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

JESSICA CERITELLO ALVES

GRAMMATICAL COMPLEXITY IN A LEARNER CORPUS: ASSESSING STUDENTS' DEVELOPMENT THROUGH A LONGITUDINAL STUDY

BELO HORIZONTE
2022

JESSICA CERITELLO ALVES

GRAMMATICAL COMPLEXITY IN A LEARNER CORPUS: ASSESSING STUDENTS' DEVELOPMENT THROUGH A LONGITUDINAL 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

2022

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

FOLHA DE APROVAÇÃO

GRAMMATICAL COMPLEXITY IN A LEARNER CORPUS: ASSESSING STUDENTS' DEVELOPMENT THROUGH A LONGITUDINAL STUDY

JESSICA CERITELLO ALVES

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 2022, pela banca constituída pelos membros:

Prof(a). Deise Prina Dutra - Orientadora

UFMG

Prof(a). Lucia de Almeida Ferrari

UFMG

Prof(a). Ana Elisa Pereira Bocorny

UFRGS

Belo Horizonte, 25 de fevereiro de 2022.

Documento assinado eletronicamente por **Lucia de Almeida Ferrari**, **Professora do Magistério Superior**, em 25/02/2022, às 17:37, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Deise Prina Dutra**, **Professora do Magistério Superior**, em 25/02/2022, às 17:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Eliza Pereira Bocorny**, **Usuária Externa**, em 03/03/2022, às 10:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1225148** e o código CRC **D77FEF57**.

Referência: Processo nº 23072.205770/2022-12 SEI nº 1225148

Alves, Jessica Ceritello.
A474g Grammatical Complexity in a Learner Corpus [manuscrito] : assessing student's development through a longitudinal study / Jessica Ceritello Alves. – 2022.

146 f., enc.: il., grafs, tabs, color.

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. 123-127.

Apêndices: f. 128-146.

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

CDD: 425.07

Ficha catalográfica elaborada pela bibliotecária Priscila Oliveira da Mata – CRB/6-2706

To my family. To the world. With love.

ACKNOWLEDGMENTS

To God, for having given me life, this amazing journey of discoveries and learning, and for never letting me give up on my goals and dreams.

To my dear advisor, prof. Deise, for walking by my side for so long, always sharing her expertise and encouraging me, even when I was faced with obstacles and uncertainties. Thank you so much!

To my son, Raul, who has participated in all this process with me since the belly. I want your company forever, my darling, in the happy and sad moments. I love you!

To my husband, Filipe, for showing so much love and patience during this master's course, for always supporting and encouraging me, even when things were hard. I love you!

To my parents, Silvana and João, who, even without schooling, came to know the value of education and encouraged me to follow my dreams. I am here today because of you. I love you so much!

To my grandma, Elza, my grandma, Maria (in memoriam), my in-laws, my godparents, my dear brother, Victor, my cousins, Carol and Nathy, and all my family. Thank you so much! Without you, I would not have had the strength to carry on. Thank you for all your support, always.

To my colleagues from the GECEA group, for all the discussions, patience, and willingness to help me. Our meetings and discussions were always rewarding and have been key components of my human and academic formation. Thank you!

To prof. Andressa Gomide, for her willingness to help me.

To prof. Bethany Gray, for showing so much patience, willingness, and helpfulness.

To prof. Bárbara Dias, for helping me with everything and being the best sister-in-law ever.

To my friends from Faculdade de Letras, especially Angélica, Emmanuelle, Martim (and Lu), Rubens, and Solaine, who have been sharing their respect, love, and understanding to this very day. Thank you for your friendship, for all your love, support, and cheerfulness in life. I love you all!To my professors at Faculdade de Letras, who have always taught me so much and inspired me many times.

"*Oh, tell me where your freedom lies*
The streets are fields that never die
Deliver me from reasons why
You'd rather cry
I'd rather fly"

The Doors, The Crystal Ship.

ABSTRACT

Grammatical complexity, which is defined "as the addition of structural elements to 'simple' phrases and clauses" (BIBER et al., 2020, p. 5) in written and spoken texts, has been studied from the perspective of first (L1) and second languages (L2) (e.g., STAPLES et al., 2016, BIBER et al., 2020). Unlike previous research grounded on the same measures for both speaking and writing (e.g., WOLFE-QUINTERO et al., 1998), Biber et al. (2011) present a hypothesized developmental index with five different stages containing phrasal and clausal features. According to that index, low proficiency students begin writing by adopting features more commonly found in speech (i.e., finite adverbial clauses), and, as their proficiency increases, they come to rely on phrasal features as well (i.e., nouns as pre-modifiers), which are more typical in written discourses. To the best of our knowledge, no research has investigated the development stages among Brazilian English learners longitudinally. For example, Queiroz (2018) has analyzed complex noun phrases (NPs) in Brazilian university students' written texts cross-sectionally. Thus, in order to fill this gap, this study aims at capturing the development of grammatical complexity of Brazilian EAP students,' according to the framework proposed by Biber et al. (2011). To this end, a longitudinal subcorpus from CorIFA (Corpus of English for Academic Purposes) was collected. The subcorpus contains texts written by 13 students (n = 13) who attended the EAP courses for three semesters. Each student wrote three texts, six months apart, so this research analyzes three moments in time. The results of the longitudinal study demonstrate that not all features from the framework presented statistical significance over time, but most of the phrasal features expected to increase in frequency showed a positive outcome from Time 1 to Time 3 (e.g., the use of the attributive adjective increased significantly). However, even not statistically, some clausal features showed a slight increase over time instead of decreasing (e.g., finite adverbial clauses). Moreover, a quasi-longitudinal analysis of register and academic division revealed differences in the preference for features, which are often related to the registers' communicative purposes and the academic divisions' specificities. At last, when compared with Staples et al., (2016) the L1 study results show that the development of Brazilian and native learners is not completely similar, especially in the scope of clausal features. Indeed, the use of such constructions among natives decreased, but this did not happen for all clausal features in our subcorpus.

Keywords: grammatical complexity, longitudinal study, quasi-longitudinal study, learner corpus

RESUMO

A complexidade gramatical definida “como a adição de elementos estruturais a frases e orações 'simples'” (BIBER et al., 2020, p. 5), em textos escritos e falados, tem sido estudada tanto na primeira língua (L1) quanto na segunda língua (L2) (STAPLES et al., 2016, BIBER et al., 2020). Diferentemente de pesquisas anteriores, que se baseiam nas mesmas medidas tanto para fala quanto para escrita (WOLFE-QUINTERO et al, 1998), Biber et al. (2011) apresentaram um índice de desenvolvimento hipotético com cinco diferentes estágios, contendo frases e orações. De acordo com esse índice, os alunos de baixa proficiência começam a escrever utilizando características mais comumente encontradas na fala (por exemplo, orações adverbiais), e, à medida que sua proficiência aumenta, eles passam a introduzir também mais traços frasais (por exemplo, substantivos como pré-modificadores), que são mais comuns na escrita. Até onde sabemos, não há pesquisa que tenha investigado longitudinalmente os estágios de desenvolvimento em textos produzidos por aprendizes brasileiros de inglês. Queiroz (2018), por exemplo, estudou transversalmente sintagmas nominais complexos (SNs) em redações argumentativas de estudantes universitários brasileiros. Desse modo, para preencher essa lacuna, esta pesquisa tem como objetivo captar o desenvolvimento da complexidade gramatical dos alunos brasileiros do IFA, de acordo com o referencial proposto por Biber et al. (2011). Para tanto, foi coletado um subcorpus longitudinal do CorIFA (Corpus do Inglês para Fins Acadêmicos). O subcorpus contém textos escritos por 13 alunos diferentes (n = 13) que frequentaram os cursos do IFA por um ano e meio. Cada aluno escreveu três textos com seis meses de intervalo; portanto, esta pesquisa analisa três pontos no tempo. Os resultados do estudo longitudinal demonstram que nem todos os elementos do índice apresentaram significância estatística ao longo do tempo, mas a maioria dos elementos frasais que se previa o aumento da frequência tiveram um resultado positivo do Tempo 1 para o Tempo 3 (por exemplo, o aumento estatisticamente significativo de adjetivos atributivos). No entanto, ainda que não estatisticamente, alguns tipos de orações, ao invés de decrescerem, apresentaram um pequeno aumento ao longo do tempo (por exemplo, orações adverbiais). Além disso, uma análise quasi-longitudinal sobre os registros e as divisões acadêmicas dos alunos mostra diferenças na preferência por características muitas vezes relacionadas às finalidades comunicativas dos registros e das áreas. Por fim, uma comparação com resultados do estudo de Staples et al. (2016) sobre L1 demonstra que o desenvolvimento de aprendizes brasileiros e nativos não é semelhante, principalmente no que diz respeito à utilização de orações, pois os

nativos diminuíram no uso de tais construções, o que não aconteceu para todos os tipos de orações em nosso subcorpus.

Palavras-chave: complexidade gramatical, estudo longitudinal, estudo quasi-longitudinal, corpus de aprendiz

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Attributive adjective
ABs	Abstract
AC	Finite adverbial clause
AD	Adverbs as adverbials
AEs	Argumentative Essay
BAWE	British Academic Written English Corpus
BHS	Biological and Health Sciences
BNC	British National Corpus
CL	Corpus linguistics
COCA	Corpus of Contemporary American English
COEP	Comitê de Ética em Pesquisa
CorIFA	Corpus de Inglês para Fins Acadêmicos
E	Edited
EAP	English for Academic Purposes
FCC	Finite complement clause
HA	Humanities and Arts
IFA	Inglês para Fins Acadêmicos
IQR	Interquartile range
L1	First language
L2	Second Language
LONGDALE	Longitudinal Database of Learner English
MD	Multidimensional
NE	Non-edited
NFC	Non-finite complement clause
NFTC	Non-finite to complement clause
NP	Noun phrase
OP	Of phrases
PP	Prepositional phrase
PSE	Physical Sciences and Engineering
RC	Relative clause
SD	Standard deviation

SOP	Statement of Purpose
SSE	Social Sciences and Education
SUM	Summary
TRC	That-relative clause
UFMG	Universidade Federal de Minas Gerais
WRC	Wh-relative clause

