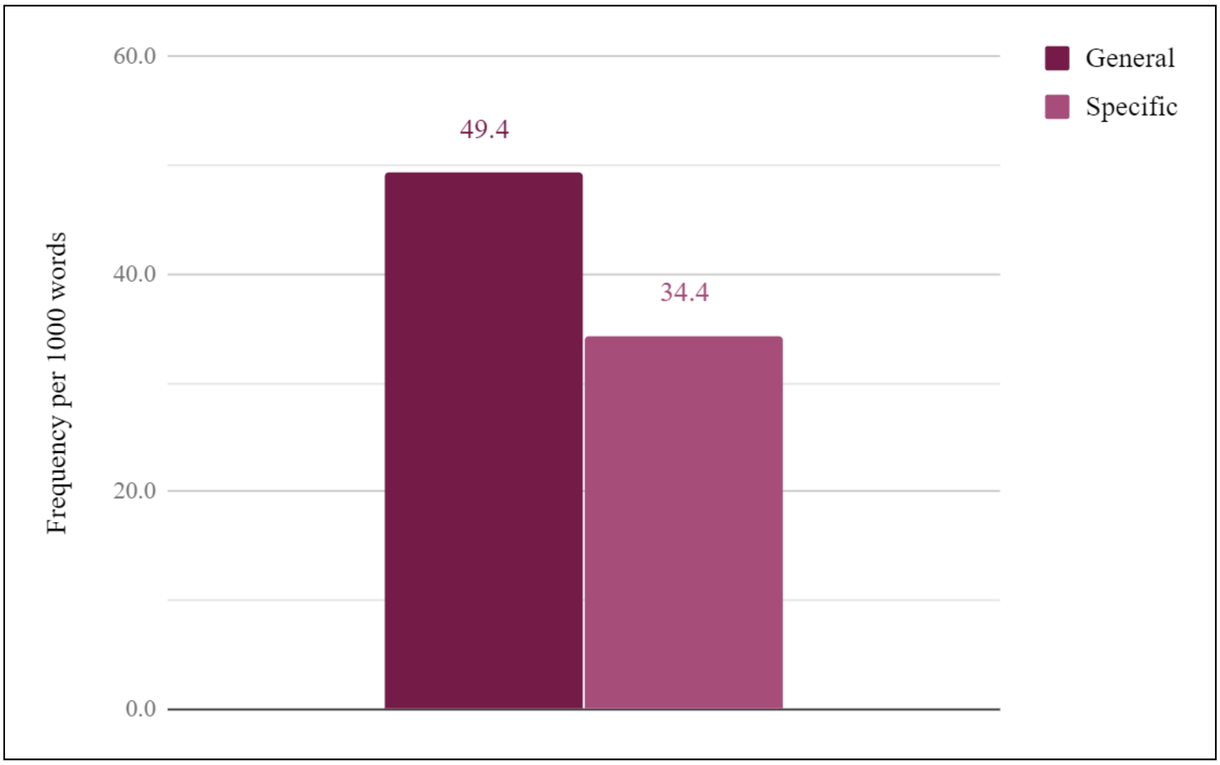
After the brief overview of the results obtained with the careful application of the research methodology, more details about simple NPs as produced by Brazilian learners are given in the next section.

4.2 Simple noun phrases

Simple NPs, as defined in Chapter 2, are those NPs which have a noun as headword by itself or accompanied by one or more determiners to its left. It was made clear in section 4.1 that this type of NP was not more frequent than complex NPs in the essays written by Brazilian learners. If we look at the proportion of simple NPs obtained when compared to the total of NPs analyzed in our research corpus, we find out that only 35% of NPs analyzed are simple.

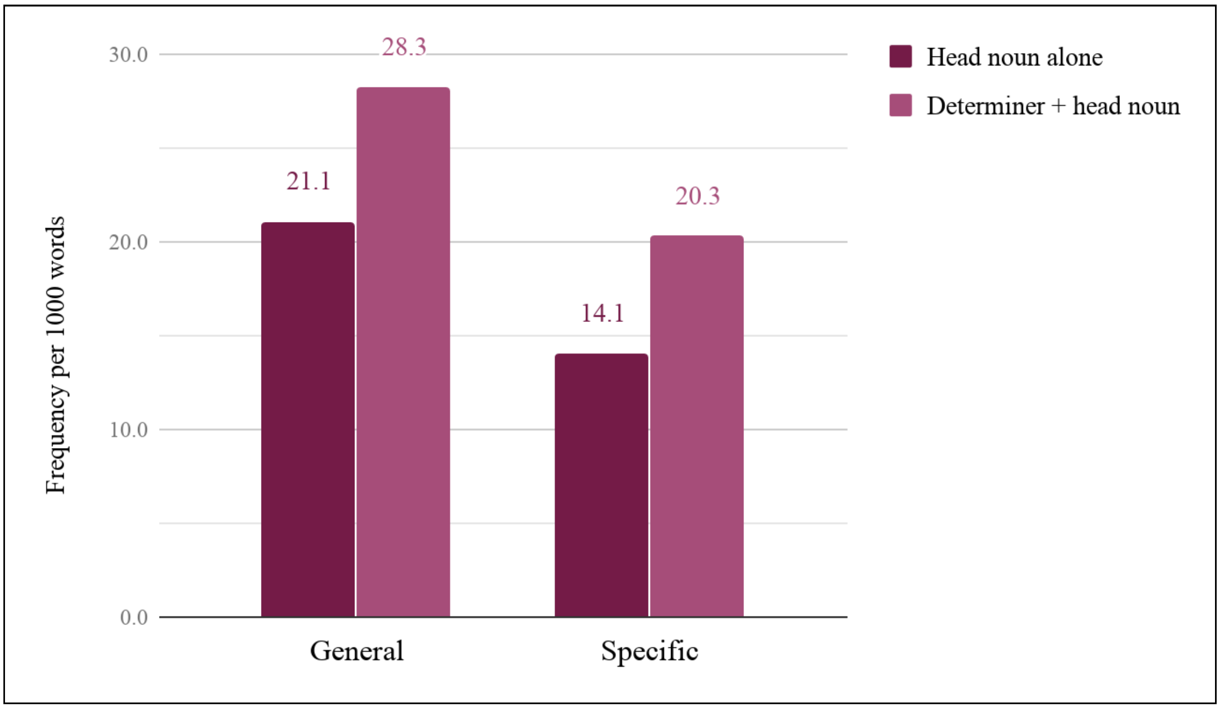
In the general topic subcorpus, simple NPs occur more frequently than they do in the specific topic subcorpus (see GRAPH 4.4 and TABLE 4.1). In terms of simple NP constituents, it was found that Brazilian learners produce NPs with determiners more often than NPs with head nouns by themselves (see GRAPH 4.5). Proportionately, 58.3% of simple NPs have at least one determiner while 41.7% have only a headword. This result could have been expected as nouns can serve as a means of referential specification about the text, which usually requires the use of determiners (BIBER *et al.*, 1999, p. 232), but that should be further analyzed in the following subsections.

Graph 4.4 – Simple NPs (per subcorpus)



Source: Designed by the author, 2019.

Graph 4.5 – Simple NPs: Head noun alone vs. Determiner + head noun (per subcorpus)

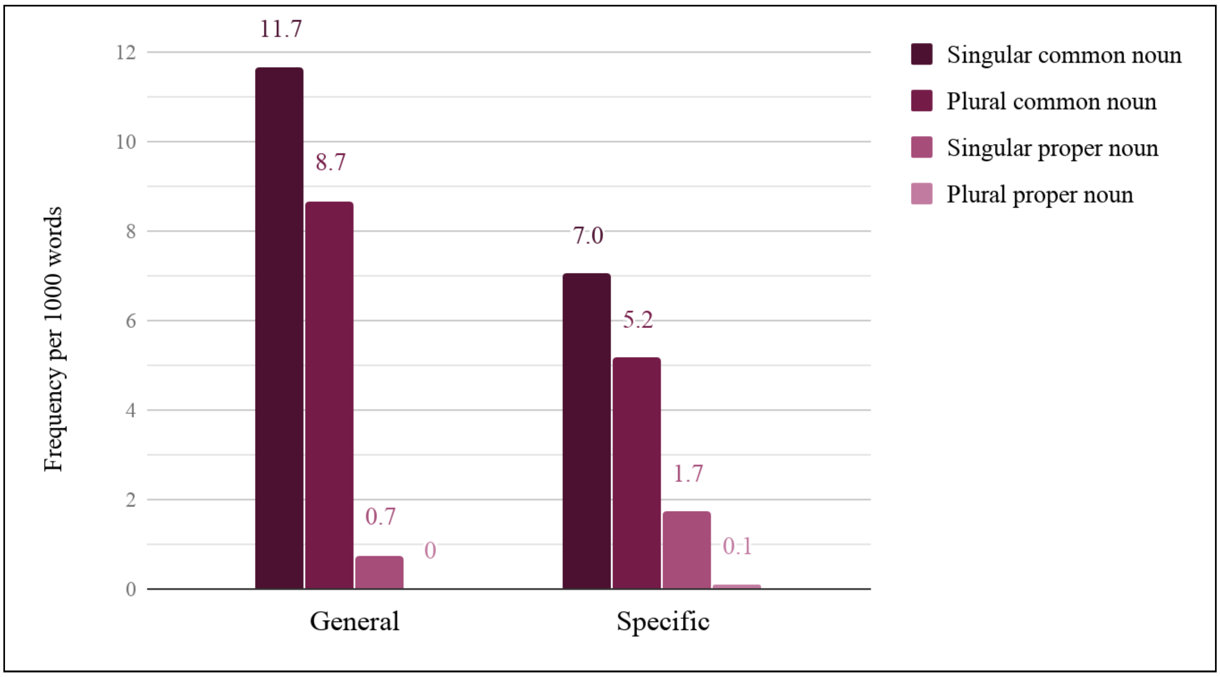


Source: Designed by the author, 2019.

4.2.1 Simple noun phrases without determiners

As shown in Graph 4.5, in both subcorpora, the frequency of NPs with a head noun by itself is a little lower than when it is contrasted with NPs that have both a determiner and a head noun. It seems necessary then to evaluate the types of head nouns that have been produced by learners in their essays, which could be: common or proper nouns, used in their singular or plural forms. To distinguish these categories, the Penn Treebank word tags NN, NNS, NNP, and NNPS served as guides. Graph 4.6 presents the frequency of use of these four categories of head nouns in the subcorpora of this study.

Graph 4.6 – Simple NPs: Types of head nouns alone (per subcorpus)



Source: Designed by the author, 2019.

As anticipated by Biber *et al.* (1999), common nouns are more frequently used and their singular forms as well in both subcorpora. According to Biber *et al.* (1999), singular forms of common nouns are more frequent than plural forms in English but, in written registers, there is also a quite common, even though small, use of plural forms. The use of plural nouns could be a reflection of the writers' preoccupation "with generalizations that are valid more widely" (BIBER *et al.*, 1999, p. 291). In example (1), taken from the first paragraph of an essay, the plural nouns *dentists* and *people* (the NPs illustrated are in between square brackets [ ] and head noun marked in **bold**) represent the author's concern with making a general statement about dentists and people who visit them, probably based on common knowledge or personal experience.

1) Because [**dentists**] do not have much time in their clinics to answer a lot of questions, many of them tends to say anything at all to explain why the procedure is needed and how they will get it done. That is why [**people**] do not like to go or do not trust dentists in general.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0646.0416 (Specific)

Some frequent singular common nouns used by learners were *talent* (24), *technology* (24), *life* (7), and *information* (7), in the general topic subcorpus, and *research* (6), *energy* (3), *population* (3), and *school* (3), in the specific topic subcorpus, where it should be mentioned that many simple NP heads were used a few times in the same text. Some regularly used plural common nouns were *people* (61), *children* (24), and *kids* (9), in the general topic subcorpus, and *people* (9), *experiments* (4), and *humans* (4), in the specific topic subcorpus.

As for singular proper nouns, *English* (4) was the most used one in the general topic subcorpus, probably because one of the topics assigned to students was about foreign language teaching and learning, and *Brazil* (15) was widely used in the specific topic subcorpus. No plural proper noun was produced in the general topic subcorpus and, in the specific topic subcorpus, *Brazilians* (3) was the most frequent one but by one person only. These results, though, cannot be generalized since it would be necessary to have a larger learner corpus and a wide variety of topics to do so.

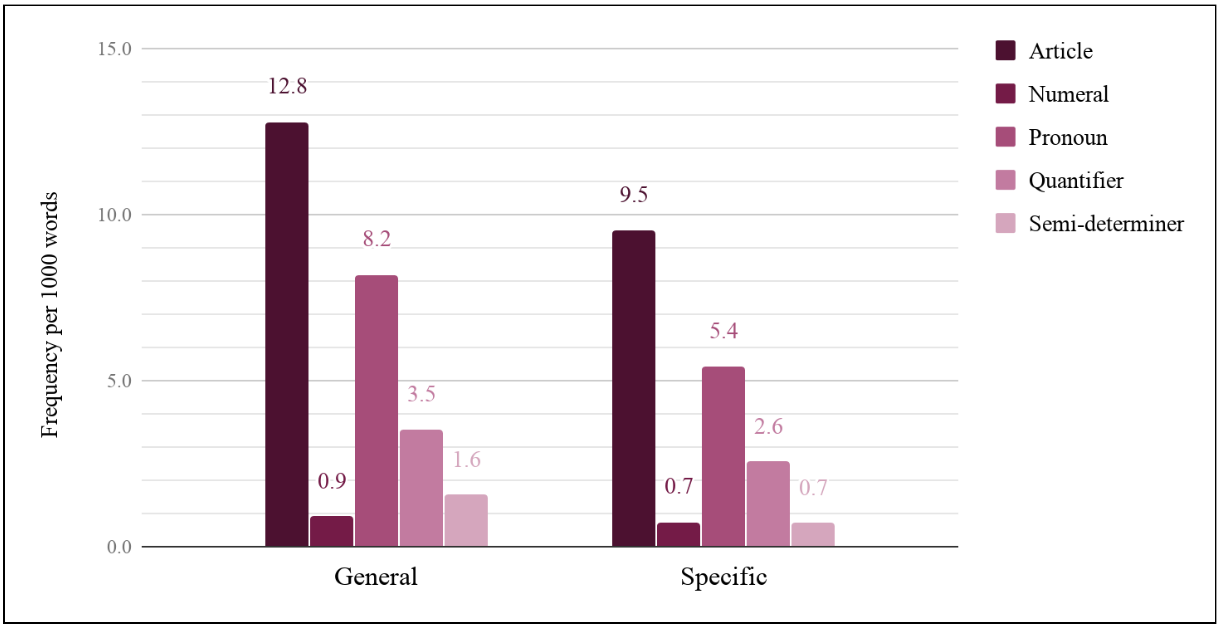
It is convenient to mention that a few corrections had to be done during the categorization just proposed, because of minor parser mistakes during the POS tagging. For instance, the word *nature* was tagged as a singular proper noun when it should have been tagged as a singular common noun. This was slightly recurrent in cases where common nouns were capitalized, e.g. the word above was written as *Nature*, and perhaps that led the parser to make the mistake. In such cases, the words wrongly tagged were added manually to the right category.