LIST OF FIGURES

LIST OF GRAPHS AND CHARTS

LIST OF TABLES

SUMMARY

1 – INTRODUCTION

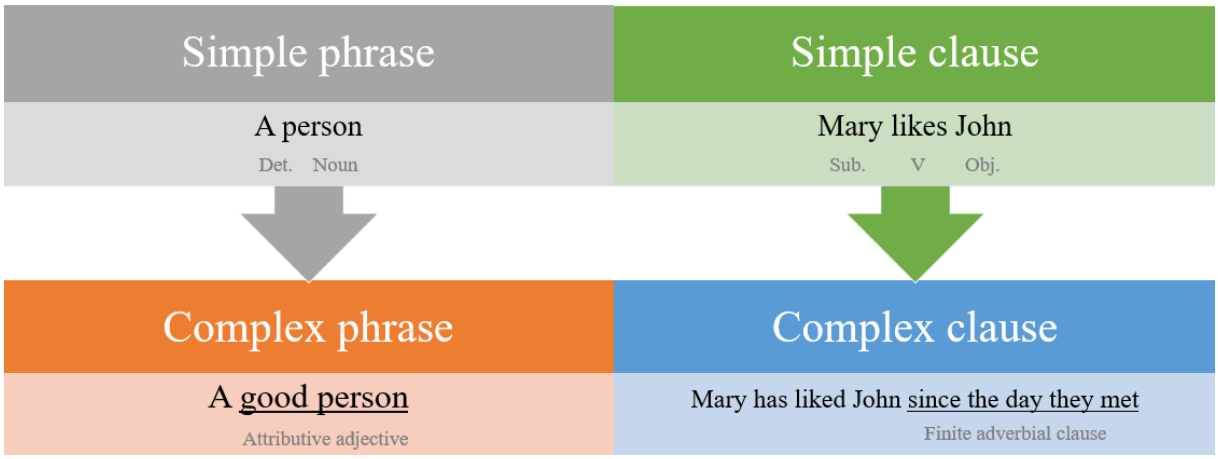
Since its emergence in modern linguistics, corpus linguistics (CL) has taken a significant role in language description studies, with the primary aim to document authentic language production. For example, studies on the variations across writing and speech (BIBER et al., 1999; BIBER, 1991) have gone beyond this purpose and, besides presenting authentic language production, also revealed characteristics that had not been described before, such as the use of lexical bundles across registers (BIBER et al., 1999). CL can be used in several different studies to fit various research purposes, including register variation (e.g., BIBER, 1988; 2012), learner corpus (e.g., GRANGER, 1998), and academic writing (RÖMER et al., 2020), among others.

Academic writing is one of the most analyzed registers in corpus-based studies, due to its uniqueness and specificity. Often seen as "deliberately complex, and more concerned with impressing readers than communicating ideas" (BIBER; GRAY, 2016, p. 1), such register may pose a challenge for both first (L1) and second language (L2) writers, requiring explicit instruction and several years of practice. Many studies addressing the specific features of academic writing focused on complex noun phrases, which are formed by nouns and their pre and/or postmodifiers and are mostly found in academic writing excerpts (BIBER; GRAY, 2016). For example, Parkinson and Musgrave (2014) assessed noun phrase complexity in English for Academic Purposes (EAP) texts; Krayer and Schaub (2018) investigated the development of noun phrase complexity by German intermediate learners of English longitudinally; Queiroz (2019) examined noun phrase complexity in essays in a corpus of Brazilian learners; Dutra et al. (2020) discussed premodification with adjectives in Applied Linguistics and Chemistry papers, and Mattos (2020) explored hyphenated premodifiers in a specialized corpus of Biology research articles. Moreover, differences in trends within academic writing registers and/or students' disciplinary fields have been found, such as in Staples et al. (2016), and Biber et al. (2020), thus showing the importance of extensive analysis in this specialized form of writing.

When dealing with the academic register, we come across some issues, such as how one can measure complexity development in academic writing. In this thesis, we adopt the concept of grammatical complexity as defined by Biber et al. (2020, p. 5) "as the addition of structural elements to 'simple' phrases and clauses." For example, a simple phrase refers to the headword,

and a simple clause is formed by the subject, the verb, and the object. If we add more information to these structures, they become more complex. Figure 1.1 below illustrates this definition.

Figure 1.1 – Simple and complex phrases and clauses



Source: Prepared by the author, 2022.

The analysis of text complexity has relied on different measures. Clausal subordination measures, such as the T-unit[1], have been used for a long time, for both speech and writing registers (Larsen-Freeman, 2006; Nelson & Van Meter, 2007, *apud* BIBER et al., 2011). Thanks to corpora-based research, register structural/syntactic differences, specifically between conversation and academic writing, were identified, showing that the "kinds of complexity common in academic writing are fundamentally different from the kinds of complexity common in conversation." (BIBER et al., 2011, p. 29). As academic writing complexity relies on phrases, whereas conversation complexity relies on clauses, the use of T-units is not an appropriate perspective to measure complexity in different registers.

With this in mind, Biber et al. (2011), created a hypothesized index with the most common grammatical features found in academic writing texts through a corpus-based analysis. This index presents features divided by stages, representing students' line of progression, and complexity as a multidimensional construct. This means that students may start by applying typical features mostly found in speech, such as finite complement clauses, and evolve to features more commonly found in academic writing, such as nouns as premodifiers (which were also approached in Biber and Gray, 2016). Even though the index has already been implemented

[1] Defined as "a main clause and all associated dependent clauses." (BIBER et al., 2011, p. 7).

in many studies, it has never been analyzed in texts of Brazilian learners of English. This issue will be fully discussed in chapter 2.

If we want to better understand and help learners to improve their writing, it is paramount to investigate how their academic writing improves over time, so that they allegedly become more proficient writers (BIBER et al., 2011). Hence, this study aims at investigating the development of academic writing by Brazilian learners of English, through the analysis of grammatical complexity features present in a subcorpus compiled from the Corpus of English for Academic Purposes (*Corpus de Inglês para Fins Acadêmicos*, CorIFA). Such analysis seems to be of crucial relevance to understanding learners' writing development, especially because, to the best of our knowledge, no longitudinal corpus-based research on higher education students' writing has been conducted in Brazil.

1.1 Justification for the Research

English has been my passion since I was a little girl, even though I could not speak it back then. When I had the opportunity to join an English class, I started from the intermediate level because I had already learned how to communicate on my own. As time went by, I grew more and more passionate about this language and was even lucky enough to study abroad by taking part in an exchange program. I was only 16, and that completely changed my life. The experience helped me grow as a person, and, of course, as an L2 English speaker. I believe opportunities are extremely important, but not everybody is granted the same ones, especially here in Brazil.

Therefore, programs such as English for Academic Purposes (EAP), which are freely available in some federal universities in Brazil, provide university students with the opportunity of becoming more proficient in the English language used in the academic context. This allows them to take part in exchange programs, such as *Minas Mundi*[2] and Science without Borders[3], apply for seats at universities abroad, present papers in international conferences, and submit

[2] Minas Mundi is one of the exchange programs at *Universidade Federal de Minas Gerais* (UFMG). For more information, access the 2021-2022 call at https://www.ufmg.br/dri/wp-content/uploads/2021/09/Edital-005_2021_UNIFICADO_MOBILIDADE-INTERNACIONAL-2021-2022-modificado-pela-Errata-01.pdf

[3] *Ciências sem Fronteiras*, in Portuguese. This program was created in 2011, with the aim to exchange science, knowledge, and technology, by granting Brazilian undergraduate and graduate students scholarships to study in competitive universities abroad, and promote internationalization. For more information, access http://cienciasemfronteiras.gov.br/web/csf/o-programa.

articles to international journals. Since English is the most spoken language in the world, and the most used in scientific publications, EAP courses are overly important.

At UFMG the EAP course is called IFA (*Inglês para Fins Acadêmicos*[4]). IFA subjects, divided into groups numbered from I to V, are designed to cover the four primary skills, that is, listening, writing, reading, and speaking, at proficiency levels ranging from B1 to C1[5]. IFA's courses usually host students from several different areas, both undergraduate and graduate, and are focused on the language used in academic contexts. Their writings have been compiled to compose CorIFA.

CorIFA has already been addressed for myriad different purposes, such as identifying the most frequent academic verbs used in argumentative essays, compared to a native corpus (GUEDES, 2017); analyzing linking adverbials by contrasting them to native corpora (DUTRA et al., 2017; DUTRA et al., 2019); investigating contrastive conjunctions by comparing the results to a native corpus (SANTOS, 2008); conducting grammatical complexity analysis of argumentative essays, with a focus on noun phrases (QUEIROZ, 2019); and examining transitivity in passive *that-clause* in abstracts, by comparing the results to a Lingua Franca Corpus (ORFANÓ; NUNES, 2020).

Nonetheless, no longitudinal study has been carried out on this corpus to assess learners' writing development over time. Besides the local importance inherent to our analysis, this study should also contribute to the learner corpus field in general as longitudinal studies in this scope are scarce (except for Biber et al., 2020; Gray et al., 2019; Bestgen and Granger, 2018; Barron, 2018; Krayer and Schaub, 2018; and Huat, 2015). Hence, this study will answer vital issues concerning grammatical complexity in learners' writing over 18 months, as texts were collected at three different moments, six months apart. Among them is the question: do students develop grammatical complexity in writing over time? Section 1.3 below will address the research objectives of this project.

[4] Translated as English for Academic Purposes (EAP).

[5] CorIFA's proficiency categories are based on the Common European Framework of Reference for Languages.

1.2 Research Objectives

This study is based on a longitudinal CorIFA subcorpus and has the general objective of analyzing the grammatical complexity development of EAP students. More specifically, it aims at achieving the following objectives:

1. To assess students' development of grammatical complexity in three points over time, totaling one year and a half;
2. to compare the results of grammatical complexity development with the stages from the Hypothesized Developmental Index proposed by Biber, Gray, and Poonpon (2011);
3. to analyze the extent of variation across CorIFA registers in terms of grammatical complexity features;
4. to verify the presence of variation across students' academic divisions in terms of grammatical complexity features;
5. to compare the results found with Staples et al. (2016), which was based on first language (L1) English university students.

1.3 Research Questions

Given the objectives laid out above, we aim to answer the following research questions:

1. Does the use of grammatical complexity features among Brazilian students develop over time?
2. To what extent do the variations observed in the subcorpus comply with or differ from the hypothesized developmental index proposed by Biber, Gray, and Poonpon (2011)?
3. To what extent can variation across registers be observed in the CorIFA subcorpus when grammatical complexity features are considered?
4. To what extent can variation across disciplines be observed in the CorIFA subcorpus when grammatical complexity features are considered?
5. Is there a difference in the use of grammatical complexity features in BAWE and the CorIFA subcorpus?

1.4 Research Hypotheses

According to the research questions above, five research hypotheses were formulated:

1. Over time, students will increase the use of phrasal features from the second stage onwards and decrease the use of finite and non-finite clausal features from the first and second stages.
2. Over time, students will follow the hypothesized developmental index, thus showing an increase in the use of features from the later stages.
3. There will be variations in the use of certain features across registers.
4. There will be variations in the use of certain features across academic divisions.
5. There will be differences between the texts written by Brazilian and British university students in the scope of the development of certain complexity features.

1.5 Outline

This master's thesis is divided into five chapters, including this introduction. The next chapter is the Literature Review, which details important concepts and previous studies related to the subjects addressed herein, such as learner corpus, grammatical complexity, and longitudinal studies. The third chapter (Methodology) elucidates data collection and analysis procedures. Results and Discussion is the fourth section and focuses on the research objectives, by answering all research questions and presenting the results through extensive data analysis. Finally, the fifth and closing chapter (Conclusion), summarizes all information available in this thesis and presents a brief overview of the results found. References and appendices are presented following the final chapter.

2 – LITERATURE REVIEW

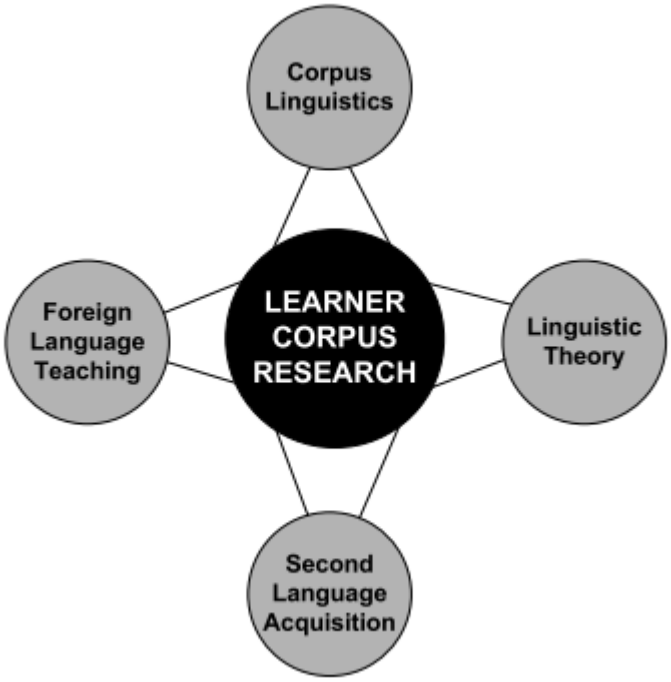
In this section, previous studies on corpus linguistics, learner corpus, grammatical complexity, and longitudinal corpus are presented and discussed, to shed light on the literature on these subjects. Sections are divided into 2.1 Learner corpus, 2.2 Grammatical complexity, and 2.3 Longitudinal studies. In addition, a paramount feature of this thesis' methodology will also be introduced in subsection 2.2.1, which is the Hypothesized Developmental Index (BIBER et al., 2011), since it is a key component of data extraction.

2.1 Learner corpus

Learner corpus, also known as "electronic collections of foreign or second language learner texts assembled according to explicit design criteria" (GRANGER, 2009, p. 2) has been of major relevance for theoretical and pedagogical implications, due to its myriad research possibilities. Since the compilation of the first learner corpus, the International Corpus of Learner English[6] (ICLE) in the 1990s, studies about/with learner corpus have gained prominence by addressing several issues and have become crucial not only for language analysis but also for the pedagogical implications that its findings encompass. Granger (2009) created a figure to better display the fields existing in the scope of learner corpus research.

[6] An updated version of this corpus was released in 2020 with over 5.5 million words (GRANGER et al., 2020).

Figure 2.1 – Learner corpus research fields



Source: Granger (2009, p. 2)

In addition to a large amount of empirical basis that learner corpora provide (and also considering the compilation criteria, as all corpora in general need to have strict criteria for their compilation), the analysis of students' texts is of paramount importance for several reasons, such as to "identify the phrases that are typically used by foreign or second language (L2) learners" (BESTGEN; GRANGER, 2018, p. 2), to "investigate how novice writers develop over time to produce such complex language." (STAPLES et al., 2016, p. 152), and to "enable[s] researchers to tackle a much wider range of topics and hence [to] bring[s] to light a much more diversified view of learner language." (GRANGER, 2009, p. 3).

For example, when carrying out contrastive interlanguage analysis[7] (GRANGER, 1996) with a native corpus, more specific learner writing characteristics are acknowledged, such as learners' preference on the use of terms not so used by native speakers (overuse); learners'

[7] A contrastive interlanguage analysis (CIA), first acknowledged by Granger (1996) is a comparative methodology designed for learner corpus research, divided between two different types of analysis: native language (E1) x learner language (E2) and learner language (E2) x learner language (E2). In the second type of approach, E2 x E2, learners can have different mother tongues (L1).

infrequent use of terms typically used by native speakers (underuse); and the inappropriate use of certain terms (misuse). Thus, learner corpus research combined with contrastive interlanguage analysis is useful to expand the research scope, which can focus on a native or another learner corpus.

A topic of investigation that suits learner corpus purposes and can be conducted by adopting the contrastive interlanguage analysis methodology or not is grammatical complexity analysis. This is true because "grammatical complexity is a multidimensional construct, with different types of complexity features serving different discourse functions, and different registers being complex in different ways" (BIBER et al., 2020, p.7). In this way, a learner's grammatical complexity analysis allows researchers to discover a wide range of features in learners' writing, thus enabling comparisons among learners' proficiency level, discipline, or register, as each variable can present distinct complexity features.

Furthermore, it is important to bear in mind the best measures that will fit the research's interest, which will be able to capture all types of grammatical complexity. Section 2.3 will present an overview of grammatical complexity focused on academic writing, as well as a description of investigations already performed, which included the use of learner corpus on the grammatical complexity analyses.

2.2 Grammatical Complexity

According to Bulté and Housen (2012, p. 22), "there is no commonly accepted definition of complexity[8]". Therefore, it is important to clearly explain how complexity is defined and treated in this research. As mentioned earlier, the meaning of complexity adopted here will be the same as the definition for grammatical complexity coined by Biber et al. (2020), which refers to every extra element added to simple phrases or clauses. A simple phrase, in this sense, contains only a head, which is usually a noun; and a simple clause contains the combination of subject + verb + object, or complement (BIBER et al., 2020). Hence, complexity will be treated according to the features from the Hypothesized Developmental Index (BIBER et al., 2011, p. 30-31), which will be addressed below.

Grammatical complexity has been a topic of analysis for a long time, most specifically since the 1930s, when the first studies on L1 writing began to be carried out, such as Frogner

[8] Emphasis added.

(1933), LaBrant (1933), and Anderson (1937) (*apud* BIBER et al., 2011). These focused mainly on primary and secondary school students, but from the late 1960s onwards, the focus shifted to college students (e.g., HUNT, 1970; JAKOBOVITS, 1969, *apud* BIBER et al., 2011). At the same time, scholars became interested in grammatical complexity in L2 writing and started analyzing students' writing development, such as Cooper (1976) and Ferris and Politzer (1981), and this interest has remained to this day (*apud* BIBER et al., 2011).

At that time, most studies on grammatical complexity in writing relied on clausal subordination, which was described by scholars such as Wolfe-Quintero et al. (1998) as the best measure to analyze complexity. Indeed, scholars sincerely believed that "longer units and more subordination reflect[ed] greater complexity." (*apud* BIBER et al., 2011, p. 7). This resulted in a profusion of works examining both first (L1) and second language (L2) writing based on the T-unit, which refers to the main clause and all its dependent clauses (Hunt, 1965 *apud* BIBER et al., 2011, p. 7). This unit has prevailed to this day (indeed, Ho-Peng, 1983, analyzed writing proficiency of university ESL students; Beers and Nagy, 2007, examined words per clause and clauses per T-unit in two genres produced by middle school students; and Knoch et al., 2015, analyzed grammatical complexity in texts produced by 31 advanced university students longitudinally). In the scope of those investigations, two T-unit measures were applied most frequently: The MLTU, which measures the mean length of the T-unit), and the C/TU, which measures the number of clauses per T-unit, averaged by all T-units in a text (BIBER et al., 2011).

Meanwhile, empirical corpus-based studies challenged the idea that T-units were the ideal measure for assessing grammatical complexity in writing (BARDOVI-HARLIG, 1992, *apud* BIBER et al., 2011). This was primarily the case of studies contrasting oral and written language extracts, most specifically academic writing, since they are "produced in circumstances where language is carefully planned and edited, detailed and specific, and produced in a concise format." (STAPLES et al., 2016, p. 151). Thus, empirical studies found that while clausal subordination is a typical feature in speech, noun phrases (NPs) are more important and better predictors of writing quality in academic writing (e.g., BIBER, 1985; 1986; BIBER et al., 2011; KYLE; CROSSLEY, 2018).

2.2.1 Hypothesized Developmental Index (BIBER et. al., 2011)

Focusing on this discovery about clausal and phrasal features, Biber et al. (2011) proposed a hypothesized developmental index with five (5) stages, comprising both L1 and L2

writing. Contrary to other studies based solely on theoretical grounds, Biber et al. (2011) is an empirical study of authentic language extracts. According to this hypothetical index, students' academic texts incorporate typical features of speech at the first stage, such as "finite complement clauses (*that* and *WH*) controlled by extremely common verbs (e.g., think, know, say)" (p. 30). Later, at the second stage, they evolve to the use of "noun modification features starting with simple modification through attributive adjectives and participle pre-modifier" (ANSARIFAR et al., 2018, p. 60).

From the third stage on, students begin to use more complex structures, such as nouns as premodifiers and prepositional phrases. In the fourth stage, they are expected to incorporate nouns modified by non-finite clauses into their writing, as well as prepositional phrases as noun postmodifiers. In the last stage, students are expected to use overly complex structures, such as "appositive noun phrases and complement clauses as noun modifiers in addition to multiple phrasal embeddings'" (ANSARIFAR et al., 2018, p. 60). An expanded model of the developmental index is presented below.