In the next subsection, the simple NPs containing determiners are examined.

4.2.2 Simple noun phrases with determiners

Simple NPs which have at least one determiner are, as it has been seen in Graph 4.5, the most frequent configuration of simple NP produced by Brazilian learners. It is already known that the determiners that could possibly occur together with a head noun are articles, numerals, pronouns, quantifiers, and semi-determiners (see more details about each determiner in Chapter 2). Graph 4.7 gives an overview of the distribution of simple NPs which have one of these five determiners as a constituent in each subcorpus and examples (2) and (3) show them as they are used by learners (the NPs illustrated are in between square brackets [ ], determiners are underlined and head nouns are marked in **bold**).

Graph 4.7 – Simple NPs: Types of determiners (per subcorpus)



Source: Designed by the author, 2019.

In both subcorpora, the determiners are produced in similar proportions. As expected, articles are more frequently used as determiners in simple NPs, as they are most common in the academic register (BIBER *et al.*, 1999, p. 267). More specifically, in both subcorpora, the definite article *the*, which usually specifies a referent as with *the environment* in (3), occurs more often than the indefinite article *a*/*an*, which often indicates an indefinite reference as with *a person* in (2), a difference that is also common in written English in general (BIBER *et* *al.*, 1999, p. 267) and a similar trend that is found in our research corpus. *The* occurs almost three times more often than *a*/*an* in the general topic subcorpus (171 definite article vs. 68 indefinite article) and almost five times more often in the specific topic subcorpus (255 definite article vs. 55 indefinite article).

2) I agree children should begin learning another language as soon as they start school, because this type of learning process is much more effective and easily done when [a **person**] is under 12 years old, according to several studies promoted by neuroscientists all over the world.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0606.0380 (General)

3) Moreover, this illumination is considered sustainable, since its various technical features make it noticeably less harmful to [the **environment**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0632.0404 (Specific)

The use of pronouns as determiners was also high in both subcorpora, even though demonstrative and possessive pronouns are not as common in the academic prose as they are in the other registers investigated by Biber *et al.* (1999, p. 270). Demonstratives, as generally anticipated by Biber *et al.* (1999), were more frequently produced by learners than possessives but the difference in use was very small in the general topic subcorpus (80 demonstrative vs. 73 possessive) and a little higher than the double in the specific topic subcorpus (124 demonstrative vs. 52 possessive). In the subcorpora, *this* (126) is more commonly used as a demonstrative determiner of simple NPs and *their* (95) is more frequently used as a possessive determiner. These determiners often have an anaphoric function, quite clear in examples (4) and (5), in which *their lives* and *this phenomenon* are referring back to *children* and *introversion*, respectively.

4) Therefore, many children has contact with English very early in [their **lives**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0599.0373 (General)

5) Therefore, it is clear that the introversion have to be more studied. Professionals in field of education and heath should be motivated to study and to know [this **phenomenon**] throughout their professional formation.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1001.0525 (Specific)

Numerals, quantifiers, and semi-determiners were consistently used in the essays written by the learners but not as frequently as the other determiners. Numerals, particularly cardinal numbers, as in examples (6) and (7), were more commonly used as determiners probably due to their precision and the informational purpose of academic texts (BIBER *et al.*, 1999). In the specific topic subcorpus, there was a higher variation in the cardinal numbers used and that could have happened because of the specificity of topics, which usually involved the presentation of students' own research, allowing them to be more precise about certain issues, as with *16 analyses* in (7).

6) First of all, the ability of learn is bigger for a child than for an adult and some studies say that if a kid learn [one **language**], this child is more able to learn another language easily.
   Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0612.0237 (General)

7) The litochemical study includes 46 litochemical analysis, trace and major elements, [16 **analyses**] are unreleased and the others were compiled from studies of Pedrosa-Soares (1984, 1995) and Grossi Sad e Motta (1991).
   Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0689.0329 (Specific)

Brazilian learners used a wide range of quantifiers when producing simple NPs. However, in both subcorpora, *some* (50) and *many* (36) occurred more frequently. These two quantifiers are also frequent in academic texts, since they can express generalizations (BIBER *et al.*, 1999, p. 277). From examples (8) and (9), it should be noticeable that the use of *some* and *many* helps writers make generalizations about *researchers* and *governments*, respectively.

8) However, [some **researchers**] defend that video games in reasonable doses have a quite powerful and positive effects on many different aspects of children behavior.
   Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509 (General)

9) From there, [many **governments**] started to research about genetically modified food, better know as transgenic foods, but they not yet proven whether they are safe to human health a long term.
   Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0692.0433 (Specific)

In both subcorpora analyzed, *other* (27) and *another* (20) were the most commonly used semi-determiners in simple NPs. This result could be expected, as *other* is one of the most frequent semi-determiners in academic texts (BIBER *et al.*, 1999, p. 282), used to specify, add an indefinite meaning to referents, or refer back to something previously mentioned in the text. In example (10), *another reason* is apparently referring to one or more reasons already given to the assertion that technology is not harmful to children. In example (11), *other substances* is possibly adding an indefinite reference to substances that could combine with Teflon.

10) [Another **reason**] is that, with the proper supervision, no technology is harmful to kids.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543 (General)

11) Thus, modification of PTFE by irradiation with electron beams and gamma rays has been heavily researched, since it results in the formation of reactive groups on the surface of the polymer, allowing the combination of Teflon with [other **substances**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0649.0418 (Specific)

It must be pointed out that the categorization of some quantifiers, such as *many*, *more*, and *much*, certain semi-determiners, such as *other*, and ordinal numerals had to be done carefully and manually. These specific determiners were not tagged as determiners but quite often as adjectives. That is not a problem because these words can be and are often analyzed as adjectives as they co-occur with articles, but it seems more reasonable to consider them as determiners as they do not change or add any new meaning to the head nouns in the NP (BIBER *et al.*, 1999).

In the case of simple NPs that have more than one determiner in their arrangement, there were a few occurrences in the research corpus. However, they were not as frequent as NPs with only one determiner. In the whole corpus, while 54.9% of simple NPs had one determiner, only 3.3% had two or more determiners. What was most remarkable in those few occurrences was the occurrence of the definite article in 66.7% and 69.8% of the NPs with two determiners in the general topic subcorpus and in the specific topic subcorpus, respectively. Example (12) illustrates the use of the definite article before the quantifier *most* and example (13) shows its use after the quantifier *all*.

12) It is considered that [<u>the</u> <u>most</u> **illnesses**] can be prevented with simple actions and a new style of life.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0953.0557 (General)

13) In this description, [<u>all</u> <u>the</u> **results**] can be explain without use coherence in optical regime, in other words, the explanation did not use the analogy with classical behavior because the Fock states has no classical analogue, they are purely quantum.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1000.0582 (Specific)

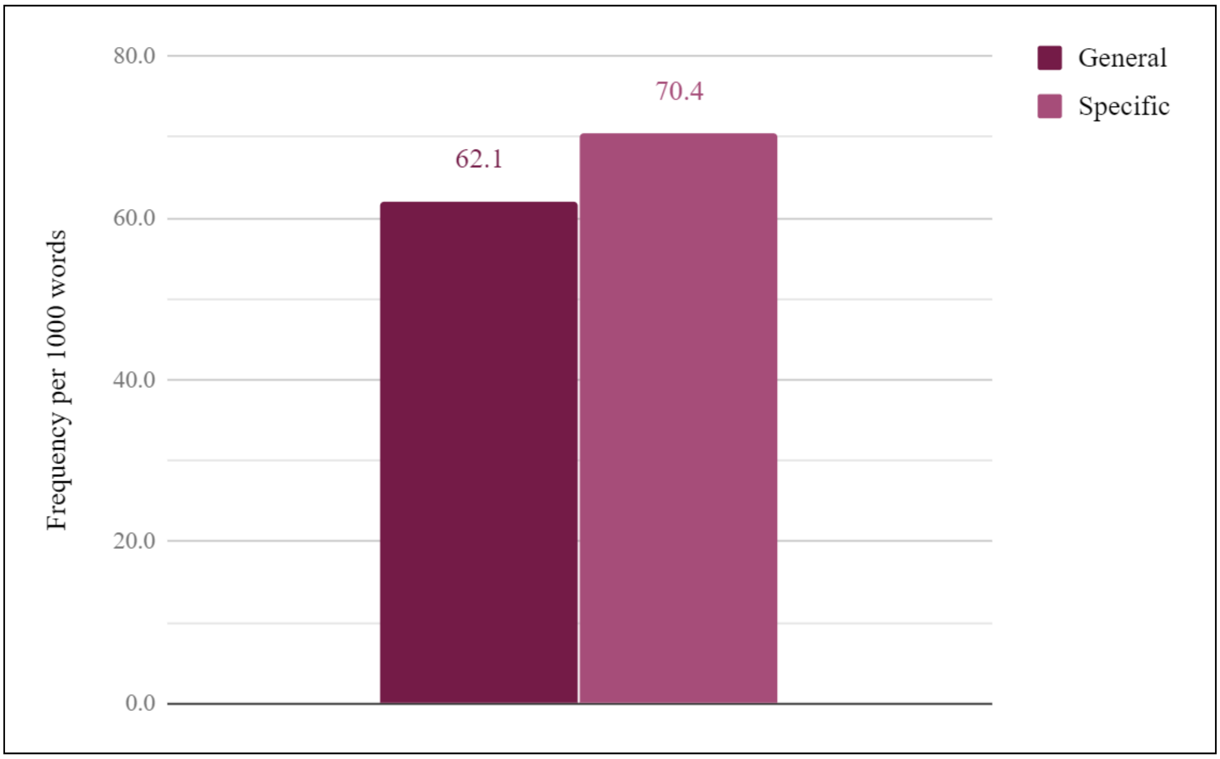
This is the description of simple NPs as they occurred in the research corpus. From here on, more details about the complex NPs produced by Brazilian learners will be given.

4.3 Complex noun phrases

Complex NPs, as defined in Chapter 2, are those NPs which have a noun as headword accompanied by one or more modifiers to its left or right. Section 4.1 showed that these NPs were more frequent than simple NPs in Brazilian writers' essays in our two CoIFA subcorpora. If we look at the proportion of complex NPs obtained when compared to the total of NPs analyzed in our research corpus, we find out that 59.3% of NPs analyzed are complex. This result is quite interesting because it is, generally speaking, closer to what is found in professional academic writing, where "almost 60% of all noun phrases have some modifier" (BIBER *et al.*, 1999, p. 578). It could also be implied that Brazilian learners at an upper intermediate proficiency level are capable of producing complex structures in their writing.

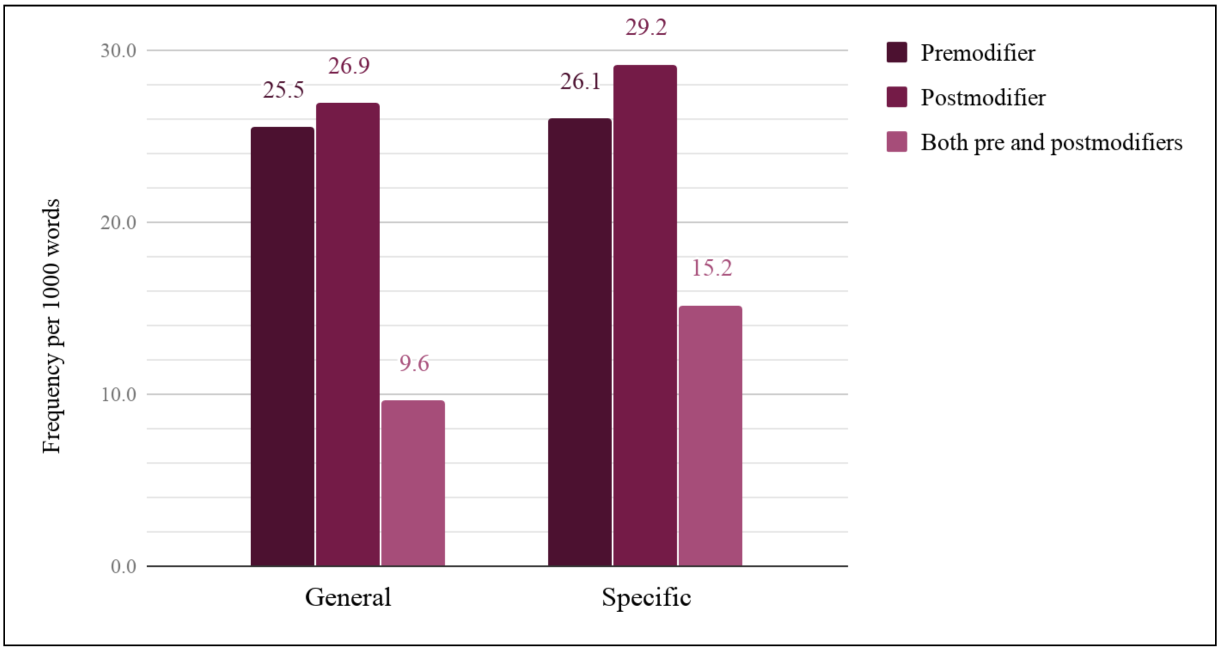
Graph 4.8 shows that complex NPs are slightly more common in the specific topic subcorpus than they are in the general topic subcorpus. These occurrences should be analyzed in detail in view of a learner corpus in the following subsections, but it can be observed from Graph 4.9 that NPs with postmodifiers are slightly more frequent than NPs with premodifiers, while both types of complex NP are more commonly used than NPs with both pre- and postmodifiers. In the whole research corpus, NPs with postmodifiers represent 42.1% of complex NPs, NPs with premodifiers 38.4%, and NPs with both pre- and postmodifiers 19.5%.