Table 2.1 – Hypothesized Developmental Stages for Complexity Features

Stage	**Feature**	**Examples**
1	Finite complement clauses (that and WH) controlled by extremely common verbs (e.g., *think, know, say*)	we never quite **know** what to make of him (conv)
2	Finite complement clauses controlled by a wider set of verbs	I'd **forgotten** that he had just testified on that one (conv)
	Finite adverbial clauses	If you're sitting next to me and you want ninety degrees, and I want sixty degrees, we're just gonna be battling each other… (conv)
	Non-finite complement clauses, controlled by common verbs (especially *want*)	I don't **want** to fight with them about it (conv)

	Phrasal embedding in the clause: adverbs as adverbials	He's so confused anyway (conv)
	Simple phrasal embedding in the noun phrase: attributive adjectives	It certainly has a nice flavor (conv)
3	Phrasal embedding in the clause: prepositional phrases as adverbials	He seems to have been hit on the head (fict)
	Finite complement clauses controlled by adjectives	It seemed quite **clear** that no one was at home (fict)
	Non-finite complement clauses controlled by a wider set of verbs	The snow began to fall again (fict)
	That-relative clauses, especially with animate head nouns	…the guy that made that call (fict)
	Simple phrasal embedding in noun phrases: nouns as premodifiers	…some really obscure cable channel (fict)
	Possessive nouns as premodifiers	Tobie's voice (fict)
	Of phrases as postmodifiers	**editor** of the food section (fict)
	Simple PPs as postmodifiers, especially with prepositions other than *of* when they have concrete/locative meanings	**house** in the suburbs (fict)
4	Non-finite complement clauses controlled by adjectives	These will not be **easy** to obtain (acad)
	Extraposed complement clauses	**It is clear** that much remains to be learned… (acad)
	Non-finite relative clauses	…**the method** used here should suffice… (acad)

	More phrasal embedding in noun phrases = attributive adjectives, nouns as premodifiers	The prevalence of airway obstruction and self-reported disease status (acad)
	Simple prepositional phrases as postmodifiers, especially with prepositions other than of when they have abstract meanings	with half of the subjects in each age/ instructional condition receiving each form (acad)
5	Preposition + non-finite complement clause	The **idea of** using a Monte Carlo approach (acad)
	Complement clauses controlled by nouns	The **hypothesis** that female body weight was more variable (acad)
	Appositive noun phrases	The **CTBS** (the fourth edition of the test) was administered in 1997–1998 (acad)
	Extensive phrasal embedding in the NP: multiple prepositional phrases as postmodifiers, with levels of embedding	The [presence of layered [[structures] at the [[[borderline]] of cell territories]]] (acad)

Source: Adapted from Biber et al., (2011, p. 30-31).

This developmental index, which hypothesizes the process through which students go in academic writing as their proficiency increases, is not always observed when put into practice. For this reason, several studies have found complexity increase in phrasal features as students' proficiency level rises, but very few strictly follow the stages from the index proposed by Biber et al. (2011). The subsection below will embody this statement further.

2.2.2 Grammatical Complexity Studies

Following the proposal of Biber, Gray, and Poonpon's (2011) developmental index, grammatical complexity studies began to be conducted. Parkinson and Musgrave (2014) investigated NP complexity using the developmental index. The authors analyzed two groups, one consisting of EAP L2 international students coming from several countries, such as China, Mexico, Japan, and another formed by international L2 students enrolled in a master's program (MA) in TESOL, who had achieved scores over 6,5 in IELTS and thus had a higher level of

proficiency compared to the EAP group. Furthermore, only the features from the index related to NP complexity were considered for the analysis, such as prepositional phrases (PPs), attributive adjectives (AAs), relative clauses. Their findings corroborated the index, as the EAP group relied more frequently on AAs, a feature considered to be acquired early in academic writing, whereas the MA group used more nouns as premodifiers (NPs).

Ansarifar et al. (2018) compared abstracts from Persian learners of English from MA to Ph.D. levels and published writers from the field of Applied Linguistics. They selected NP features from the developmental index proposed by Biber et al. (2011), from the 2nd stage onwards, totaling 16 features. The results showed that the most common premodifiers were attributive adjectives and nouns, whereas prepositional phrases were the most common post-modifiers in all three corpora. However, when comparing the three corpora, four significant differences between the MA and PhD-level and the expert writers were found. Indeed, the latter group used nouns as premodifiers more frequently, along with -ed participles as post-modifiers, and adjective/noun combinations as premodifiers. In turn, prepositional phrases were more frequent in the MA-level students' texts.

Lan et al. (2019) is another interesting study that analyzed Chinese learners of English through a corpus of argumentative essays with 50 high-proficiency students' texts and 50 low-proficiency ones. They also selected only NPs of the index from the 2nd stage onwards, totaling 11 noun modifiers. Their findings show that the 11 noun modifiers represent 4.3% of the variance in L2 writing proficiency, which is a considerable figure since it is a single category. Although their study revealed a weak association between academic writing and writing proficiency for all noun modifiers in their statistical comparison, four noun modifiers contributed the most to the association, namely (1) attributive adjectives, (2) premodifying nouns, (3) relative clauses, and (4) *of* prepositional phrases. The low-proficiency group showed fewer occurrences of attributive adjectives and relative clauses, and more occurrences of premodifying nouns and prepositional phrases than expected, which is in opposition to the high-proficiency group.

Studies on the development of grammatical complexity across disciplines and genres (besides those focusing on the level of schooling) have also been conducted. Examples include Staples et al. (2016), which analyzed university-level L1 writers in the BAWE[9] corpus by

[9] The British Academic Written English Corpus (BAWE)

investigating not only phrasal but clausal features as well. Their work was based on the developmental index and other features described by Biber et al. (2014, *apud* STAPLES et al., 2016). Their findings support their hypothesis that phrasal features would increase as students' proficiency grows; however, this is not a contiguous progression, since certain such as attributive adjectives only showed improvement from level 3 to 4.

Regarding grammatical complexity variations among disciplines and genres, the authors' findings show that the development across disciplines is analogous to the results concerning one's level of schooling. This means that the phrasal features increase in complexity regardless of the discipline or field of knowledge in question, except for Social Sciences, which did not show an increase in noun + *of* phrases features. Among different genres, the results were similar to those across disciplines, as they did not contain the same number of texts for each genre analyzed. Hence, certain trends in specific genres of specific disciplines emerged. The citation below better illustrates the results:

> Explanations and Case Studies, which were found most in the Life and Physical Sciences, used more pre-modifying nouns than Essays and Critiques (…). *Of* genitives and nominalizations were used the most in Essays, which were found primarily in Arts and Humanities and Social Sciences. However, attributive adjectives were used most frequently in Case Studies, which were found most commonly in Life and Physical Sciences. (STAPLES et al., 2016, p. 169)

Mattos (2020) examined NP complexity with a focus on hyphenated premodifiers in a corpus of published Biology research articles (RA). Her findings confirm that academic writing texts are "more compressed and less explicit, grammatically and semantically" (p. 126), due to the great dependence on compression devices such as hyphenation and acronyms in biology RAs. The author also introduces interesting EAP pedagogical suggestions designed specifically for Brazilian learners of English and states, based on her findings, that it would be beneficial if teachers started to have their students work with compound and hyphenated premodifiers from level A2.

A recent study by Biber et al. (2020) also approached grammatical complexity variation across diverse levels of education and disciplines. The authors analyzed 22 university-level L2 students from different L1 backgrounds, including Greek, Vietnamese, Bengali, Russian, German, French, and Turkish. They also grounded the measurement of complexity on the hypothesized developmental index, and their findings support the indexing hypothesis, which

postulates “a decline in the use of dependent clause complexity features and an increase in the use of phrasal complexity features” (BIBER et al., 2020, p. 1) as students’ proficiency increases. Different from the other studies previously discussed in this thesis so far, Biber et al. (2020) adopted a longitudinal corpus design framing two years of students’ writings. Section 2.3 delves deeper into this type of approach, which is similar to the one to be applied herein.

2.2.3 Measuring Grammatical Complexity of Brazilian Learners of English

Only a single study has analyzed grammatical complexity among Brazilian learners of English, but it was not based on the index proposed by Biber et al. (2011). Instead, it assessed specific features that are typically associated with academic writing. Queiroz (2019) analyzed the English NP in argumentative essays written by intermediate to upper intermediate EAP Brazilian learners studying English. The author classified NPs into *simple* and *complex* categories, similarly to Longman Grammar (BIBER et al., 1999). The simple NP was formed solely by a determiner and a head noun, and the complex NP featured postmodifiers, such as prepositional phrases. The results showed that Brazilian learners use complex NPs more often than simple ones, more specifically NPs with adjectives as premodifiers and NPs with prepositional phrases as postmodifiers.

2.3 Longitudinal Studies

The employment of a longitudinal design can be considered a relatively new perspective in learner corpus research. And as the lack of longitudinal corpora may stem from the difficulty in compiling them, some scholars have favored the pseudo-longitudinal approach, which consists of a comparison of students’ texts with different proficiency levels over a given period to assess different topics. Indeed, Gotz and Mukherjee (2017) analyzed German learners of English investigating the Study Abroad variable, and Maden-Weinberger (2015) conducted a study investigating subjunctives in the Corpus of Learner German (CLEG). Another method that follows similar guidelines is the quasi-longitudinal design, which focuses on texts rather than students, contrary to a true longitudinal analysis, which basically focuses on students’ variables.

As Bestgen and Granger (2018, p. 11) point out, although pseudo-longitudinal studies can lead to results similar to those found through longitudinal design, longitudinal corpus-based studies are still necessary to assess the actual development of a given student or group of students. Gass (2013, p. 37) draws a comparison between the two designs, the pseudo-longitudinal and the longitudinal, and concludes that “(…) differences in proficiency level

(beginners instead of intermediate/advanced) or the timeframe of the data collection (a few months rather than three years) will lead to radically different results." (*Apud* BESTGEN; GRANGER, 2018).

Therefore, longitudinal designs have become increasingly popular among researchers, although not all longitudinal designs have the same purpose. The vast majority aims to analyze students' development, such as Bestgen and Granger (2018), for example, who analyzed the development of collgrams[10] through a longitudinal corpus-based analysis of French learners of English. In their study, they relied on a subcorpus extracted from the Longitudinal Database of Learner English corpus (i.e., LONGDALE), consisting of 178 argumentative essays from 89 undergraduate students from the University of Louvain, all of whom were learners of English. Students had to write two texts about the same topic, one in their freshmen year in college and the other one during their junior year. To analyze students' development on the use of collgrams, the authors compared the results found in the learner corpus to a native reference corpus (British National Corpus, BNC). The results showed that the writing of most students progressed (67%), from the freshmen to the junior year; also, there were fewer occurrences of infrequent collgrams, categorized as "non-collocational" in the junior year, compared to the freshmen year. Furthermore, the authors found out that students tend to use more creative combinations of collgrams in their junior year compared to their freshmen year, such as *ban violence*, *a smartphone*, *emotional intelligence*, *anatomically precise*, *candle-lit nights*, and *emotionless violence*, although it is unfortunate that these combinations fail to appear in the native corpus.

Another study that approached students' development longitudinally was Krayer and Schaub (2018), which investigated the development of NP complexity longitudinally, in a corpus consisting of texts written by intermediate German learners of English. In their analysis, they employed two different approaches: the first one was based on global measures of complexity, such as length and number of modifiers per 1,000 words, whereas the second one was to compare their results of NP-modification structure changes over time with the results found in Biber et al. (2011), and Parkinson and Musgrave (2014). Their findings attested that the global measures of complexity remained stable over time, presenting a non-significant increase. They also demonstrated differences in the extent of NP uses among participants, as

[10] This term, acknowledged by Bestgen and Granger (2015), refers to the junction of collocations and lexical bundles.

they did not rely on NPs as advanced learners and expert writers from those two studies. Through these findings, the authors emphasized the importance of instruction for learners to acquire these writing characteristics similar to expert and advanced writers.

In addition to works analyzing students' writing, longitudinal studies have also been employed to analyze speaking development, such as Baron (2018), which examined a spoken corpus of Irish learners of German longitudinally, to attest the extent of development of routine apologies in a group of 33 learners, in the context of "studying abroad." Questionnaires were collected twice in the same year, which means that the analysis addresses two different moments in time. The results revealed that aspects of the learners' routine apologies remained stable, such as "the high use of explicit apologies" (p. 100), as well as developments concerning the increase of decrease of more appropriate features, and non-linear developments, such as a "decrease in L2-like routines" (p. 100).

Gray et al. (2019) performed an interesting twofold analysis of the longitudinal development of grammatical complexity of spoken and written responses by EFL learners in China, in the context of the TOEFL iBT Test. They did a multidimensional analysis followed by a grammatical complexity analysis to compare the results. The analyzed features were also taken from the developmental index by Biber et al. (2011) over nine months, from Time 1 to Time 2. The results of their MD analysis showed that learners developed according to what is expected by the index, both in speaking and writing. In turn, the grammatical complexity analysis led to conflated results, which signal that learners may have still been in the initial stages of the hypothesized index.

In addition to analyzing students' development, some longitudinal studies tried to find out whether EAP writing instructions could reduce learners' lexico-grammatical errors over time, such as Crosthwaite (2018). To test the hypothesis, the author used a longitudinal corpus of L2 EAP exclusively containing essays and reports. The texts were extracted from collections of three key data points and therefore encompassed three proficiency levels. The results showed a decrease in the lexico-grammatical errors produced by the students over time. Nonetheless, the individual analysis attested the opposite; that is, students continue to produce the same errors, even after receiving instructions and feedback from their teachers. Therefore, the results are inconclusive.

Unlike the aforementioned longitudinal studies, in which all participants were university students, Chau (2015) developed a longitudinal corpus-based analysis of the development of L2-English secondary school students. The author investigated the English competence development of Malaysian secondary school students longitudinally, over 24 months, based on four different points in time. To analyze the development properly, the author selected three closed-class words, namely *that, to,* and *of*, besides also comparing the texts' length and structure. Results show that the frequency of *that* increased over time, *to* remained highly frequent, and *of* showed variation in frequency. Also, text length seemed to increase over time, along with a shift in structure, as the students' narratives started to incorporate a larger number of characters compared to their freshmen year and became more redundant according to the central idea of the narrative in question. This result shows that the language users developed their discursive skills over time.

Those papers display the extent of the longitudinal perspective, as it suits a range of different purposes, not only focusing on students' development but also teachers' instructions. The results often pinpoint variations over time, even though sometimes they may not be significant. However, there have been few corpus-based longitudinal studies so far, especially concerning grammatical complexity. The most closely related studies are the ones by Biber et al. (2020), which is a longitudinal analysis of grammatical complexity in the writing of university-level L2-English students (already presented in section 2.2), and Gray et al. (2019), which is an analysis on the longitudinal grammatical complexity and MD of TOEFL speaking and writing responses by L2-English students.

This gap is even more conspicuous when it comes to Brazilian learners of English, as there has been no corpus-based longitudinal study analyzing the grammatical complexity in texts written by Brazilian learners of English texts. Furthermore, only a single corpus-based longitudinal study was found. Lima Jr (2019) investigated the acquisition of six English vowels by Brazilian undergraduate students of English, at four different points in time. The results indicate the diversity of forms of learners' development, as some vowel contrasts[11] were created over time, whereas others were lost. The author emphasizes that all students presented individual variables, such as different backgrounds and motivations, variations in L2 exposure, and so on, which may have led to this discrepancy.

[11] Contrasts = difference.

Therefore, a longitudinal corpus-based analysis of university-level Brazilian EAP students who are learners of English is of paramount importance to fill the gap in the literature.

3 – METHODOLOGY

This section describes the material and methods used in this study. First, the accessed corpus is presented along with information about the compiled texts and the students who wrote them. Second, the parameters chosen to create the subcorpus for this masters' thesis analysis are detailed. Third, the steps that were taken to conduct the analysis, such as the subcorpus tagging, and data extraction are outlined. Since this thesis aims to analyze the grammatical complexity features found in students' texts quantitatively and qualitatively, it can be categorized as mixed-methods research (CRESWELL, 2015). This type of research integrates the quantitative and qualitative methods and combines them to make assertions about research questions.

3.1 CorIFA

As it was previously presented in Section 1.2, the subcorpus chosen for this research was compiled from *Corpus de Inglês para Fins Acadêmicos*[12] (CorIFA), a learner corpus consisting of Brazilian[13] university students, learners of English, enrolled in one of the IFA courses from UFMG. CorIFA features texts from undergraduate and graduate students. The corpus has been compiled at UFMG, with a subcorpus of a few texts from Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP)[14], since 2013.

Providing practice in the four language skills, IFA courses elect specific genres[15] to be covered according to each proficiency level (more on genre selection in Dutra et al., 2019). In the case of writing skills, six different registers have been the focus throughout the years, namely Abstract, Summary, Statement of Purpose, Argumentative Essay, Literature Review,

[12] Literally translated as Corpus of English for Academic Purposes (COEAP).

[13] CorIFA also has a subcorpus of texts written by non-Brazilian students' learners of English, such as Spanish, and French, due to exchange programs at the university, but these will not be part of this research, as they do not suit the research purposes.

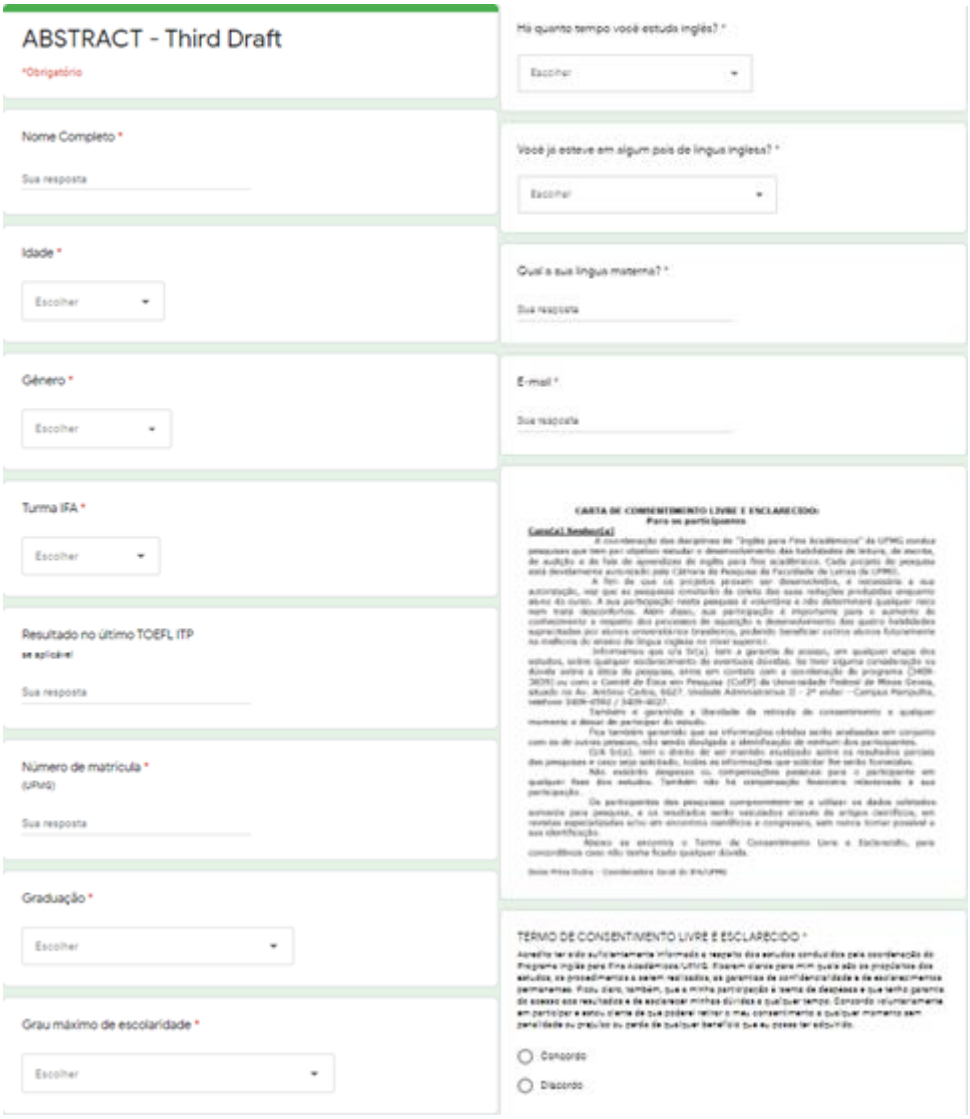
[14] Translated as Paulista State University "Júlio de Mesquita Filho".

[15] It is important to distinguish the terms register and genre, as they refer to different perspectives (BIBER; CONRAD, 2019). The register perspective, which is the perspective adopted in this thesis, "combines an analysis of linguistic characteristics that are common in a text variety with analysis of the situation of use of the variety" (BIBER;CONRAD, 2019, p. 2), whereas the genre perspective, although also including analysis of linguistics characteristics and situation of use, focuses on the conventional structures of each type of text, such as how an essay is written, from beginning to end.

and Research Article, divided into five subjects numbered from IFA I to IFA V. Furthermore, this research analysis should shed light in the way such registers have been approached in the courses. To this end, and to improve language learning and development by elaborating appropriate materials and new courses, a corpus called CorIFA was created to document students' writings (more about CorIFA in Dutra et al., 2022).

As with any long-term corpus compilation, changes were made to CorIFA's compilation process, until the process was standardized. Therefore, its texts dated from 2013 and 2014 were excluded from the corpus, due to a lack of students' metadata and text format (many texts from 2014 were delivered in handwritten form). Thus, even though its compilation began in 2013, its texts are from the second semester of 2015 onwards. The current corpus compilation follows some specific guidelines, such as requiring that all texts are delivered in electronic format and that all students must fill in a form with their metadata and agree or disagree with a consent form. Figure 3.1 below shows an example of a form used for the corpus compilation.

Figure 3.1 – Example of the form used for the corpus compilation



Source: Adapted from CorIFA, 2019.