Graph 4.8 – Complex NPs (per subcorpus)



Source: Designed by the author, 2019.

Graph 4.9 – Complex NPs: Premodifiers vs. Postmodifiers vs. Both (per subcorpus)



Source: Designed by the author, 2019.

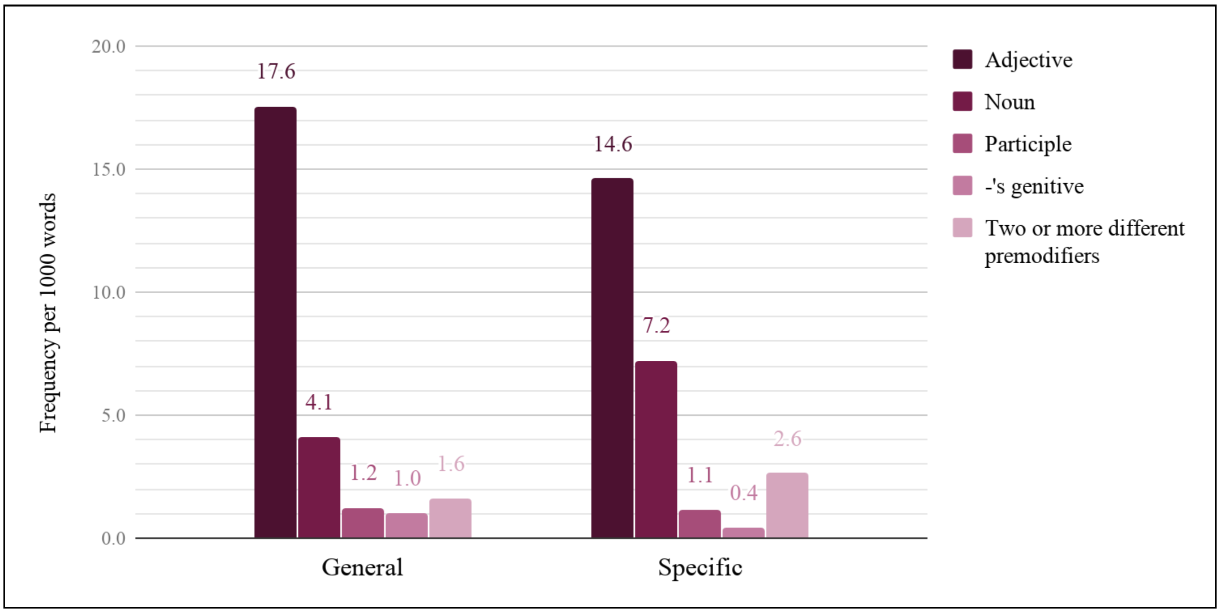
The higher percentage in the use of postmodifiers confirm our second hypothesis that learners would use postmodifiers more often than premodifiers. This finding is broadly opposite to the results reported by Nitsch (2017) regarding Brazilian learners' general topic essays, which had more premodifiers than postmodifiers (p. 99), and the results given by Biber *et al.* (1999) regarding professional academic writing in which NP premodifiers are more commonly used than postmodifiers.

Let us further analyze our results in the following subsections.

4.3.1 Complex noun phrases with premodifiers

There are three types of premodifiers commonly identified and analyzed in research on the grammatical complexity of English NPs, i.e. adjectives, nouns, and participle forms (cf. BIBER *et al.*, 1999; CARTER; McCARTHY, 2006). Complex NPs with two or more of the same premodifiers, e.g. an NP with two premodifying adjectives, were placed under the same category of NPs which have only one of the premodifiers. For this study, we also opt for analyzing the use of the *'s* genitive and cases where two or more premodifiers that are from different word classes occur together with a head noun.

Graph 4.10 – Complex NPs: Types of premodifiers (per subcorpus)



Source: Designed by the author, 2019.

Graph 4.10 presents the frequencies of NPs containing each type of premodifier categorized in this study and examples (14) to (25) are given to represent each category (the NPs illustrated are in between square brackets [ ], premodifiers are underlined and head nouns are marked in **bold**). It is necessary to mention that complex NPs with or without determiners were not categorized separately.

In both subcorpora, there is a higher use of adjectives in premodifying position, confirming our third hypothesis that learners use more adjectives as NP premodifiers. These NPs which contain one or more premodifying adjectives correspond to 60.6% of NPs with premodifiers. Such finding was expected as adjectives in premodifying position are more commonly found in written registers (BIBER *et al.*, 1999, p. 506) as well as in journal registers in general (GRAY, 2015), in applied linguistics abstracts (ANSARIFAR *et al.*, 2018), in essays written by EAP learners (PARKINSON; MUSGRAVE, 2014), and in essays written by Brazilian learners (NITSCH, 2017). These NPs can have as its constituents a head noun preceded by an adjective alone as in *multicultural families* in example (14), by a determiner and an adjective in *a future journalist* in (15), or by two or more adjectives separated by a coordinating conjunction in *a new and efficient model* in (16). Other configurations are possible but these were the ones most commonly seen in the learner corpus under analysis.

14) It already happens to [multicultural **families**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0609.0383 (General)

15) However, [a future **journalist**] will discuss, also, theories about Communication field, the relation between communication and popular culture and ethical implications about this work.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0696.0436 (Specific)

16) Specialists must discuss exhaustively and propose a progressivily transition to [a new and efficient **model**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0995.0577 (Specific)

Nouns premodifying another noun are also frequent in our learner corpus. 23.6% of the NPs with premodifiers contain one or more nouns in this position. This finding, in which premodifying nouns are not as frequent as attributive adjectives, is quite common in L2 learner writing (e.g. ANSARIFAR *et al.*, 2018; NITSCH, 2017). The contrary is, in fact, a characteristic of expert writing as Biber and Gray (2016) and Staples *et al.* (2016) reveal. Likewise, these complex NPs occur more frequently in our specific topic subcorpus. That might happen because complex NPs with one or more premodifying nouns compress meanings particular to academic discourse communities, creating (more) technical and specialized expressions such as the NP *blood stem cells* in example (18) as opposed to *brain stimulation* in example (17). It is easier to find NPs with premodifying nouns with more specialized meanings, which demand specialized knowledge, in the specific topic subcorpus than in the general topic subcorpus.

17) Researches has established that skills are born of [brain **stimulation**], indicating discoveries which support hard work defenders.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0724.0456 (General)

18) As an illustration, we can cite bone marrow transplantation that a patient receiving [blood stem **cells**] from a donor.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0698.0437 (Specific)

NPs with premodifying participles, as illustrated in *acquired skills* in (19) and in *a sounding board* in (20), are not as frequent as the other categories analyzed. They represent only 4.5% of the NPs with premodifiers. This finding is similar to other studies which consider participle forms separately from adjectives, such as Parkinson and Musgrave (2014), in which participial premodifiers have a lower frequency in their corpus. Differently from our results, in Nitsch (2017), premodifying participles were more commonly produced by Brazilian learners than premodifying nouns. Furthermore, the increase in the use of noun-participle forms, e.g. *corpus-based*, demonstrated by Biber and Gray (2016) could not be seen in our learner corpus since no similar structure was produced by learners.

19) So as to [acquired **skills**] it is necessary so much effort.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0737.0468 (General)

20) Today we have a hundreds of variations to instruments similar to it and almost certainly they came to the same place lost in the history, when a man discover that a string can produce a sound and a shell, or [a sounding **board**] make by wood, can propagate.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0647.0417 (Specific)

As anticipated for published academic texts, in the research corpus, NPs containing the *’s* genitive, exemplified in (21) and (22), do not occur as frequently as NPs with other premodifiers. In our learners’ essays, only 2.5% of NPs with premodifiers have the *’s* genitive. That could be part of a major change in English academic writing, in which there has been a decrease in the use of *’s* genitive and, as an alternative, an increase in the use of *of*-phrases or premodifying nouns (BIBER; GRAY, 2016, p. 171-172).

21) Health both from mind and body in the old age are important aspects to consider, Merzenich, a neuroplastic scientist said that although you can reach life expectancy around late eighties when you are eighty-five, there is a fourty-seven percent chance that you will have [Alzheimar’s **disease**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0954.0558 (General)

22) [This quota’s **politic**] is an attempt to give underprivileged Brazilians better chances of getting free higher education and thus access to better jobs and consequently to reduce social inequality in Brazil.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0714.0331 (Specific)

In both subcorpora, complex NPs with two or more different premodifiers are more commonly produced by Brazilian learners than the last two configurations just mentioned (participle and *’s* genitive), representing 8.8% of NPs with premodifiers. Sometimes, these premodifiers have a coordinating conjunction between them as in *the hybrid and electrical vehicles* in (25), but other times they come one after the other, as can be seen in *a social network boom* in (23) or *the longest confirmed human lifespan* in (24). The last configuration, for instance, in which there is an adjective followed by a participle and a noun premodifying the head noun, shows the compression of information in an NP.

23) Beyond the release of gadgets, we have faced [a social network **boom**].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0735.0466 (General)

24) Also, this community is living longer, [the longest confirmed human **lifespan**] died at the age of 123.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0953.0557 (General)

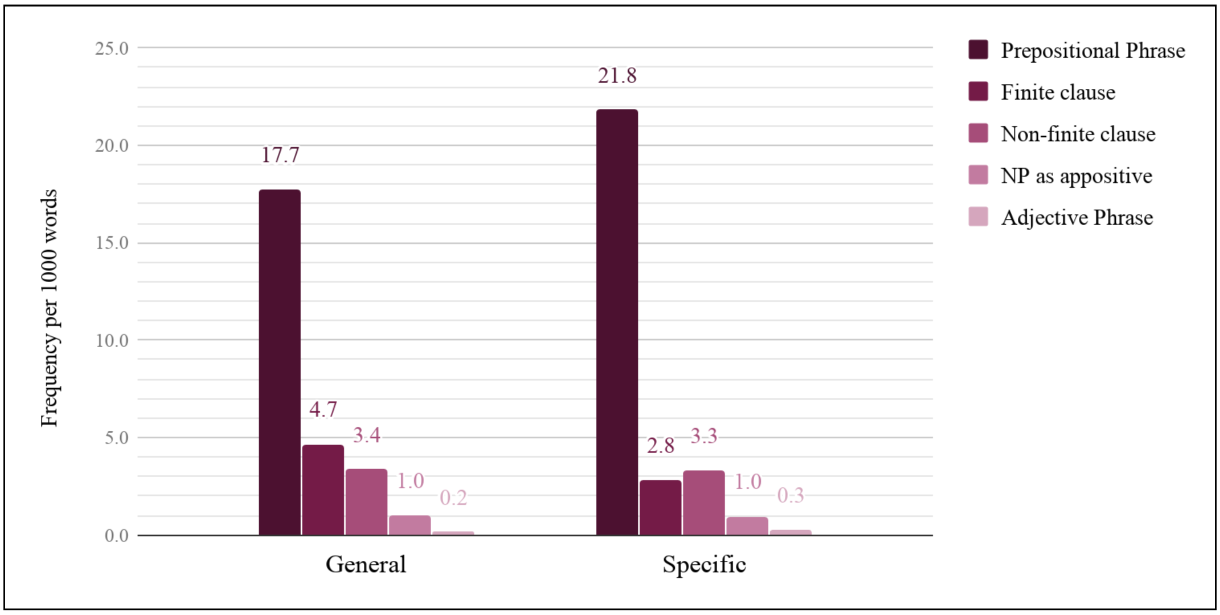
25) Therefore, due to all the benefits that this new idea can cause, [the hybrid and electrical **vehicles**] are becoming a tendency in the world car market and suggest the beginning of a transition time forward a full electric age in automotive field.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0997.0579 (Specific)

Having presented the complex NPs with premodifiers, the next subsection details the occurrences of NPs with postmodifiers.

4.3.2 Complex noun phrases with postmodifiers

As it has been observed in section 4.3, NPs with postmodifiers were the most frequent configuration of complex NPs found in the research corpus. The types of postmodifiers usually identified in studies of NP complexity are four: PPs, finite clauses, non-finite clauses, and appositive NPs. Based on our perception and findings after parsing the research corpus, we decided to include cases where AdjPs occur as postmodifiers. Even though these are not frequent, they are interesting as they may represent a common learner mistake.

Graph 4.11 – Complex NPs: Types of postmodifiers (per subcorpus)



Source: Designed by the author, 2019.

In order to identify all these postmodifiers, we looked at the first postmodifier occurring right after a head noun. That means, we did not analyze NPs with multiple postmodifiers differently. Graph 4.11 and the examples (26) to (44) should give an idea of the postmodified NPs produced by Brazilian learners (the NPs illustrated are in between square brackets [ ], postmodifiers are <u>underlined</u> and head nouns are marked in **bold**). Overall, 71.6% of NPs were postmodified by PPs, 12.3% by finite clauses, 11.8% by non-finite clauses, 3.4% by appositive NPs, and 0.8% by AdjPs.

Data analysis shows that the frequency of PPs is much higher than that of other postmodifiers, confirming our fourth hypothesis that learners use more PPs as NP postmodifiers. This is an expected finding as postmodifying PPs are common in written expository registers, particularly in the case of NPs containing PPs headed by the preposition *of* (BIBER *et al.*, 1999, p. 635). In general, postmodifying PPs are also most frequent in journal registers (GRAY, 2015) and in applied linguistics abstracts (ANSARIFAR *et al.*, 2018). Tables 4.2 and 4.3 show in detail the frequency of the prepositions heading postmodifying PPs.

Among the PPs used in complex NPs with postmodifiers produced by learners *of*-phrases were the most frequent ones, used 57.4% of the times in the general topic subcorpus and 75.9% in the specific topic subcorpus. That is similar in Nitsch's (2017) study, in which Brazilian learners produced more complex NPs postmodified with *of*-phrases than with other prepositions (p. 112). Examples (26), *a reality of a different culture*, and (27), *the reproduction of Amazonian turtles*, illustrate the use of the preposition *of* with a meaning comparable to that of genitives.

26) Thus, through develop knowledge in language children can get in though with [a **reality** <u>of a different culture</u>], consequently, opening the way a kid can think.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0612.0237 (General)

27) These environments are important for [the **reproduction** <u>of Amazonian turtles</u>].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0988.0570 (Specific)

In-phrases, similarly to what is found by Biber *et al.* (1999) concerning professional academic writing, were the second type of PPs most frequently used as postmodifiers by learners in both subcorpora. They occurred 15.7% of the times in the general topic subcorpus and 9.2% in the specific topic subcorpus. In example (28), *ageism in an interpersonal level*, the preposition *in* is used to identify the phenomenon of ageism in relation to one specific level among other levels affected by it. In example (29), *techniques in each powerplant that could make one more feasible than other*, the postmodifying PP used compresses the information that techniques will be applied to power plants, marking that these are the semantic patient of the process (BIBER; GRAY, 2016, p. 196). It is worth mentioning that there is an embedded *that* clause in (29) adding another layer of modification to the head noun *techniques*.