Furthermore, to enter the corpus all texts must reach a minimum number of words according to each different register. This figure can vary from 200 words, for the Abstract register, to 1,500 words, for Research Article. The registers' selection concerns academic needs and also the student's level. Moreover, IFA instructors are free to decide which register they want to work on with their students, based on their level and needs. This also applies to topics, since CorIFA does not contain predefined topics for each different register, which, in turn, gives students the freedom to write about whatever they want, especially about their related disciplines and/or majors.

Corpus registers, students with different disciplinary fields, and texts with diverse topics enable various research possibilities, such as a comparison between texts from different disciplinary fields, which is one of the objectives of this thesis. Table 3.1 below presents IFA's subjects, their corresponding registers, and the students' proficiency level, according to the Common European Framework of Reference for Languages (COUNCIL OF EUROPE, 2001).

Table 3.1 – IFA's courses, registers, and proficiency levels

Course	**IFA I**	**IFA II**	**IFA III**	**IFA IV**	**IFA V**
Register	Statement of Purpose or Summary	Abstract	Argumentative Essay	Literature Review	Literature Review or Research Article
Proficiency level	B1	B1+	B2	B2+	C1

Source: Prepared by the author, 2022.

It is important to point out that students are required to take a language test to assess their reading, listening skills, and grammar competency before entering IFA, to place them at the right level (particularly because IFA courses starts from level B1). However, there is no test to assess whether the proficiency level of students increases throughout the term, from one course to the next. Thus, once students enter IFA, there is no way to ensure that they reached the following levels, except for their own in-class abilities, which comprise the four main skills (speaking, writing, reading, and listening).

IFA students usually write three versions of each register per semester, but only two are selected to enter the corpus: the first version, which features no comments by instructors, and the third or last version, which has feedback from the teachers. These two versions are labeled Non-edited (NE) and Edited (E) corpus texts, respectively. Nevertheless, not all students contribute to the second or last version, which results in a difference between the total number of NE and E texts. Table 3.2 shows the total number of texts per register and version, and the total number of words from the corpus.

Table 3.2 – CorIFA's total number of texts and number of words per register and version (2015-2019-1)

Register and level	**Total amount of NE texts**	**Total amount of E texts**	**Total amount of texts**	**Total amount of words**
Abstract B1+	269	232	501	112,945
Argumentative Essay B1	41	-	41	5,097
Argumentative Essay B2	233	193	426	200,543
Literature Review B2+	60	35	95	63,274
Literature Review C1	64	57	121	72,518
Research Article B2+	11	9	20	31,290
Statement of Purpose B1	214	161	375	166,462
Summary B1	49	40	89	21,086
Total	941	727	1,668	673,215

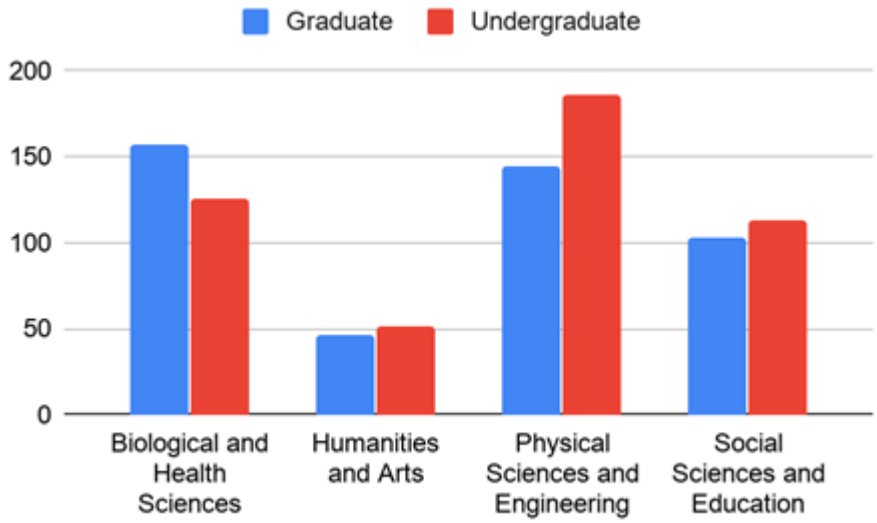
Source: Prepared by the author, 2022.

The difference between the amount of non-edited and edited versions does not interfere with the corpus purpose, that is, to assess the students' writing. In addition, the CorIFA compilation also documents and provides important metadata from the participants, under the guidelines laid out by the UFMG research ethics committee, also known as COEP[16]. Students' metadata are useful for many different types of research, such as sociolinguistic analysis,

[16] COEP is the acronym for *Comitê de Ética em Pesquisa*, translated as Research Ethics Committee. More information at: https://www.ufmg.br/bioetica/coep/

besides providing a detailed picture of the corpus content. Graph 3.1 below demonstrates CorIFA participants' level of education and their respective fields of study.

Graph 3.1 – CorIFA participants' level of education and fields of study (2015-2019)



Source: Adapted from Dutra et al. (2022).

Since all IFA subjects can be taken in two and a half years, from IFA I to IFA V, some students can start an IFA class while enrolled in their undergraduate courses and continue taking another IFA course after starting a graduate program. Moreover, other metadata available in the corpus allows the investigator to know participants' age, gender, how long they have been studying English, and if they have been to an English-speaking country before. For this reason, students' metadata analysis is essential to understand CorIFA's content as well as to expand research possibilities.

3.2 Longitudinal and quasi-longitudinal designs

Since the primary purpose of this research is to analyze the development of grammatical complexity of students over different periods, in addition to checking the presence of variations across registers and academic divisions, two designs were selected in this thesis methodology: the longitudinal and the quasi-longitudinal. According to Biber et al. (2020), the longitudinal design or "true longitudinal design" (p. 48), treats each student as an observation in a longitudinal perspective. This means that according to this method, each student is analyzed separately as the independent variable to measure the extent of variation from one point of the dependent variables (or time) to another (or others). In turn, the "quasi-longitudinal design" (BIBER et al., 2020, p. 49) treats each text as an object of observation, which suits a register or

a disciplinary variation analysis of longitudinal students' texts, for example. This means that this method does not solely consider the variation of each individual (or a group) in different periods, as it happens with a longitudinal design. It also considers the comparison of data from different individuals (samples) in the same period, such as a cross-sectional design. Thus, this design cannot be considered longitudinal, as it allows comparisons from a range of perspectives.

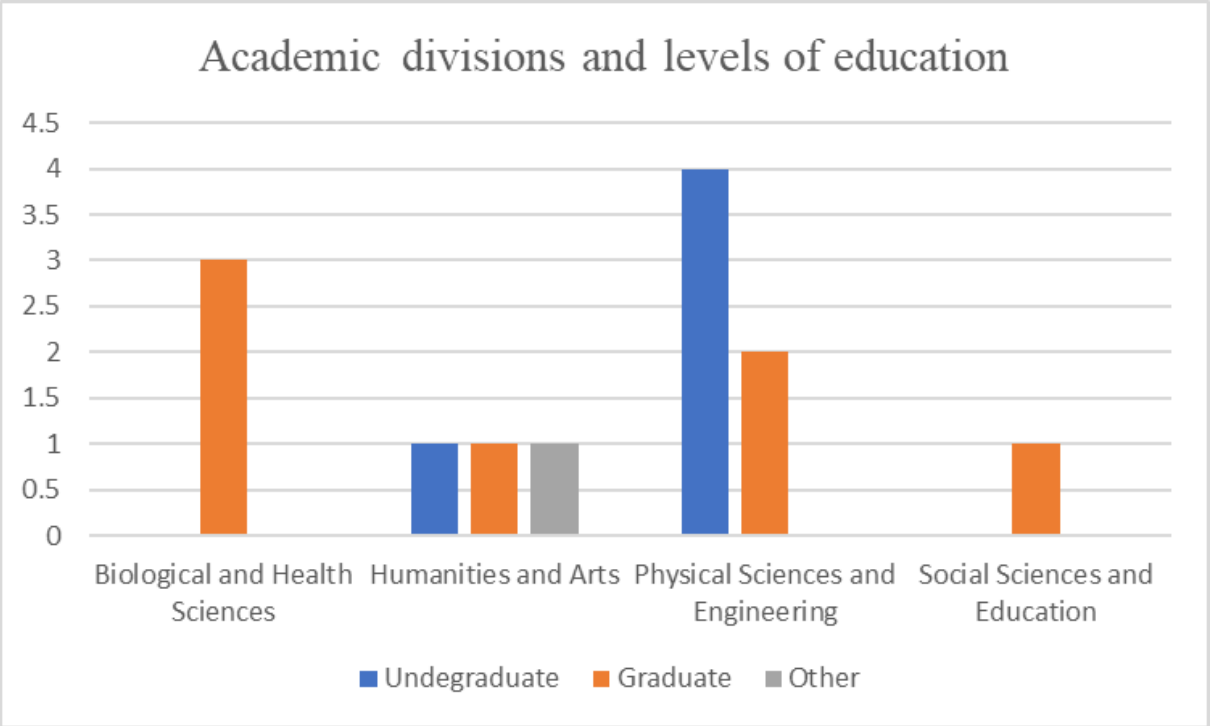
3.2.1 Longitudinal subcorpus used in the research

A subcorpus from CorIFA was compiled for the analysis to verify grammatical complexity development in learners' writing across three different moments in time, through registers and academic divisions. The information about subcorpus' participants and data are as follows.

3.2.1.1 Participants

The subcorpus contains texts from students who remained in IFA's courses for three semesters in a row (IFA I, II, and III). This parameter was set, so that a longitudinal analysis of each student could be done as well as a group analysis across register and academic divisions (quasi-longitudinal). Therefore, the subcorpus contains texts written by 13 different students (n = 13). Since an analysis across academic divisions is to be performed as well, it is important to present the related subcorpus information. Graph 3.2 below presents participants' fields of study divided by level of study.

Graph 3.2 – Subcorpus participants' academic divisions and levels of education



Source: Prepared by the author, 2022.

Graph 3.2 presents an interesting perspective, almost identical to the data presented in Graph 3.1. The number of students enrolled in the areas of Physical Sciences and Engineering is the highest in the subcorpus, as it is in the entire corpus, followed by Biological and Health Sciences. However, while in the complete corpus, Humanities and Arts is the academic division with the smallest number of students enrolled in IFA classes, Social Sciences and Education is the academic division with the smallest number of samples in the subcorpus. Therefore, the frequency of the independent variable academic division will have to be normalized for adequate data comparison and analysis. Section 3.4 carefully describes the steps that were conducted for data extraction and analysis.