Table 4.2 – Complex NPs: PPs as postmodifiers (general topic subcorpus)

Preposition	raw	per 1000 words	%
of	190	10.2	57.4
in	52	2.8	15.7
for	22	1.2	6.6
with	21	1.1	6.3
between	8	0.4	2.4
to	8	0.4	2.4
about	7	0.4	2.1
on	5	0.3	1.5
as	4	0.2	1.2
from	4	0.2	1.2
at	2	0.1	0.6
*other prepositions**	8	0.4	2.4
TOTAL	331		100

Source: Designed by the author, 2019.

* Each preposition (*across after*, *around*, *like*, *over*, *through*, *against*, *without*) occurred only once in the general topic subcorpus.

28) Those assumptions would be considered ageism on a personal level but, when those simple believes evolve to actions or language, for example, to speak loudly because one thinks elderly have bad hearing, it is considered [**ageism** in an interpersonal level].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0950.0554 (General)

29) One and another way of production has its points, and some aspects can be minimized with [**techniques** in each powerplant that could make one more feasible than other].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1008.0589 (Specific)

Other prepositions commonly used by learners were *for*, *with*, *between*, *to*, *about*, *on*, and a few others. Example (30), *time between being a kid and learning*, shows the use of the preposition *between*, perhaps chosen because of the verb *dividing*, and example (31), *a partnership with the Colombia construction companies*, represents the use of the preposition *with*, perhaps frequent with the head noun *partnership*.

Table 4.3 – Complex NPs: PPs as postmodifiers (specific topic subcorpus)

Preposition	raw	per 1000 words	%
of	538	16.5	75.9
in	65	2.0	9.2
for	20	0.6	2.8
with	17	0.5	2.4
to	16	0.5	2.3
about	12	0.4	1.7
on	11	0.3	1.6
from	8	0.2	1.1
between	8	0.2	1.1
as	3	0.1	0.4
like	3	0.1	0.4
at	2	0.1	0.3
*other prepositions**	6	0.2	0.8
TOTAL	709		100

Source: Designed by the author, 2019.

* Each preposition (*among*, *around*, *over*, *than*, *via*, *without*) occurred only once in the specific topic subcorpus.

30) This learning process has to be something natural, according to what both parents and kids want, dividing [**time** between being a kid and learning].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0622.0395 (General)

31) They made a Real Estate Fair to Colombians at Spain, in [a **partnership** <u>with the Colombia construction companies</u>].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1014.0593 (Specific)

Considering the use of clauses as NP postmodifiers, there is a slight difference between the subcorpora analyzed. On the one hand, in the general topic subcorpus, finite clauses are more frequent than non-finite clauses, being 58% of the postmodifying clauses finite and 42% non-finite. On the other hand, in the specific topic subcorpus, non-finite clauses occur more frequently than finite clauses, being 54% of the clauses in postmodification non-finite and 46% finite.

Once more, that contrast in use of postmodifying clauses in the subcorpora could be due to the fact that non-finite clauses are more compressed than finite clauses and their use might reflect these tendency towards compression in academic prose, specially in texts with more specialized topics (BIBER; GRAY, 2016). It can be assumed that the specific topic subcorpus has more texts with (more) specialized themes and the Brazilian learners could perhaps have produced more compressed structures such as the postmodifying non-finite clause because of that. The lower production of postmodifying finite clauses is also observed by Staples *et al.* (2016) in higher academic levels. Nevertheless, there is also a disciplinary variation demonstrated by Gray (2015), in which non-finite clauses are more frequent in the quantitative studies while finite clauses are more frequent in the qualitative and humanities studies (p. 126). However, that would require another investigation of the complex NPs produced by our learners according to their major or research field.

Table 4.4 details the use of postmodifying clauses in complex NPs in each subcorpus. It can be noticed that the different types of postmodifying finite clauses are used with similar proportions in both subcorpora, which means that the *that* clause is more common than the *wh-* clause, which is in turn, more frequent than the zero relativizer clause.

Table 4.4 – Complex NPs: Finite and non-finite clauses as postmodifiers (per subcorpus)

	General topic subcorpus			Specific topic subcorpus		
	raw	per 1000 words	%	raw	per 1000 words	%
Finite clauses						
that clause	52	2.8	34.7	58	1.8	29.0
wh- clause	26	1.4	17.3	31	1.0	15.5
zero relativizer	9	0.5	6.0	3	0.1	1.5
TOTAL	87		58.0	92		46.0
Non-finite clauses						
participle clause	31	1.7	20.7	69	2.1	34.5
to clause	32	1.7	21.3	39	1.2	19.5
TOTAL	63		42.0	108		54.0

Source: Designed by the author, 2019.

The relative pronoun *that* can be used in several contexts, especially when conveying restrictive meanings. Such interpretation is clear in *some problems that change a little the reality* in (32), in which the *that* clause identifies the existence of specific problems, and in *an anisotropy that depends of the crystalline direction* in (33), in which the *that* clause establishes a particular type of the property *anisotropy*. This type of postmodifier is also found to be frequent in Nitsch (2017), where *that* clauses are the second most common postmodifying clause in Brazilian learners' essays. It should be noted that some complex NPs containing a human head noun, e.g. *people*, were modified by *that* clauses instead of *wh-* clauses. That occurred at least 12 times in the research corpus, being 11 of those in the general topic subcorpus.

32) It can be true, but most of people who survive to old age faces [some **problems** that change a little the reality].

    Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0952.0556 (General)

33) The results obtained through the magneto optometry indicates that for a 10 monolayers thickness, the magnetization is pointing in plane, in contrast with the results obtained in the Pd, but the hysteresis curves indicates that we have [an **anisotropy** that depends of the crystalline direction].

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1010.0591 (Specific)

The *wh-* clause is somewhat two times less frequent than the *that* clause in both subcorpora. As for particular *wh-* pronouns used by Brazilian learners, *who* (26) and *which* (19) are more common than the pronouns *when* (3), *where* (3), and *why* (2), for instance. The use of *who*, restricted to human reference as in the example (34), is the third most frequent relativizer in professional academic writing (BIBER *et al.*, 1999) but the first most common in Brazilians' general topic essays (NITSCH, 2017). The relativizer *which*, as shown in example (35), is the most frequently used in professional academic texts (BIBER *et al.*, 1999) and often represents a stylistic choice that might not be known by learners, justifying its lower frequency in this study.

34) I agree with many authors when they say: in most of careers, [**people** who hardly work] is more important than people has talent.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0737.0468 (General)

35) One of the most present objects in this domain are [**books**, which are, even in a connected world, a powerful tool to consolidate knowledge].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1011.0592 (Specific)

NPs postmodified by zero relativizer clauses are the least frequent ones produced by our Brazilian learners. Still, they are used as structures which explicitly add information to head nouns even without having a relative pronoun to join them. Example (36) shows that the use of the postmodifying clause *we produce and consume energy* without a relativizer is permitted because of the head noun *way*, which has an adverbial gap, and the lack in English of a relative adverb to mark manner (BIBER *et al.*, 1999, p. 621).

36) This new approach of the electrical system can bring a revolution to [the **way** we produce and consume energy].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0942.0547 (Specific)

From Table 4.4, it is seen that the different types of postmodifying non-finite clauses are produced with similar proportions in both subcorpora, meaning participle clauses occur more frequently than *to* clauses. There are though some differences that should be reported.

The use of NPs postmodified by participle clauses differed in each subcorpus. In the general topic subcorpus, *-ed* clauses (21) were two times more common than *-ing* clauses (10). In the specific topic subcorpus, the contrary was found, *-ing* clauses (55) were almost three times more frequent than *-ed* clauses (14). This last result was similar in the academic register (BIBER *et al.*, 1999) and in other Brazilian learners' essays (NITSCH, 2017). As these clauses are considered more compressed and economical structures, examples (37) and (38) are presented to give an idea of that characteristic.

37) Although there are studies on the case, [the **tests** performed] are not conclusive because they do not take into account the various factors that may alter their test result, that is, they do not present only one variant between the control group and the test group.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0944.0549 (General)

38) In addition, with regard to the thoracoabdominal motion, the response of the abdominal rib cage to different postures it was not evaluated and this is considered an important distinguishing factor of this instrument, in order that [the **forces** acting on the upper rib cage (adjacent to lungs)] are quite different from those acting on its bottom (adjacent to diaphragm).
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0707.0443 (Specific)

As the least common non-finite postmodifier in English, *to* clauses were also not commonly produced by learners. Likewise, the essays analyzed by Nitsch (2017) had just a few occurrences of that postmodifier. In example (39), *to reduce loneliness* is used to express the purpose of the head noun *strategies*. In (40), there is an example of complement clause, *to synthesize a heterodox approach to Urban Economics* controlled by the head noun *attempt*.

39) [**Strategies** to reduce loneliness] can encourage people to develop more social skills and reduce loneliness, like programs and support groups for those who face social media anxiety and loneliness.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0742.0472 (General)

40) In this panorama, this essay is [an **attempt** to synthesize a heterodox approach to Urban Economics].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0730.0462 (Specific)

As expected, NPs in apposition were not frequently used by learners. Studies on learner writing usually do not find high frequencies of this type of postmodifier (e.g. PARKINSON; MUSGRAVE, 2014) even though this is a frequent and distinctive postmodifying structure of academic texts' grammatical complexity (BIBER; GRAY, 2016). In both of our subcorpora, a little more than 3% of the NPs with postmodifiers were appositive NPs. When used, they mostly characterized a person, such as *Merzenich* in example (41), explained a technical term, as with *leishmaniose* in (42), or itemized group members. That is a result of the varied use of NPs as appositive elements in the academic prose (BIBER *et al.*, 1999).

41) Health both from mind and body in the old age are important aspects to consider, [**Merzenich**, a neuroplastic scientist] said that although you can reach life expectancy around late eighties when you are eighty-five, there is a fourty-seven percent chance that you will have Alzheimar's disease.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0954.0558 (General)

42) For example, there are researches try to find a vaccine for [**leishmaniose**, a tropical parasitic disease that affects more dogs than humans,] it is transmitted for humans by the bite of the mosquito that first bites the infected dog.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0711.0425 (Specific)

AdjPs as postmodifiers were also not common in the research corpus. They represent less than 1% of complex NPs with postmodifiers. However, it should be remarked that this structure of postposed adjectives exists in English, particularly for certain adjectives such as *available* and for heavy/long AdjPs, and learners are capable of using them. That can be seen in example (43). Still, there are a few learners who make mistakes and use the adjective as a postmodifier when it should be used as a premodifier, as in *another issue considerable* in example (44), probably because of the word order of nouns and adjectives in Portuguese (their L1), which tends to be the opposite of what we have in English.

43) The effluent of UASB reactor is anaerobic, which means that there is [no **oxygen** available to sulfide oxidation].

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0634.0406 (Specific)

44) [Another **issue** considerable] is that the use of technology as the main platform to work leads to reduce the capacity of the workers to pass trough uncomfortable situations and reach objectives by them selves.

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0721.0453 (General)

In the next subsection, the complex NPs that contain premodifiers as well as postmodifiers are analyzed.

4.3.3 Complex noun phrases with both pre- and postmodifiers

From the complex NPs found in our research corpus, 19.5% of them were NPs that contained pre- and postmodifying elements. As could be seen in Graph 4.9 in section 4.3, this configuration of complex NP is the least common one found in the learner corpus, being more commonly used in the specific topic subcorpus than in the general topic subcorpus.

Table 4.5 – Complex NPs: Possible configurations of NPs with pre- and postmodifiers (general topic subcorpus)

Sequences	raw	per 1000 words	%
1 pre + N + 1 post	125	6.7	69.4
1 pre + N + 2 post	24	1.3	13.3
1 pre + N + 3 post	10	0.5	5.5
1 pre + N + 4 post	3	0.2	1.7
1 pre + N + 5 post	3	0.2	1.7
1 pre + N + 6 post	1	0.1	0.6
2 pre + N + 1 post	9	0.5	5.0
2 pre + N + 2 post	3	0.2	1.7
2 pre + N + 3 post	2	0.1	1.1
TOTAL	180		100

Source: Designed by the author, 2019.

When analyzing these NPs, we decided to write down in the spreadsheets where they were stored, next to each NP, the number and type of premodifiers and postmodifiers used. After that procedure, it was discovered that the general topic subcorpus had 9 possible combinations of pre- and postmodifiers in complex NPs and the specific topic subcorpus had 18 possibilities. Tables 4.5 and 4.6 present these combinations of head nouns (represented by the letter N) and pre- and postmodifiers in each subcorpus and their frequencies of use. Excerpts (45) to (48) should exemplify some of these structures (the NPs illustrated are in between square brackets [ ], pre- and postmodifiers are underlined and head nouns are marked in **bold**).

It is evident that a wider range of combinations of pre- and postmodifiers in NPs was produced in the specific topic subcorpus, containing sequences with more premodifiers as well as with more postmodifiers than the general topic subcorpus. That could have been Brazilian writers’ attempts to compress more information in NPs due to their (more) specialized topics. Nonetheless, it is interesting that learners in both subcorpora produced complex NPs with numerous layers of phrasal embedding, with six, seven, and even ten postmodifiers. The embedding of multiple postmodifiers, in which two or more postmodifiers are used, is actually much more recurrent than the co-occurrence of several premodifiers. Multiple postmodifiers represent 25% of NPs with both pre- and postmodifiers in the general topic subcorpus and 39.9% in the specific topic subcorpus (see TABLE 4.7 for more details). That demonstrates learners’ capability and perhaps necessity of embedding information in NPs, just like professional writers (BIBER; GRAY, 2016). In example (45), which has four premodifiers and four different postmodifiers, it is fascinating to identify each modifier adding meaning to the head noun *ship*.