3.2.1.2 Data

The subcorpus' participants wrote a total of 39 non-edited texts, from 2015 to 2018. Participants also have edited versions of texts, but only the non-edited versions were employed in the grammatical complexity analysis. This happens because the primary objective of this study is to analyze students' development, and students did not write different texts but only drafts of the same text in the same term. Table 3.3 below presents subcorpus information. All data were collected using the AntConc (ANTHONY, 2019) concordance software, and Microsoft Excel®.

Table 3.3 – Subcorpus information

Time	No. of Texts	Total No. of Types	Total No. of Tokens	Mean Words Length	Max Words Length
1	13	1,137	4,032	299.4	587
2	13	1,172	3,478	257.7	493
3	13	1,530	5,865	433	673
Total	39	2,763	13,375	330.1	673

Source: Prepared by the author, 2022.

The subcorpus has a total of 2,763 types and 13,375 tokens (Table 3.4). In addition, subcorpus texts were divided into four (4) different registers based on the six (6) registers available in the corpus. Table 3.4 below demonstrates subcorpus registers, amount of texts, and amount of tokens per register.

Table 3.4 – Subcorpus registers and amount of texts

Registers	No. of Texts	No. of Tokens
Summary	2	287
Statement of Purpose	10	3,557
Abstract	13	3,346
Argumentative Essay	14	5,667

Source: Prepared by the author, 2022.

As an analysis across academic divisions will also be performed, it is paramount to present text information divided by academic divisions. Table 3.5 below demonstrates subcorpus academic divisions, number of texts, and amount of words per division.

Table 3.5 – Subcorpus academic divisions and amount of texts

Academic Divisions	No. of Texts	No. of Tokens
Biological and Health Sciences	9	3,021
Humanities and Arts	9	2,870
Physical Sciences and Engineering	18	6,085
Social Sciences and Education	3	881

Source: Prepared by the author, 2022.

3.3 Corpus Tagging

In the following subsections, the corpus annotations such as the *Biber Tagger* and the *Complexity Tagger* software will be introduced, along with the steps taken to check and correct some specific tags semiautomatically.

3.3.1 Biber Tagger

All of CorIFA's texts were already tagged by Biber tagger (BIBER, 1988), an automatic software that loads about 129 part-of-speech (POS) tags[17] for several lexico-grammatical features, presenting morphological, syntactic, and a few semantic information (more about this software in Biber and Gray, 2013). For this step, no manual analysis for possible tag errors was performed, as texts would be tagged again by the complexity tagger (GRAY et al., 2019), and a manual analysis would be conducted then. The complexity tagger complements the tagging process as some grammatical complexity features are not tagged by the Biber Tagger. More about this software will be approached below. In Biber and Gray (2013, p. 15-18) a fine-grained explanation is presented for the measure of this tagger's accuracy as they investigate the discourse characteristics of the TOEFL iBT responses corpus. In such assessment, the majority of linguistic features were accurately identified 90% of the time, for both precision and recall. This accuracy will be considered in this thesis as well because the TOEFL corpus used in their research is similar to CorIFA, which is the corpus adopted herein. Furthermore, they analyzed writing and speech registers separately in their assessment, to avoid interference in their results.

3.3.2 Complexity Tagger

All the subcorpus texts tagged by Biber tagger were sent to Prof. Bethany Gray of the University of Iowa, who willingly accepted our request to tag the subcorpus with her complexity tagger (GRAY et al., 2019). We carried out a Zoom meeting to discuss the analysis and the features in question in more detail. This software tags almost all features from the Hypothesized Developmental index (Table 1 above) that are not tagged by the Biber tagger, such as finite complement clauses controlled by common verbs. However, it does not tag two features from stage 5, namely the appositive noun phrases, and multiple prepositional phrases as post-modifiers. Thus, those two features were not considered in this analysis, primarily because they are most commonly found in specialized informational writing as research articles,

[17] This is the amount of tags that were tagged in CorIFA. New versions may contain about 131 tags.

and since the subcorpus analyzed herein is a learner subcorpus, it contains texts of a less specialized nature.

The complexity tagger relies on the Longman Grammar of Spoken and Written English (BIBER et al., 1999) to help in the tagging process, for instance, the definition of "extremely common verbs" from stage 1, which are verbs that occurred in the finite complement clauses a 100 times per million words (Biber et al., 1999, pp. 685–686 *apud* GRAY et al., 2019, p. 8). Appendix A presents an extensive explanation of how the complexity tagger works, as well as its operational definition. Table 3.6 below presents a summarized version of this appendix, with the analyzed features and their corresponding tags.

Table 3.6 – Analyzed features and their corresponding tags

Stage	Feature	Tag
1	Finite *that*-complement clauses controlled by common verbs	vcmpth-1a
	Finite *wh*-complement clauses controlled by common verbs	vcmpwh-1a
2a	Finite *that*-complement clauses controlled by other verbs	vcmpth-2a
	Finite *wh*-complement clauses controlled by other verbs	vcmpwh-2a
2b	Finite adverbial clauses with single and multiword subordinators	fadvl-2b
2c	Non-finite *to* complement clauses controlled by common verbs	vcmpto-2c
	Non-finite *ing* complement clauses controlled by common verbs	vcmping-2c
2d	Adverbs as adverbials	adv-2d
2e	Attributive adjective as nominal premodifier	jatrb-2e
	Attributive adjectives occurring with other premodifiers	jatrb-2e4d
3a	Prepositional phrases as adverbials	ppadvl-3a
3b	Finite *that*-complement clauses controlled by adjectives	jcmpth-3b
3c	Non-finite *to*-complement clauses controlled by a wider set of verbs	vcmpto-3c
	Non-finite *ing*-complement clauses controlled by a wider set of verbs	vcmping-3c
3d	Finite *that*-relative clauses	finrel-3d
	Finite *wh*-relative clauses	finrel-3d
3e	Nouns as single nominal premodifiers	npnm-3e
	Nouns as nominal premodifiers with multiple premodifiers	npnm-3e4d
	Possessive nouns (genitive nouns) as single nominal premodifier	npnmgen-3f

	Possessive nouns (genitive nouns) as noun premodifiers with multiple premodifiers	npnmgen-3f4d
3f	*Of* phrases as postmodifiers	ppnof-3g
3g	Prepositional phrases with concrete/locative meanings	ppnc-3h
4a	Non-finite *to* complement clauses controlled by adjectives, simple (i.e., non-extraposed)	jcmpto-4a
4b	Non-finite relative clauses	nfrel-4c
4c	Relative clauses with a *wh*-relativizer	whrel-4b
4d	Prepositional phrases with abstract meanings	ppna-4e
5a	Noun + preposition + non-finite *ing*-complement clauses	ppning-5a
	Preposition + non-finite *ing*-complement clauses	ppxing-5a
5b	*That*-complement clauses controlled by nouns	ncmpth-5b
	To-complement clauses controlled by nouns	ncmpto-5b
	Wh-complement clauses controlled by nouns	ncmpwh-5b
	Ing-complement clauses controlled by nouns	ncmping-5b

Source: Adapted from Gray et al. (2019, p. 39-44).

After texts were tagged by the complexity tagger, some features, such as the nominal premodifiers, *that* complement clauses, and prepositional phrases had to be semiautomatically checked and corrected using FixTag software. Then, the texts were forwarded to Prof. Bethany Gray again, for the generation of the final tag count. Appendix B presents a broad explanation of the tags' correction process of all these features. In this process, spelling mistakes in students' texts were also corrected, as they could interfere in the tagging process, thus leading to tagging errors.

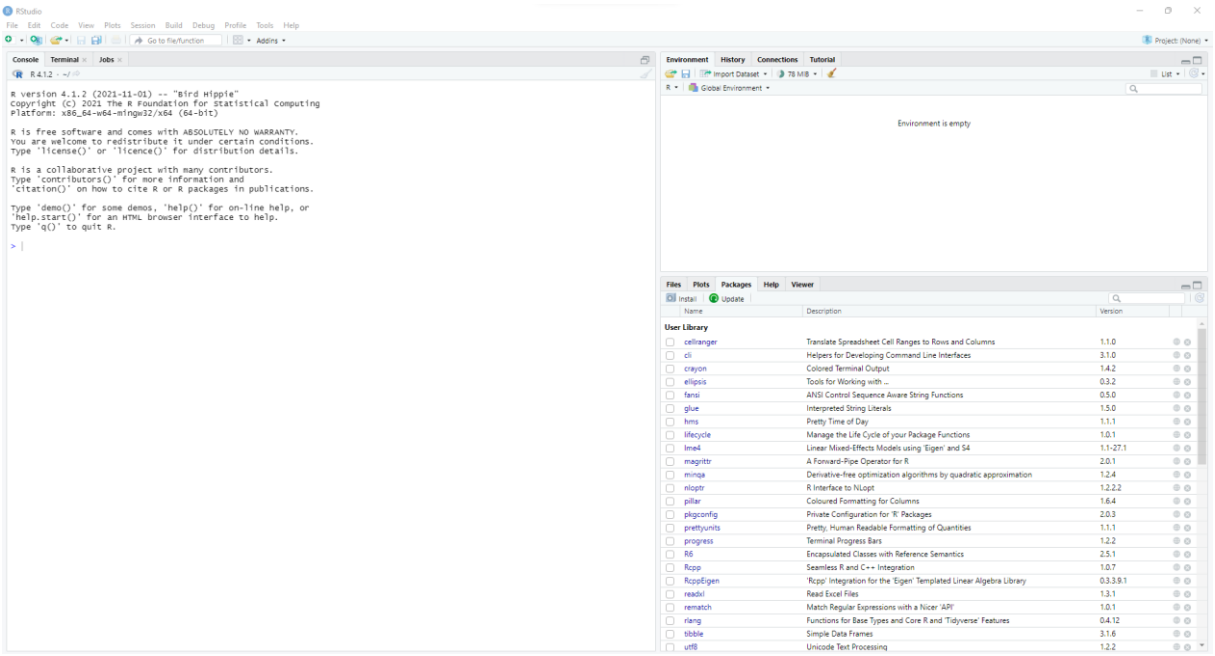
Besides tagging all texts, the Complexity Tagger, like the Biber Tagger, creates a .xml file format with the tag counts of every feature from each text, in its raw and normalized frequency. Nonetheless, since the subcorpus went through some changes after texts had already been tagged by the Complexity Tagger (such as an exclusion of a few texts), the raw frequency generated by the Complexity Tagger had to be manually normalized by 1,000 words, using Microsoft Excel, to fit the subcorpus total number of words.

3.4 Data extraction and analysis

The steps regarding data extraction and analysis of the true longitudinal design and the quasi-longitudinal design of this research will be described in this section.