45) An example widely used in engineering schools is [the Liberty class naval cargo **ship**, built in the U.S. during World War II, when a large number of them sank during crossings in the Atlantic Ocean].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0651.0420 (Specific)

Table 4.6 – Complex NPs: Possible configurations of NPs with pre- and postmodifiers (specific topic subcorpus)

Sequences	raw	per 1000 words	%
1 pre + N + 1 post	265	8.2	53.6
1 pre + N + 2 post	92	2.8	18.6
1 pre + N + 3 post	43	1.3	8.7
1 pre + N + 4 post	24	0.7	4.9
1 pre + N + 5 post	7	0.2	1.4
1 pre + N + 6 post	4	0.1	0.8
1 pre + N + 7 post	2	0.1	0.4
1 pre + N + 8 post	1	0.03	0.2
1 pre + N + 10 post	1	0.03	0.2
2 pre + N + 1 post	27	0.8	5.5
2 pre + N + 2 post	15	0.5	3.0
2 pre + N + 3 post	1	0.03	0.2
2 pre + N + 4 post	3	0.1	0.6
2 pre + N + 5 post	1	0.03	0.2
2 pre + N + 7 post	1	0.03	0.2
3 pre + N + 1 post	5	0.2	1.0
3 pre + N + 2 post	1	0.03	0.2
4 pre + N + 4 post	1	0.03	0.2
TOTAL	494		100

Source: Designed by the author, 2019.

Also in both subcorpora, the most frequently used configuration is the one which contains one premodifier and one postmodifier, corresponding to 69.4% of the NPs in the general topic subcorpus and 53.6% in the specific topic subcorpus. These complex NPs are illustrated in example (46), which has one premodifying adjective and one postmodifying PP, and example (47), which is longer than (46) but which also has one adjective as premodifier and one *that* clause as postmodifier.

46) The psychologist realized, later, that the brain of mices in the first situation developed more and made [a higher **number** of synapses].

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0724.0456 (General)

47) This is [a primordial **subject** that should have be treated with more importance than of the last decades].

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0690.0431 (Specific)

NPs with one premodifier and two postmodifiers are also very common in both subcorpora, representing 13.3% and 18.6%, respectively. Excerpt (48) illustrates the use of one premodifying adjective and two PPs, the *of*-phrase followed by the *on*-phrase.

48) [The main **impact** of video games on the brain] is related to the brain network that control attention.

Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509 (General)

Table 4.7 – Complex NPs: Frequencies of pre- and postmodifiers when both are used (per subcorpus)

	General topic subcorpus			**Specific topic subcorpus**		
	raw	per 1000 words	%	raw	per 1000 words	%
Premodifiers						
Adjective	135	7.2	75.0	349	10.7	70.6
Noun	15	0.8	8.3	55	1.7	11.1
Participle	11	0.6	6.1	31	1.0	6.3
's genitive	5	0.3	2.8	4	0.1	0.8
Two or more premodifiers	14	0.7	7.8	55	1.7	11.1
TOTAL	180		100	494		100
Postmodifiers						
Prepositional phrase	93	5.0	51.7	206	6.3	41.7
Finite clause	19	1.0	10.5	23	0.7	4.7
Non-finite clause	18	1.0	10.0	42	1.3	8.5
NP as appositive	3	0.2	1.7	20	0.6	4.0
Adjective	2	0.1	1.1	6	0.2	1.2
Two or more postmodifiers	45	2.4	25.0	197	6.1	39.9
TOTAL	180		100	494		100

Source: Designed by the author, 2019.

In terms of the types of premodifiers and types of postmodifiers used in these complex NPs, in both subcorpora, the most common premodifier found was the adjective and the preferred postmodifier was the PP (see TABLE 4.7). In the general topic subcorpus, in NPs that had one premodifier and one postmodifier, 75% of them had an adjective as premodifier and 51.7% had a PP as postmodifier. Similarly, in the specific topic subcorpus, 70.6% of the same type of NPs had a premodifying adjective and 41.7% had a postmodifying PP. Interestingly, these results are equivalent to the ones obtained for complex NPs with premodifiers (see GRAPH 4.10) and NPs with postmodifiers (see GRAPH 4.11), suggesting that learners follow the same preferences when producing the different types of complex NPs. They also confirm once more our two hypotheses concerning the common use of adjectives and PPs as NP modifiers.

The use of two or more postmodifiers in both subcorpora was also considerable, 25% in the general topic subcorpus and 39.9% in the specific topic subcorpus. In those cases, we analyzed the postmodifier that came right after the head noun in the NP, independently of the number of postmodifiers in the NPs. We discovered that, in both subcorpora, more than 50% of the first postmodifiers were PPs and, in the general topic subcorpus, 42.3% and, in the specific topic subcorpus, 34% was a finite or a non-finite clause (see TABLE 4.8). Again, these results are similar to the other ones given in this thesis concerning the use of postmodifiers in NPs.

Table 4.8 – Complex NPs: Frequencies of postmodifiers in first position after a noun head (per subcorpus)

	General topic subcorpus			**Specific topic subcorpus**		
	raw	per 1000 words	%	raw	per 1000 words	%
Postmodifiers						
Prepositional phrase	24	1.3	53.3	116	3.6	58.9
Finite clause	12	0.6	26.7	38	1.2	19.3
Non-finite clause	7	0.4	15.6	29	0.9	14.7
NP as appositive	2	0.1	4.4	9	0.3	4.6
Adjective	-	-	-	5	0.2	2.5
TOTAL	45		100	197		100

Source: Designed by the author, 2019.

To summarize, section 4.3 has explored the complex NPs produced by Brazilian L2 writers and has showed that learners use slightly more postmodifiers than premodifiers as complex NP constituents. In the research corpus as a whole, premodifying adjectives are the most frequent word class used in complex NPs, whereas postmodifying PPs are the most common type of NP postmodifier used. It should be pointed out that, in the general topic subcorpus, the frequencies of premodifying adjectives and postmodifying PPs are virtually the same (17.6 adjectives vs. 17.7 PPs) (compare GRAPH 4.10 and GRAPH 4.11), whereas, in the specific topic subcorpus, the frequencies of both categories are quite different (14.6 adjectives vs. 21.8 PPs). There are also several possibilities of pre- and postmodification in the same NP, but the use of adjectives and PPs in them are common. Overall, the results indicate learners' strong reliance on these modifiers to compress information in NPs.

Having explored the last category of complex NPs proposed in this study, let us look at NPs that are coordinated.

4.4 Coordinated head nouns

Among the NPs analyzed in this thesis, it was asserted that 94.3% of them were simple or complex NPs. The other 5.7% (334) of NPs analyzed were actually categorized differently, because they were phrases in which head nouns were coordinated by means of the conjunctions *and* and *or*. Three categories have been proposed for these coordinated heads since there was the coordination of: a) two or more simple head as in (49); b) two or more heads where both heads were modified by the same pre- or postmodifier as in (50) or each head was modified by different pre- or postmodifiers as in (51); and c) at least one simple head together with at least one complex head as in (52) (the NPs illustrated are in between square brackets [ ], pre- and postmodifiers, when occurring, are <u>underlined</u> and head nouns are marked in **bold**). Graph 4.12 presents the frequencies of use of these coordinated noun heads per subcorpora.

49) Someone who works hard will get [**experience** and **skills**] because of the practice, and practice makes person perfect, and perfection is a form of success.

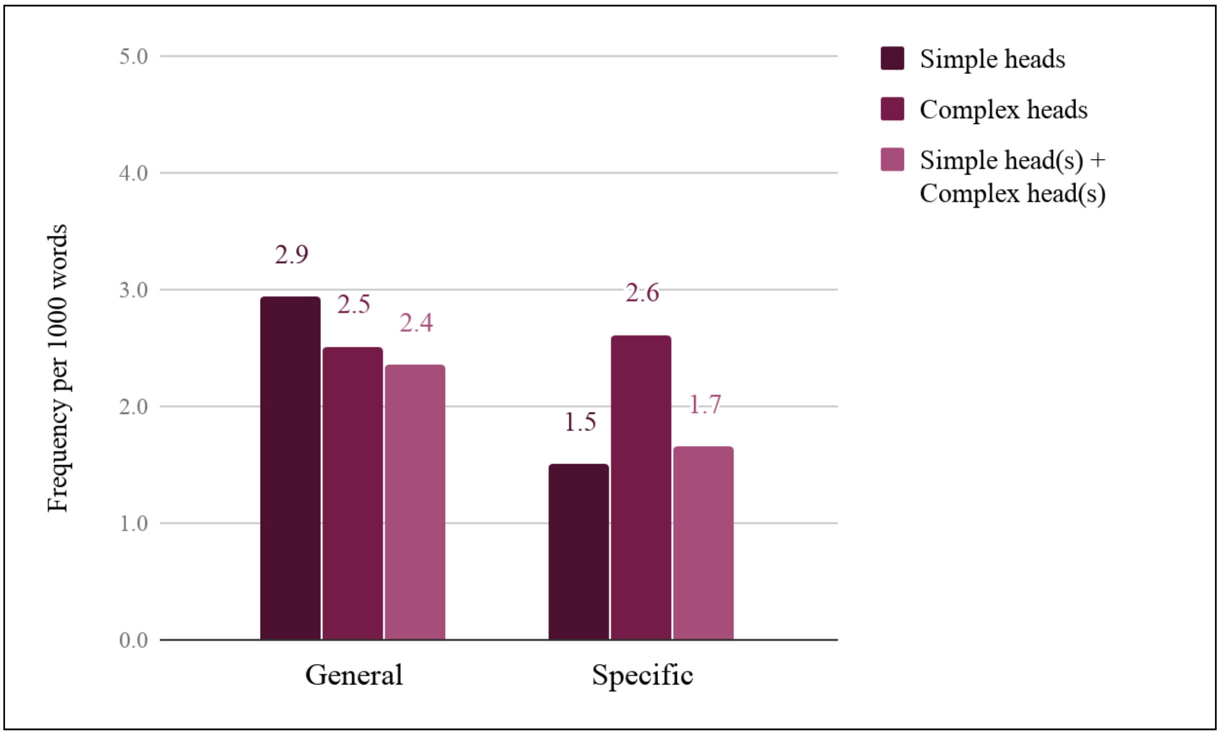
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0732.0463 (General)

50) Research has shown that if pregnant women receive [periodontal **care** and **treatment**], premature births may be reduced by about 45,000 each year.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0992.0574 (Specific)

51) Then, I use data from the Northern Vector of the Metropolitan Area of Belo Horizonte to test the hypothesis that [land **rent**, urban **convention** and the **behavior** of entrepreneurs] are essential dimensions of the urban economy dynamics.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0730.0462 (Specific)

52) In the Enlightenment of the 18th century, [**philosophers** and European **economists**] spread their knowledge, and they deemed themselves propagators of light, comparing this with reason.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0632.0404 (Specific)

Graph 4.12 – Coordinated head nouns (per subcorpus)



Source: Designed by the author, 2019.

In the whole research corpus, 39.5% of these coordinated heads are formed of complex heads, 31.1% are simple heads, and 29.3% are a combination of simple and complex heads. In the general topic subcorpus, coordinated simple heads were more frequent than the others, and, in the specific topic subcorpus, coordinated complex heads were preferred.

Another aspect examined in the NPs with coordinated head nouns was the number of headwords in each phrase (see TABLE 4.9). In the general topic subcorpus, it was found NPs with two to five simple heads, NPs with two to three complex heads, and NPs with two to four heads being at least one simple and one complex. In the specific topic subcorpus, it was found NPs with two to four simple heads, NPs with two to four complex heads, and NPs with two to eight heads, being at least one simple.

All in all, in both subcorpora and in all three defined categories, the majority of NPs had two headwords. In the general topic subcorpus, 86.3% of NPs had two heads while 13.7% had three or more heads. In the specific topic subcorpus, 79.3% had two heads and 20.7% had three or more head nouns. The higher use of coordinated heads in the specific topic subcorpus, for instance, could be seen as a way to modify two or more nouns at the same time, making it a more compressed structure.

Table 4.9 – Coordinated head nouns: Number of heads coordinated (per subcorpus)

		Number of heads					
Subcorpus	**Types of coordinated heads**	2	3	4	5	6	8
General topic	Simple heads	47	6	1	1	-	-
	Complex heads	45	2	-	-	-	-
	Simple and complex heads	34	8	2	-	-	-
		86.3%	11.0%	2.1%	0.7%	-	-
Specific topic	Simple heads	45	3	1	-	-	-
	Complex heads	69	13	1	-	2	-
	Simple and complex heads	35	8	4	5	1	1
		79.3%	12.8%	3.2%	2.7%	1.6%	0.5%

Source: Designed by the author, 2019.

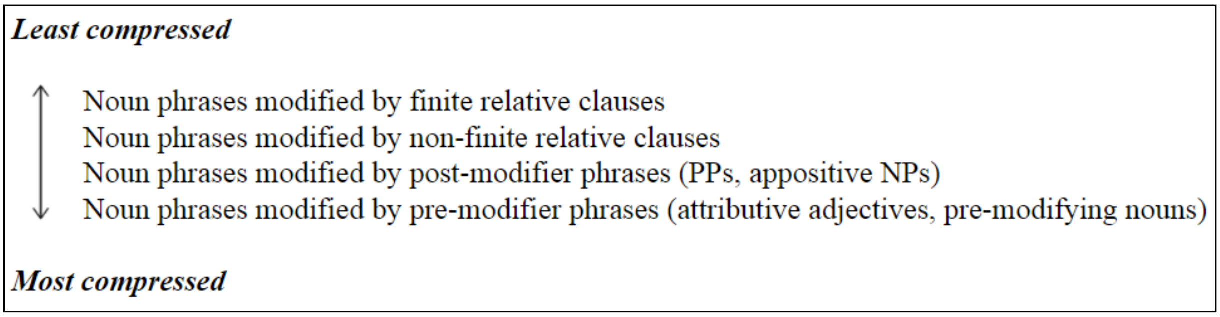
4.5 Structural compression in noun phrases

Phrasal devices, such as the many forms of complex NPs seen previously, are to a great extent the structures that define the grammatical complexity in written academic English (BIBER; GRAY, 2016). As reviewed in Chapter 2, some authors analyze the linguistic features produced by language users in relation to their elaboration or compression (e.g. GRAY, 2015). Those features involve not only phrasal (i.e. NPs) but also clausal (i.e. complement and adverbial clauses) structures, making it possible to determine if a text or register has a more/less elaborated or more/less compressed style. As far as researchers know, academic writing changed over the centuries from a more elaborated style to a more compressed one, which is more economical and efficient to read (BIBER; GRAY, 2010, 2016). In our study, it is not possible to compare the use of elaborated and compressed structures, because we only examined the use of NPs by Brazilian learners. However, future research would benefit from investigating both structures as together they can provide a more comprehensive understanding of grammatical complexity.

In regards to complex NPs, they are considered structures of compression. This means that academic writers would attempt to modify a head noun as much as possible as a way of compressing information in one large phrase. Using noun premodification usually makes an NP more compressed than noun postmodification does, because fewer words are being used in the NP, making the phrase more condensed and the meaning relations inside it less explicit.

In Chapter 2, a cline of structural compression proposed by Biber and Gray (2016) was presented (see it reproduced in this chapter as FIGURE 4.1). In it, some of the NP configurations discussed in this study are evaluated based on their structure as more or less compressed. On the one hand, complex NPs that contain at least one finite clause as postmodifier is considered less compressed, as in excerpt (35) reproduced again as (53), because several words are used to specify the reference of the head noun *books*. On the other hand, NPs that contain one or more premodifying elements are defined as more compressed, such as in example (18) reproduced again as (54), in which only two premodifying nouns add information and perhaps implicit meanings to the headword *cells*.

Figure 4.1 – Cline of structural compression



Least compressed

↕ Noun phrases modified by finite relative clauses
Noun phrases modified by non-finite relative clauses
Noun phrases modified by post-modifier phrases (PPs, appositive NPs)
Noun phrases modified by pre-modifier phrases (attributive adjectives, pre-modifying nouns)

Most compressed

Source: BIBER; GRAY, 2016, p. 207.

53) One of the most present objects in this domain are [**books**, which are, even in a connected world, a powerful tool to consolidate knowledge].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1011.0592 (Specific)

54) As an illustration, we can cite bone marrow transplantation that a patient receiving [blood stem **cells**] from a donor.
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0698.0437 (Specific)

By analyzing the compression of NPs produced by Brazilian learners, we can affirm whether these NPs tend to be more or less compressed in regards to the NPs used by expert writers, which tend to be more compressed (BIBER; GRAY, 2016). We have not counted or categorized our data in terms of their structural compression, but the fact that most of the NPs produced by learners were complex and had premodifying adjectives and postmodifying PPs shows that learners' NPs tend to be more compressed. That tendency could be explained by Brazilian learners' proficiency level (upper intermediate) and perhaps by their contact with specialized texts in English from their own disciplines, which could influence learners' production of more compressed NPs. Still, it should be remarked that learners' complex NPs with pre- and postmodifiers used multiple postmodification, showing they also rely on these less compressed structures, such as example (45) reproduced again as (55), in which four postmodifiers are used to add meaning to the head noun *ship*.

55) An example widely used in engineering schools is [the Liberty class naval cargo **ship**, built in the U.S. during World War II, when a large number of them sank during crossings in the Atlantic Ocean].
Example taken from CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0651.0420 (Specific)

CHAPTER 5

CONCLUSION

Recent studies on English have been interested in understanding the grammatical complexity across registers, in particular the academic register. Most of them have investigated the use of clausal and phrasal level devices in written texts in order to find if their grammatical complexity has more linguistic features of one level or the other. Biber *et al.* (1999), Gray (2015), Biber and Gray (2016), Staples *et al.* (2016), and so many others have demonstrated that the complexity of professional academic written texts is most frequently defined by the use of phrasal features, particularly concerning the use of complex NPs.

In this thesis, we proposed a descriptive and formal analysis, in which we examined the production of simple and complex NPs in a Brazilian university student corpus, i.e. CorIFA. These learners were part of an upper intermediate EAP class and wrote argumentative essays. After the automatic parsing and extraction of NPs from our research corpus and the manual organization of NPs into simple and complex categories, we were able to quantify and analyze this data according to the research objectives defined in Chapter 1.

We discovered that Brazilian learners use more complex NPs than simple NPs. In fact, 59.3% of the NPs analyzed were complex, 35% were simple, and 5.7% involved NPs with coordinated head nouns. That finding refutes our first hypothesis:

1) Brazilian upper intermediate learners use more simple NPs than complex NPs in their English written production.

By identifying each configuration of NP produced by learners, we found out that simple NPs with determiner(s) are the most common type of simple NP, representing 58.3% of these NPs. Concerning complex NPs, 42.1% of them had postmodifier(s), 38.4% had premodifier(s), and 19.5% had both pre- and postmodifiers. From that, our second hypothesis can be confirmed:

2) Brazilian upper intermediate learners produce more NPs with postmodifier(s) than NPs with premodifier(s) or NPs with both pre- and postmodifiers.

As for commonly used NPs by Brazilian writers, complex NPs premodified by adjectives were more frequent than NPs with premodifying nouns and participle forms. Moreover, complex NPs postmodified by PPs were more common than NPs containing postmodifying clauses and appositive NPs. These results were overall anticipated by other studies (e.g. GRAY, 2015; STAPLES *et al.*, 2016; PARKINSON; MUSGRAVE, 2014; NITSCH, 2017; ANSARIFAR *et al.*, 2018), confirming our third and fourth hypotheses:

3) Brazilian upper intermediate learners use more adjectives as NP premodifiers.
4) Brazilian upper intermediate learners use more PPs as NP postmodifiers.

All things considered, it can be assumed that Brazilian learners, due to their proficiency level, the academic context of writing, and the probable contact with specialized texts in English from their own disciplines, are capable of using structurally complex and compressed phrasal structures, often characteristic of professional academic writing. In other words, considering that students have to write an academic argumentative essay, they are expected to try to adequate their writing to the writing they find in the different academic texts they read in their academic life as well as in their search for reference material for the essay. Knowing professional academic writing makes higher use of complex NPs, it is anticipated that learners will attempt to use similar grammatical structures.

Even though Brazilian upper intermediate learners frequently use complex NPs in their texts, EAP classes could still work on NPs premodified by nouns and NPs postmodified by appositive NPs, which are considered the most compressed structures in academic writing (BIBER; GRAY, 2016). They should also focus on explicitly teaching these NP structures and having learners interested in looking for common patterns in their own research field papers.

A final word should be given regarding the limitations of this study and suggestions for improvements. First, our research corpus was rather small and only analyzed its object of study in one genre, i.e. the argumentative essay. Future research could attempt to analyze NPs based on a larger corpus, with several proficiency levels and genres represented. In that same perspective, it would have been interesting to compare the production of NPs depending on the educational level of the writers, i.e. undergraduate or graduate. Second, we did not apply any statistical tests of significance to support our analyses, but that should become an integral part of future studies on the same topic. Third, a good part of the methodology applied in this thesis involved manual work. We believe that it is possible to automate the categorization of NPs, based on the sequences of POS tags found in our corpus. Fourth, we could not have a comprehensive overview of the 5823 NPs analyzed, examining them as one category of NPs and having an idea, for instance, of the most frequent head nouns used. Neither did we analyze the use of both phrasal and clausal structures so as to have a thorough understanding of grammatical complexity in learner writing. Finally, it would be interesting to have more studies on the use of NPs in essays written in Brazilian Portuguese by university students[30] so we could have an idea of L1 tendencies that could have an influence in L2 writing.

[30] A university corpus of Portuguese essays has already been compiled by Silero (2014).

REFERENCES

AKINLOTAN, Mayowa; HOUSEN, Alex. Noun phrase complexity in Nigerian English. *English Today*, v. 33, n. 3, p. 31-38, Sep. 2017.

ALMEIDA, Valdênia C. *Investigando colocações em um corpus de aprendiz*. 2014, 165 p. Thesis (PhD in Applied Linguistics) – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2014.

ANSARIFAR, Ahmad; SHAHRIARI, Hesamoddin; PISHGHADAM, Reza. Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, v. 31, p. 58-71, 2018.

ANTHONY, Laurence. *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. 2018.

BIBER, Douglas. *Variation across speech and writing*. Cambridge University Press,1988. 299 p.

BIBER, Douglas. *University Language: A corpus-based study of spoken and written registers*. John Benjamins Publishing Company, 2006. 261 p.

BIBER, Douglas. Corpus-based and corpus-driven analyses of language variation and use. In: HEINE, Bernd; NARROG, Heiko (ed.). *The Oxford handbook of linguistic analysis*. Oxford University Press, 2010. p. 159-191.

BIBER, Douglas; CONRAD, Susan. *Register, genre, and style*. Cambridge University Press, 2009. 344 p.

BIBER, Douglas; CONRAD, Susan; LEECH, Geoffrey. *Student grammar of spoken and written English*. Longman, 2002. 487 p.

BIBER, Douglas; GRAY, Bethany. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, v. 9, p. 2-20, 2010.

BIBER, Douglas; GRAY, Bethany. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press, 2016. 277 p.

BIBER, Douglas; GRAY, Bethany; POONPON, Kornwipa. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *Tesol Quarterly*, v. 45, n. 1, p. 5-35, 2011.

BIBER, Douglas; GRIEVE, Jack; IBERRI-SHEA, Gina. Noun phrase modification. In: ROHDENBURG, Günter; SCHLÜTER, Julia (ed.). *One language, two grammars? Differences between British and American English*. Cambridge University Press, 2009. p. 182-193.

BIBER, Douglas; JOHANSSON, Stig; LEECH, Geoffrey; CONRAD, Susan; FINEGAN, Edward. *Longman grammar of spoken and written English*. London: Longman, 1999. 1204 p.

BIES, Ann; Ferguson, Mark; Katz, Karen; MacIntyre, Robert; Tredinnick, Victoria; Kim, Grace; Mary; MARCINKIEWICZ, Ann; SCHASBERGER, Britta. *Bracketing guidelines for Treebank II style Penn Treebank projec*t. University of Pennsylvania, Jan. 1995.

BULTÉ, Bram; HOUSEN, Alex. Defining and operationalising L2 complexity. In: HOUSEN, Alex; KUIKEN, Folkert; VEDDER, Ineke (ed.). *Dimensions of L2 Performance and Proficiency*. Amsterdam: John Benjamins Publishing Company, 2012. p. 21-46.

CARTER, Ronald; McCARTHY, Michael. *Cambridge grammar of English: Spoken and written English grammar and usage*. Cambridge University Press, 2006. 973 p.

CRYSTAL, David. *How language works*. Penguin Books, 2005. 500 p.

CROSSLEY, Scott A.; McNAMARA, Danielle S. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, v. 26, p. 66-79, 2014.

DUTRA, Deise P.; BERBER SARDINHA, Tony. Referential expressions in English learner argumentative writing. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny (ed.). *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses universitaires de Louvain, 2013. p. 117-127.

DUTRA, Deise P.; GOMIDE, Andressa R. Compilation of a university learner corpus. *Brazilian English Teaching Journal*, v. 6, p. 21-33, 2015.

DUTRA, Deise P.; ORFANÓ, Bárbara M.; ALMEIDA, Valdênia C. Result linking adverbials in learner corpora. *Domínios de Lingu@gem*, v. 13, n. 1, p. 400-431, 2019.

DUTRA, Deise P.; QUEIROZ, Jessica M. S.; ALVES, Jessica C. Adding information in argumentative texts: A learner corpus-based study of additive linking adverbials. *Revista Estudos Anglo-Americanos*, v. 46, n. 1, p. 9-32, 2017.

DUTRA, Deise P.; SILERO, Rejane W. P. Descobertas linguísticas para pesquisadores e aprendizes: a Linguística de Corpus e o ensino de gramática. *RBLA*, v. 10, n. 4, p. 909-930, 2010.

ELLIS, Rod. *The study of second language acquisition*. Oxford University, 1994.

GRAFFI, Giorgio. *200 Years of Syntax: A critical survey*. John Benjamins Publishing Company, 2001. 551 p.

GRANGER, Sylviane. How to use foreign and second language learner corpora. In: MACKEY, Alison; GASS, Susan M. (ed.). *Research methods in Second Language Acquisition: A practical guide*. Blackwell Publishing Ltd., 2012. p. 7-29.

GRAY, Bethany. *Linguistic variation in research articles: When discipline tells only part of the story*. John Benjamins Publishing Company, 2015. 222 p.

GUEDES, Annallena S. *Verbos do inglês acadêmico escrito e suas colocações: um estudo baseado em um corpus de aprendizes brasileiros de inglês*. 2017, 199 p., Thesis (PhD in Applied Linguistics) – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2017.

HALLIDAY M. A. K.; MARTIN, J. R. *Writing science: Literacy and discursive power*. The Falmer Press, 1993. 309 p.

HINKEL, Eli. What research on second language writing tells us and what it doesn’t. In: ______. (ed.). *Handbook of research in second language teaching and learning*. Routledge, 2011. p. 523-538.

HYLAND, Ken. *Genre and second language writing*. University of Michigan Press, 2004. 244 p.

HYLAND, Ken. *Academic Discourse: English in a Global Context*. Continuum: London, 2009. 215 p.

HYLAND, Ken. ESP and writing. In: PALTRIDGE, Brian; STARFIELD, Sue (ed.). *The handbook of English for specific purposes*. John Wiley & Sons, 2014. p. 95-113.

HYLAND, Ken. Methods and methodologies in second language writing research. *System*, v. 59, p. 116-125, 2016.

LU, Xiaofei. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, v. 45, n. 1, p. 36-62, 2011.

LU, Xiaofei; AI, Haiyang. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, v. 29, p. 16-27, 2015.

MANNING, Christopher D.; SURDEANU, Mihai; BAUER, John; FINKEL, Jenny; BETHARD, Steven J.; McCLOSKY, David. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 55-60, 2014.

MAZGUTOVA, Diana; KORMOS, Judit. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, v. 29, p. 3-15, 2015.

McENERY, Tony; HARDIE, Andrew. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2012. 294 p.

MEYER, Charles F. *English corpus linguistics: An introduction*. Cambridge University Press, 2002. 168 p.

MONTGOMERY, Scott L. *Does science need a global language? English and the future of research*. The University of Chicago Press, 2013. 226 p.

NEWMEYER Frederick J.; PRESTON, Laurel B. Introduction. In: ______. (ed.). *Measuring grammatical complexity*. Oxford University Press, 2014. p. 1-13.

NITSCH, Vanessa C. O. W. *Complexidade dos sintagmas nominais do inglês: um estudo comparativo de corpora de aprendizes brasileiros e falantes nativos de inglês*. 2017. 168 p. Thesis (PhD in Linguistics) – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2017.

NUNAN, David. Exploring genre and register in contemporary English. *English Today*, v. 24, n. 2, p. 56-61, 2008.

OLIVEIRA, Janaína H. *O uso de advérbios intensificadores na escrita de aprendizes brasileiros de inglês como L2: um estudo baseado em corpus*. 2015, 87 p., Thesis (Master's in Applied Linguistics) – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2015.

ORTEGA, Lourdes. Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, v. 29, p. 82-94, 2015.

PALLOTTI, Gabriele. A simple view of linguistic complexity. *Second Language Research*, v. 31, n. 1, p. 117-134, 2015.

PARKINSON, Jean; MUSGRAVE, Jill. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, v. 14, p. 48-59, 2014.

PERCIVAL, W. Keith. On the historical source of immediate constituent analysis. *Syntax and semantics*, v. 7, p. 229-242, 1976.

PERINI, Mário A. *Estudos de gramática descritiva*. São Paulo: Parábola, 2008. 398 p.

POLIO, Charlene. How to research second language writing. In: MACKEY, Alison; GASS, Susan M. (ed.). *Research methods in Second Language Acquisition: A practical guide*. Blackwell Publishing Ltd., 2012. p. 139-157.

QUIRK, Randolph; GREENBAUM, Sidney; LEECH Geoffrey; SVARTVIK, Jan. *A comprehensive grammar of the English language*. New York: Longman, 1985. 1779 p.

RANSDELL, Sarah; BARBIER, Marie-Laure. An introduction to new directions for research in L2 writing. In: ______. (ed.). *New directions for research in L2 writing*. Springer, Dordrecht, 2002. p. 1-10.

SALAGER-MEYER, Françoise; SEGURA, Graciela M. L.; RAMOS, Rosinda C. G. EAP in Latin America. In: HYLAND, Ken; SHAW, Philip (ed.). *The Routledge handbook of English for academic purposes*. Routledge, 2016. p. 109-124.

SCOTT, Mike; TRIBBLE, Christopher. *Textual patterns: Key words and corpus analysis in language education.* John Benjamins Publishing, 2006. 203 p.

SELINKER, Larry. Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, v. 10, p. 209-232, 1972.

SELINKER, Larry. Interlanguage 40 years on. In: HAN, ZhaoHong; TARONE, Elaine (ed.). *Interlanguage: Forty years later*. John Benjamins Publishing Company, 2014. p. 221-246.

SILERO, Rejane W. P. *Os quantificadores a few e few: questões de interlíngua e prosódia semântica em corpus de aprendizes*. 2014. 107 p. Thesis (Master's in Applied Linguistics) – Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2014.

STAPLES, Shelley; REPPEN, Randi. Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, v. 32, p. 17-35, 2016.

STAPLES, Shelley; EGBERT, Jess; BIBER, Douglas; GRAY, Bethany. Academic writing development at university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, v. 33, n. 2, p. 149-183, 2016.

TAGUCHI, Naoko; CRAWFORD, William; WETZEL, Danielle Z. What Linguistic Features Are Indicative of Writing Quality? A Case of Argumentative Essays in a College Composition Program. *Tesol Quarterly*, v. 47, p. 420-430, 2013.

TAYLOR, Ann; MARCUS, Mitchell; SANTORINI, Beatrice. The Penn Treebank: An overview. In: ABEILLÉ, Anne (ed.). *Treebanks*: Building and using parsed corpora. Springer, Dordrecht, 2003. p. 5-22.

WIDDOWSON, H. G. *Linguistics*. Oxford University Press, 1996. 134 p.

APPENDIX A – List of Penn Treebank Project tags[31]

Clause Level

Tag	**Meaning**
S	Simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a *wh*-word and that does not exhibit subject-verb inversion.
SBAR	Clause introduced by a (possibly empty) subordinating conjunction.
SBARQ	Direct question introduced by a *wh*-word or a *wh*-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
SINV	Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
SQ	Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.

Phrase Level

Tag	**Meaning**
ADJP	Adjective phrase
ADVP	Adverb phrase
CONJP	Conjunction phrase
FRAG	Fragment
INTJ	Interjection. Corresponds approximately to the part-of-speech tag UH.
LST	List marker. Includes surrounding punctuation.
NAC	Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.
NP	Noun phrase
NX	Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.
PP	Prepositional Phrase
PRN	Parenthetical
PRT	Particle. Category for words that should be tagged RP.
QP	Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.
RRC	Reduced Relative Clause
UCP	Unlike Coordinated Phrase
VP	Verb Phrase
WHADJP	Wh-adjective Phrase. Adjectival phrase containing a wh-adverb, as in how hot.
WHADVP	Wh-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a wh-adverb such as how or why.

[31] This list was organized based on Bies *et al.* (1995) and Taylor *et al.* (2003).

WHNP	Wh-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some wh-word, e.g. who, which book, whose daughter, none of which, or how many leopards.
WHPP	Wh-prepositional Phrase. Prepositional phrase containing a wh-noun phrase (such as of which or by whose authority) that either introduces a PP gap or is contained by a WHNP.
X	Constituent of unknown or uncertain category

Word level

Tag	**Meaning**
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	infinitival *to*
UH	Interjection
VB	Verb, base form

VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive *wh*-pronoun (prolog version WP-S)
WRB	Wh-adverb
#	Pound sign
$	Dollar sign
.	Sentence-final punctuation
,	Comma
:	Colon, semi-colon
(	Left bracket character
)	Right bracket character
"	Straight double quote
‘	Left open single quote
“	Left open double quote
’	Right close single quote
”	Right close double quote

APPENDIX B – Questions from the Google Form answered by students[32]

1. **Full name**
2. **Age**
   - **a.** Less than 18 years old
   - **b.** 18-25 years old
   - **c.** 26-35 years old
   - **d.** 36-45 years old
   - **e.** 46-55 years old
   - **f.** 56-65 years old
3. **Gender**
   - **a.** Female
   - **b.** Male
   - **c.** Others
   - **d.** Prefer to not answer
4. **EAP class number**
5. **Text version**
   - **a.** 1st draft
   - **b.** 2nd draft
   - **c.** 3rd draft
6. **TOEFL ITP score**
7. **UFMG registration number**
8. **Undergraduate major or graduate program**
9. **Academic level**
   - **a.** Undergraduate
   - **b.** Graduate (Master's)
   - **c.** Graduate (Doctorate)
   - **d.** Other
10. **How long have you been studying English?**
    - **a.** Never studied
    - **b.** Less than one year
    - **c.** One year or more but less than two years
    - **d.** Two years or more but less than five years
    - **e.** Five years or more but less than ten years
    - **f.** Ten years or more
11. **Have you ever been to an English speaking country?**
    - **a.** No
    - **b.** Yes, for more than one month but less than six months
    - **c.** Yes, for more than six months but less than one year
    - **d.** Yes, for more than one year
12. **What is your mother tongue?**
13. **Email**

[32] Students' answers become CorIFA metadata (except for full name, email, UFMG registration number).

APPENDIX C – CorIFA informed consent term

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

O Sr. (a) está sendo convidado (a) como voluntário (a) a participar da pesquisa “Traços linguísticos do discurso acadêmico: um estudo de corpora de aprendiz e de textos científicos especializados”. Pedimos a sua autorização para a coleta e análise de seus textos escritos em inglês. A utilização de seus textos está vinculada somente a este projeto de pesquisa ou se Sr. (a) concordar em outros futuros. Nesta pesquisa pretendemos descrever o uso da língua inglesa por aprendizes brasileiros universitários de inglês bem como comparar esse uso com traços linguísticos presentes em artigos acadêmico-científicos. Para esta pesquisa os participantes farão tarefas em disciplinas e em cursos de inglês, utilizando seus conhecimentos prévios. Os desconfortos dos participantes são mínimos, podendo se sentir pressionados por saberem que seus textos farão parte de uma pesquisa. Esses desconfortos serão minimizados, pois as atividades de coleta de dados são atividades comumente feitas em sala de aula. Além disso, garanto que identificação de nenhum dos participantes será divulgada. A pesquisa contribuirá para o aumento do conhecimento a respeito dos processos de aquisição favorecidos com a utilização de corpora eletrônicos, podendo beneficiar outros aprendizes de inglês.

Para participar deste estudo o Sr. (a) não terá nenhum custo, nem receberá qualquer vantagem financeira. O Sr. (a) terá o esclarecimento sobre o estudo em qualquer aspecto que desejar e estará livre para participar ou recusar-se a participar e a qualquer tempo e sem quaisquer prejuízos, pode retirar o consentimento de participação na pesquisa, valendo a desistência a partir da data de formalização desta. A sua participação é voluntária, e a recusa em participar não acarretará qualquer penalidade ou modificação na forma em que o Sr. (a) é atendido (a) pelo pesquisador, que tratará a sua identidade com padrões profissionais de sigilo. Os resultados obtidos pela pesquisa estarão à sua disposição quando finalizada. Seu nome ou o material que indique sua participação não será liberado sem a sua permissão. O (A) Sr. (a) não será identificado (a) em nenhuma publicação que possa resultar.

Este termo de consentimento encontra-se impresso em duas vias originais, sendo que uma será arquivada pelo pesquisador responsável, na sala 4111 da Faculdade de Letras da UFMG, e a outra será fornecida ao Sr. (a). Os dados e materiais utilizados na pesquisa ficarão arquivados com o pesquisador responsável por um período de 5 (cinco) anos na sala 4111 da Faculdade de Letras da UFMG e após esse tempo serão destruídos. Os pesquisadores tratarão a sua identidade com padrões profissionais de sigilo, utilizando as informações somente para fins acadêmicos e científicos.

Eu, ____, portador do documento de Identidade ____________________ fui informado (a) dos objetivos, métodos, riscos e benefícios da pesquisa “Traços linguísticos do discurso acadêmico: um estudo de corpora de aprendiz e de textos científicos especializados”, de maneira clara e detalhada e esclareci minhas dúvidas. Sei que a qualquer momento poderei solicitar novas informações e modificar minha decisão de participar se assim o desejar.

( ) Concordo que os meus textos escritos em inglês sejam utilizados somente para esta pesquisa.

( ) Concordo que os meus textos escritos em inglês possam ser utilizados em outras pesquisa, mas serei comunicado pelo pesquisador novamente e assinarei outro termo de

Rubrica do pesquisador: ____________

*Rubrica do participante:*____________

Declaro que concordo em participar desta pesquisa. Recebi uma via original deste termo de consentimento livre e esclarecido assinado por mim e pelo pesquisador, que me deu a oportunidade de ler e esclarecer todas as minhas dúvidas.

Nome completo do participante: ___________________________________ Data: _________

____

Assinatura do participante

Nome completo do Pesquisador Responsável: Deise Prina Dutra
Endereço:
Telefones:
E-mail:
CPF:
RG:

Assinatura do pesquisador responsável: _______________________________ Data: _______

Nome completo do Pesquisador: Jessica Maria da Silva Queiroz
Endereço:
Telefones:
E-mail:
CPF:
RG:

Assinatura do pesquisador (mestrando ou doutorando): ________________ Data: _______

Em caso de dúvidas, com respeito aos aspectos éticos desta pesquisa, você poderá consultar:

COEP-UFMG - Comissão de Ética em Pesquisa da UFMG
Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Sala 2005.
Campus Pampulha. Belo Horizonte, MG – Brasil. CEP: 31270-901.
E-mail: coep@prpq.ufmg.br. Tel: 34094592.

APPENDIX D – Codes of the CorIFA argumentative essays used for analysis[33]

2015-2 (Total number of texts = 10)

CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0578.0353	CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0609.0383
CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0582.0356	CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0611.0385
CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0598.0372	CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0612.0237
CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0599.0373	CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0622.0395
CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0606.0380	CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0624.0397

2016-1 (Total number of texts = 18)

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0630.0402	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0646.0416
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0631.0403	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0647.0417
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0632.0404	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0648.0346
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0634.0406	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0649.0418
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0635.0407	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0650.0419
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0638.0410	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0651.0420
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0639.0411	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0652.0421
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0640.0412	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0653.0422
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0644.0415	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0654.0423

2016-2 (Total number of texts = 42)

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0688.0430	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0718.0450
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0689.0329	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0719.0451
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0690.0431	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0720.0452
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0691.0432	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0721.0453
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0692.0433	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0722.0454
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0693.0434	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0724.0456
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0696.0436	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0725.0457
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0698.0437	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0728.0460
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0699.0408	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0729.0461
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0700.0337	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0730.0462
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0702.0439	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0731.0106
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0703.0405	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0732.0463
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0705.0441	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0733.0464

[33] Texts in grey are part of the general topic subcorpus while texts in white are part of the specific topic subcorpus.

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0706.0442	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0734.0465
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0707.0443	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0735.0466
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0709.0444	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0736.0467
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0711.0425	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0737.0468
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0712.0446	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0738.0322
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0714.0331	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0739.0469
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0715.0448	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0741.0471
CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0717.0449	CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0742.0472

2017-1 (Total number of texts = 18)

CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0934.0541	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0947.0551
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0935.0542	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0948.0552
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0936.0543	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0949.0553
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0937.0509	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0950.0554
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0940.0545	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0951.0555
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0942.0547	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0952.0556
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0943.0548	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0953.0557
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0944.0549	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0954.0558
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0945.0550	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0955.0559

2017-2 (Total number of texts = 26)

CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0981.0563	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0996.0578
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0982.0564	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0997.0579
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0983.0565	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0998.0580
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0984.0566	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0999.0581
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0986.0568	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1000.0582
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0987.0569	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1001.0525
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0988.0570	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1002.0583
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0989.0571	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1003.0584
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0990.0572	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1006.0587
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0992.0574	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1008.0589
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0993.0575	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1010.0591
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0994.0576	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1011.0592
CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.0995.0577	CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1014.0593

APPENDIX E – Codes of the texts excluded from subcorpus

Texts with no metadata available (Total number of texts = 107)

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0277.0140
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0278.0141
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0279.0142
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0280.0143
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0281.0144
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0282.0145
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0283.0146
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0284.0147
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0285.0148
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0286.0149
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0287.0150
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0288.0151
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0289.0152
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0290.0153
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0291.0154
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0292.0155
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0293.0156
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0294.0157
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0295.0158
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0296.0159
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0297.0160
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0298.0161
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0299.0162
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0300.0163
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0301.0128
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0302.0164
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0303.0165
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0305.0167
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0306.0168
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0307.0169
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0309.0171
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0311.0173
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0313.0141
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0314.0142
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0315.0143
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0318.0174
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0319.0175
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0320.0146
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0495.0280
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0496.0281
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0497.0034
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0498.0282
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0499.0119
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0500.0283
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0501.0092
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0502.0284
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0503.0285
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0504.0286
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0505.0287
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0506.0011
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0507.0288
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0508.0289
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0509.0290
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0510.0291
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0511.0292
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0512.0293
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0513.0294
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0514.0295
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0516.0297
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0517.0298
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0518.0299
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0519.0300
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0520.0141
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0521.0161
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0522.0145
CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0523.0301
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0352.0140
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0353.0141
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0354.0142
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0355.0143
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0357.0145
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0358.0174
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0359.0146
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0360.0147
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0361.0148
CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0363.0150

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0321.0147

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0325.0151

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0326.0153

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0327.0154

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0334.0178

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0335.0159

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0338.0162

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0339.0163

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0341.0164

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0342.0165

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0344.0167

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0346.0168

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-1.0347.0169

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0492.0277

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0493.0278

CorIFA-UFMG-B2.Ind.Ne.AEss.2013-2.0494.0279

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0364.0151

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0365.0152

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0366.0153

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0368.0154

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0369.0155

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0370.0156

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0371.0176

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0372.0157

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0373.0160

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0374.0161

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0375.0164

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0376.0165

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0377.0168

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0379.0171

CorIFA-UFMG-B2.Ind.Ne.Ess.2013-1.0381.0173

Texts by students with L1 other than Portuguese (Total number of texts = 3)

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0701.0438

CorIFA-UFMG-B2.Ind.Ne.AEss.2017-1.0939.0544

CorIFA-UFMG-B2.Ind.Ne.AEss.2017-2.1009.0590

Texts by students who have another text in the corpus (Total number of texts = 7)

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0641.0329

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0633.0405

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0716.0407

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0636.0408

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0697.0412

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0643.0414

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-1.0657.0425

Text that was duplicated (Total number of texts = 1)

CorIFA-UFMG-B2.Ind.Ne.AEss.2015-2.0588.0362

Text that could not be automatically parsed (Total number of texts = 3)

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0694.0414

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0704.0440

CorIFA-UFMG-B2.Ind.Ne.AEss.2016-2.0710.0445

APPENDIX F – Prompts or topics assigned to EAP students

Specific topics prompt

Your essay should be about a topic of your interest, preferably in your study field.

Some specific topics chosen by students

- Abortion
- Active learning
- ADHD
- Air pollution
- Animals
- Autonomous Driving
- Belo Monte hydroelectric
- Blood donation
- Breathing
- Business
- Carbonaceous materials
- Civil construction
- Clinical pharmacist
- CO_2
- Corruption
- Dengue
- Dentists
- Digital Arts
- Educational system failure
- Electric cars
- Embryonic stem cell
- Engineering materials
- Environmental preservation
- Exercise and injury prevention
- Freedom of speech
- Graphene
- Grids
- Guitar
- Having one child
- Hydropower plants
- Inclusive Education
- Interdisciplinarity
- International remittances
- Internet and relationships
- Introversion
- Journalists
- Language skills at academia
- Light
- Magnetic systems
- Math
- Media
- Medicine production
- Minas Gerais economy
- Natural resources
- Neuroscience
- Periodontal disease and premature delivery
- Photovoltaic power
- Physics
- Poly
- Privacy
- Public and private education
- Quota
- Salinas formation
- Samarco
- Science
- Strength training
- Surveys on mammals
- Teeth Whitening
- Transgenic food
- Urban economy
- Voltage instability
- Wastewater
- Water

General topics

1. **Children and technology:** *Some people believe computer and video games are harmful to children, while others disagree. What do you think? Take a position and defend it.*

2. **Foreign language:** *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 opinion.

3. **Living longer:** *In general, people are living longer now. Discuss the causes and the implications of this phenomenon. Use specific reasons and details to develop your essay.*
4. **Talent or hard work:** *Which is more important: talent or work hard?*
5. **Technology and loneliness:** *Does technology make us more alone?*

APPENDIX G – Python script 1: Using the Stanford CoreNLP[34]

```
from stanfordcorenlp import StanfordCoreNLP
from glob import glob
import logging
import os
import sys
import shutil
import json
import re

sourcedir = r'C:\Users\jessi\Desktop\Jessica\CorIFA_Dissertation Subcorpus'
outputdir = r'C:\Users\jessi\Desktop\Jessica\CorIFA_Parsed Subcorpus'
nlp_dir = r'C:\Users\jessi\Desktop\Jessica\Mestrado UFMG\Stanford
Parser\stanford-corenlp-full-2018-02-27'

def checkdir(dir):
      # Checking whether the directories exist or not
      if not os.path.isdir(dir):
      os.makedirs(dir, exist_ok=True)

def parsetree(nlp, data):
      # Reading the parsed text
      props = {'annotators': 'ssplit, parse', 'pipelineLanguage': 'en',
'outputFormat': 'json'}
      try:
      parsed_data = nlp.annotate(data, properties=props)
      json_data = json.loads(parsed_data)
      except json.JSONDecodeError:
      return False, parsed_data  # Return failure and the unparsed data.
      filedata = ""  # Final file data
      for sentence in json_data['sentences']:  # Running for each sentence
parsed
      # Joining the several sentences for the parsed tree.
      filedata += ''.join(sentence['parse']) + '\n'
      return True, filedata

def removebrackets(data):
      # Ignoring any information in between angle brackets
      return re.sub(r'<.*?>', '', data)


def main() :
      nlp = StanfordCoreNLP(nlp_dir, logging_level=logging.WARNING) # Use DEBUG
for debug info
      shutil.rmtree(outputdir, ignore_errors=True)  # Removing all files in the
output directory
```

[34] The Python script was written by the programmer Euller Borges.

```
       result = glob(sourcedir + '/**/*.txt', recursive=True)
       failed_files = []  # Files that could not be parsed
       for filepath, i  in zip(result, range(len(result))):
       print("Processing file {0}: {1}".format(i, filepath))

       try:
             with open(filepath, 'r', encoding="utf8") as f:
             relative = os.path.relpath(filepath, sourcedir)
             treedir = os.path.dirname(relative)
             checkdir(os.path.join(outputdir, treedir))
             with open(os.path.join(outputdir, relative), 'w', encoding="utf8")
as t:
                    content = f.read()
                    content = removebrackets(content)
                    success, parsed_data = parsetree(nlp, content)
                    if not success:
                     print("Could not parse file {}".format(filepath))
                          failed_files.append(filepath)
                    t.write(parsed_data)

       except FileNotFoundError as exc:
             print("Error when opening file: " + filepath)
             sys.exit(1)

       if failed_files:
       print("\nWARNING - The following files failed to be parsed: ")
       for file in failed_files:
             print(file)

       print("\nParsing of files completed successfully with {}
errors".format(len(failed_files)))
       nlp.close()
```

APPENDIX H – Python script 2: Extracting noun phrases from parsed data[35]

```
import csv
import nltk.tree
import os
from glob import glob

PHRASE_LEVEL_IDS = ["ADJP", "ADVP", "CONJP", "FRAG", "INTJ", "LST", "NAC", "NP",
"NP-TMP", "NX", "PP", "PRN", "PRT", "QP", "RRC", "UCP", "VP", "WHADJP ",
"WHADVP", "WHNP", "WHPP", "X"]

STD_POS_LEVEL_IDS = ["CC", "CD", "DT", "EX", "FW", "IN", "JJ", "JJR", "JJS",
"LS", "MD", "NN", "NNS", "NNP", "NNPS", "PDT", "POS", "PRP", "PRP$", "RB",
"RBR", "RBS", "RP", "SYM", "TO", "UH", "VB", "VBD", "VBG", "VBN", "VBP", "VBZ",
"WDT", "WP", "WP$", "WRB"]
# These below are not really POS IDs, but we want to show them as if they were.
EXTRA_POS_LEVEL_IDS = [",", "-RRB-", "-LRB-"]
POS_LEVEL_IDS = STD_POS_LEVEL_IDS + EXTRA_POS_LEVEL_IDS  # Effective POS Level
IDs.

CLAUSE_LEVEL_ID = ["S", "SBAR", "SBARQ", "SINV", "SQ"]

IGNORE_LEVEL_LIST = ["$", ":", "``", "''", "."]

def np_filter_fn(tree):
        """
        Filters the tree to find elements that should be analyzed.
        :param tree: The tree we are currently analyzing.
        :return: True if this tree is valid, False otherwise.
        """
        if (tree.label() != "NP" and tree.label() != ""):
        # We are ignoring elements different from NP and NPs preceded by PP.
        return False
        subtree = tree
        while subtree != None:
        # Traversing the tree backwards to see if tree is an NP nested in another
NP.
        if subtree.parent() and subtree.parent().label() == "NP":
                return False
        # Continue traversing parent
        subtree = subtree.parent()
        return True

class Traversal(object):
        def __init__(self, tree):
        self.phrase_tag_seq = []
        self.pos_tag_seq = []
        self.pos_values = []
        self.last_elem_is_POS = False
        self.POS_translation_map = {
```

[35] The Python script was written by the programmer Euller Borges.

```
                "-RRB-": ")",
                "-LRB-": "("
        }
        self.traverse(tree)

        def translate(self, str):
        """
        Converts tags back to their original representation. If it is not a tag,
returns the word itself.
        :param str: The string to analyze.
        :return: None
        """
        if str in self.POS_translation_map:
                return self.POS_translation_map[str]
        else:
                return str

        def traverse(self, node):
        """
        Traverses the tree and fills the phrase level elements, POS elements,
etc.
        :param node: The node we are currently analyzing.
        :return: None
        """
        if isinstance(node, nltk.Tree):  # This is a subtree
                if node.label() in PHRASE_LEVEL_IDS:
                self.phrase_tag_seq.append(node.label())
                elif node.label() in POS_LEVEL_IDS:
                 self.pos_tag_seq.append(self.translate(node.label()))
                self.last_elem_is_POS = True
                elif node.label() not in CLAUSE_LEVEL_ID and node.label() not in
IGNORE_LEVEL_LIST:
                # We ignore clause level IDs.
                 print ("WARNING: Unexpected element '{}' in parse
tree.".format(node.label()))
                for child in node:
                self.traverse(child)
        elif self.last_elem_is_POS: # This is a leaf (string element)
                self.pos_values.append(self.translate(node))
        self.last_elem_is_POS = False  # Resetting POS status

        def get_pos_tag_seq(self):
        return " ".join(self.pos_tag_seq)

        def get_phrase_tag_seq(self):
        return " ".join(self.phrase_tag_seq)

        def count_POS_elements(self):
        # We have already translated the elements at this point using
translate(), thus we
        # must check if the elements are in the values of the translation map. We
also must check if
        # the element is not in any of the untranslated extra POS elements.
```

```
       return len([elem for elem in self.pos_values if elem not in
self.POS_translation_map.values()
                    and elem not in EXTRA_POS_LEVEL_IDS])


def sentence_to_csv(data, sentence_id, input_filename, csv_writer,
pp_np_csv_writer):
       """
       Analyzes a given sentence, filtering the NP elements, and outputs it to
the csv file.
       :param data: Parsed sentence string.
       :param sentence_id: ID of the sentence in the input file.
       :param input_filename: File name from which the sentence was extracted.
       :param csv_writer:  The configured CSV writer used to output to the CSV
file.
       :param pp_np_csv_writer: The configured PP-NP CSV writer used to output
to the CSV file
    :return: None
       """
       try:
       tree = nltk.ParentedTree.fromstring(data)
       except ValueError as exc:
       print("Error while parsing {}: {}".format(input_filename, str(exc)))
       return
       for s in tree.subtrees(np_filter_fn):
       traversed = Traversal(s)
       assert(len(traversed.pos_tag_seq) == len(traversed.pos_values))
       if s.parent().label() == "PP" and s.parent().parent().parent().label() ==
"ROOT":
             writer = pp_np_csv_writer
             s = s.parent()
           else:
             writer = csv_writer
       writer.writerow({"Corpus File": input_filename,
                          "NP extracted from parser": s,
                          "Sentence number": sentence_id,
                          "NP (phrase tag sequence)":
traversed.get_phrase_tag_seq(),
                          "NP (POS tag sequence)": traversed.get_pos_tag_seq(),
                          "NP (word sequence)": ' '.join(traversed.pos_values),
                          "Number of words per NP":
traversed.count_POS_elements()})

def process_directory(dir, csv_output_path, pp_np_csv_output_path):
       """
       Processes a given directory with files with the parsed trees and writes a
CSV file to the configured output path.
       :param dir: Directory to parse.
       :param csv_output_path: Path to which to write the CSV.
       :param pp_np_csv_output_path: Path to which to write the CSV with the
PP-NP information.
       :return: None
       """
       result = glob(dir + '/**/*.txt', recursive=True)
       num_files = len(result)
       try:
       csv_output_file = open(csv_output_path, 'w', newline='')
       pp_np_csv_output_file = open(pp_np_csv_output_path, 'w', newline='')
       except Exception as exc:
       print("Error opening output file {}: {}".format(csv_output_path,
str(exc)))
       exit(1)

       fieldnames = ["Corpus File", "Sentence number", "NP extracted from
parser", "NP (phrase tag sequence)", "NP (POS tag sequence)", "NP (word
sequence)", "Number of words per NP"]  # CSV header
       csv_writer = csv.DictWriter(csv_output_file, fieldnames=fieldnames,
delimiter=";")
       pp_np_csv_writer = csv.DictWriter(pp_np_csv_output_file,
fieldnames=fieldnames, delimiter=";")
       csv_writer.writeheader()
       for filepath, i in zip(result, range(num_files)):
       print("Processing file {0}/{1}: {2}".format(i+1, num_files, filepath))
       with open(filepath, "r") as file:
              content = file.read()
              sentences = content.split("\n###\n")
              for sentence, id in zip(sentences, range(len(sentences))):
              if sentence:
                     sentence_to_csv(sentence, id, os.path.basename(filepath),
csv_writer, pp_np_csv_writer)

       csv_output_file.close()

def main():
       sourcedirs = [r'C:\Users\jessi\Desktop\Jessica\CorIFA_Parsed
Subcorpus\General topic', r'C:\Users\jessi\Desktop\Jessica\CorIFA_Parsed
Subcorpus\Specific topic']
       for sourcedir in sourcedirs:
       process_directory(sourcedir,
                          os.path.join(sourcedir, os.path.basename(sourcedir) +
" extracted.csv"),
                          os.path.join(sourcedir, os.path.basename(sourcedir) +
" pp_np" " extracted.csv"))

if __name__ == "__main__":
       main()
```