As previously described in subsection 3.3.2, raw results of each complexity feature were normalized by 1,000 words and saved as spreadsheets with text information, such as student's number, time, register, and academic division. For the longitudinal quantitative analysis, the students' texts were combined into each corresponding point in time, namely Time 1, Time 2, and Time 3. To carry out the statistical analysis, the Ph.D. Economist Bárbara Dias assisted us conduct some tests. Both software R[18] and RStudio[19] had to be installed on the computer. Figure 3.2 below presents a screenshot of the software RStudio.

Figure 3.2 – Screenshot of RStudio



Source: Prepared by the author, 2022.

The selected test for comparing across points in time was the One-way Anova, as there were three samples (times), and Cohen's *f* was also used for the analysis of effect size. Text information was saved in a spreadsheet in the .csv file format, with a column for the students' number, a column for the moment in time when the text was written, and several columns for the normalized results of each feature. Table 3.6 above presents all features that were considered in the analysis. Figure 3.3 below illustrates this spreadsheet organization.

[18] https://www.r-project.org/
[19] https://www.rstudio.com/

Figure 3.3 – Screenshot of longitudinal data spreadsheet

	A	B	C	D	E	F	G	H
1	Student	Time	Feature_1	Feature_2a	Feature_2b	Feature_2c	Feature_2d	Feature_2e
2	331	1	0.07476636	0.07476636	0.074766355	0	0.14953271	0.14953271
3	516	1	0	0.14953271	0	0	0.07476636	0.37383178
4	521	1	0	0	0.074766355	0	0.44859813	1.86915888
5	525	1	0	0	0.074766355	0	0.07476636	1.04672897
6	537	1	0	0.14953271	0	0.07476636	0.37383178	0.97196262
7	563	1	0.07476636	0.07476636	0.224299065	0.07476636	0.44859813	1.34579439
8	579	1	0.07476636	0	0.224299065	0.07476636	0.29906542	1.71962617
9	713	1	0.07476636	0.22429907	0.14953271	0	0.37383178	1.64485981
10	726	1	0	0.22429907	0.224299065	0.22429907	0.37383178	2.31775701
11	774	1	0	0.07476636	0.14953271	0.14953271	0.29906542	1.4953271
12	972	1	0	0.14953271	0.074766355	0.14953271	0.29906542	1.04672897
13	973	1	0.22429907	0.14953271	0.14953271	0.14953271	0.07476636	0.6728972
14	974	1	0	0.14953271	0.074766355	0.22429907	0.22429907	1.42056075
15	331	2	0	0	0	0	0	0
16	516	2	0.07476636	0	0	0	0.22429907	1.42056075
17	521	2	0	0	0	0	0	1.4953271
18	525	2	0	0	0.074766355	0	0.14953271	1.4953271
19	537	2	0	0.07476636	0	0	0.22429907	0.29906542
20	563	2	0	0	0.14953271	0	0	0.6728972
21	579	2	0.14953271	0.22429907	0	0	0.37383178	1.42056075
22	713	2	0.22429907	0.07476636	0.074766355	0	0.14953271	2.39252336
23	726	2	0.07476636	0.07476636	0	0	0.14953271	2.99065421
24	774	2	0	0	0.224299065	0	0.74766355	1.42056075
25	972	2	0	0	0.14953271	0.14953271	0.37383178	1.42056075
26	973	2	0	0.07476636	0.074766355	0	0.07476636	0.22429907
27	974	2	0.07476636	0.07476636	0.14953271	0.22429907	0.52336449	1.12149533
28	331	3	0.07476636	0.29906542	0.074766355	0	0.6728972	4.03738318
29	516	3	0	0.14953271	0	0	0.22429907	1.94392523
30	521	3	0.07476636	0	0.074766355	0	0.07476636	1.04672897
31	525	3	0.14953271	0.07476636	0.299065421	0	0.59813084	3.28971963
32	537	3	0.14953271	0.07476636	0.224299065	0	0.52336449	0.82242991

Source: Prepared by the author, 2022.

The tests were performed for each feature using time as an independent variable since we wanted to test whether variations over time were statistically significant. After running the tests, results were saved in .txt files and then transferred to tables for the proper analysis described in Chapter 4. Other information such as mean, median, and IQR were also collected in R. Boxplots were also created comparing features divided by their structural distinctions, following Biber et al. (2011), grouping them into finite and non-finite clausal, and phrasal features per time, through the mean rate of occurrences of each feature. To this end, a new spreadsheet was designed, with a column for feature type, and columns for the mean rates of

occurrence per Times 1, 2, and 3. The packages and scripts carried out for the longitudinal analysis can be seen in Appendix C.

Figure 3.4 – Screenshot of finite and non-finite clausal features, and phrasal features spreadsheet

	A	B	C	D
1	Featurety	Time1	Time2	Time3
2	Finite	0.04	0.04	0.05
3	Finite	0.1	0.04	0.07
4	Finite	0.11	0.06	0.19
5	Non-finite	0.08	0.02	0.01
6	Phrasal	0.27	0.23	0.48
7	Phrasal	1.24	1.26	2.25
8	Phrasal	0.75	0.73	1.28
9	Finite	0	0.01	0.02
10	Non-finite	0.08	0.06	0.12
11	Finite	0.06	0.08	0.21
12	Phrasal	1	1.02	1.02
13	Phrasal	0.56	0.65	1.18
14	Phrasal	0.24	0.08	0.17
15	Non-finite	0.05	0.06	0.08
16	Non-finite	0.12	0.15	0.17
17	Finite	0.13	0.06	0.16
18	Phrasal	0.74	0.64	1.1
19	Non-finite	0.01	0.01	0.06
20	Finite	0.04	0.005	0.03

Source: Prepared by the author, 2022.

For the quasi-longitudinal analysis, the selected test was the linear mixed-effects regression. This test was also used in Biber et al. (2020), in their longitudinal analysis research of 22 L2 students. This test was chosen because our goal was to check whether register and academic division were predictors of linguistic variation, in addition to time. After all, this test can analyze the effects of predictors simultaneously. We nested students as random effects, whereas time, register, and academic division were considered fixed effects. Moreover, we selected only four features from the framework, two phrasal: attributive adjectives, and nouns as premodifiers; and two clausal: finite adverbial clauses, and finite relative clauses (*that* and *wh*). Although all features from the index were analyzed in the longitudinal analysis, we decided to analyze only these four features in the quasi-longitudinal analyses, similar to the study by Biber et al. (2020), as they “are especially important for distinguishing writing at different

levels" (p. 50), to analyze variation in registers and academic divisions in a more limited picture, due to space/time restrictions.

To run the test, a new spreadsheet was created, with columns for student, register, division, semester, and features. We also created a plot with the mean results of each text per feature, to help a better visualization of data distribution. Figure 3.5 shows a screenshot of this spreadsheet. The packages and scripts for the quasi-longitudinal analysis and the plot creation can be seen in Appendix C.

Figure 3.5 – Screenshot for the quasi-longitudinal analyses' spreadsheet

	A	B	D	F	G	H	I	J
1	Student	Register	Division	Semester	fadvl2b	frel	jatrbTotal	NPs
2	331	Aess	PSci.Eng.	Time 1	1.428571	0	2.857143	0
3	331	Abs	PSci.Eng.	Time 2	0	0	0	0
4	331	Aess	PSci.Eng.	Time 3	0.187617	1.876173	10.13133	2.439024
5	516	Summ	Hum	Time 1	0	1.273885	3.184713	6.369427
6	516	Abs	Hum	Time 2	0	1.730104	6.574394	0.692042
7	516	Aess	Hum	Time 3	0	1.079137	9.352518	3.597122
8	521	Sop	Bio.Health	Time 1	0.170358	0.340716	4.258944	6.984668
9	521	Abs	Bio.Health	Time 2	0	0.689655	6.896552	9.655172
10	521	Aess	Bio.Health	Time 3	0.4	0.8	5.6	7.6
11	525	Summ	Hum	Time 1	0.787402	0	11.02362	0.787402
12	525	Abs	Hum	Time 2	0.378788	0.378788	7.575758	1.515152
13	525	Aess	Hum	Time 3	0.655738	1.147541	7.213115	1.47541
14	537	Sop	PSci.Eng.	Time 1	0	1.612903	4.193548	4.193548
15	537	Abs	PSci.Eng.	Time 2	0	1.342282	2.684564	6.040268
16	537	Aess	PSci.Eng.	Time 3	1.094891	1.094891	4.014599	3.284672
17	563	Sop	PSci.Eng.	Time 1	1.219512	1.626016	7.317073	2.03252
18	563	Abs	PSci.Eng.	Time 2	2.083333	0	9.375	4.166667

Source: Prepared by the author, 2022.

In sum, this chapter described the details of the corpus used herein, how we prepared this study subcorpus and the methodology of analysis. The data were treated according to two methods, namely longitudinal and quasi-longitudinal analysis. Each one of them allowed us to access the data from a unique perspective. While the longitudinal study treats students as observations, with the features as dependent variables and points in time as the independent variables, the quasi-longitudinal treat texts as observations, with features as dependent variables: time, register, and academic division as independent variables. In the following chapter, the results are presented and discussed.

4 – RESULTS AND DISCUSSION

This chapter deals with the results of CorIFA subcorpus analysis, following the research objectives and questions. First, we discuss the true longitudinal design results according to the features from the developmental index (BIBER et al., 2011), and then, the quasi-longitudinal design results, divided by each variable: register and academic division. Afterward, a comparison between the results found in our subcorpus and those in Staples et al. (2016), which was based on the British Academic Written English Corpus (BAWE), will also be provided. The research questions and hypotheses are listed again as follows:

- Research questions:

1. Does the use of grammatical complexity features among Brazilian students develop over time?
2. To what extent do the variations observed in the subcorpus comply with or differ from the hypothesized developmental index proposed by Biber, Gray, and Poonpon (2011)?
3. To what extent can variation across registers be observed in the CorIFA subcorpus when grammatical complexity features are considered?
4. To what extent can variation across disciplines be observed in the CorIFA subcorpus when grammatical complexity features are considered?
5. Is there a difference in the use of grammatical complexity features in BAWE and the CorIFA subcorpus?

- Hypotheses:

1. Over time, students will increase the use of phrasal features from the second stage onwards and decrease the use of finite and non-finite clausal features from the first and second stages onwards.
2. Over time, students will follow the hypothesized developmental index, thus showing an increase in the use of features from the later stages.
3. There will be variations in the use of certain features across registers.
4. There will be variations in the use of certain features across academic divisions.

5. There will be differences between the texts written by Brazilian and British university students in the scope of the development of certain complexity features.

4.1 True longitudinal analysis

In the true longitudinal design, each student is treated as an observation (BIBER et al., 2020). To perform this analysis properly, students' texts were divided between Times 1, 2, and 3, according to the course during which they were collected, respectively IFA I, IFA II, and IFA III. Then, the mean of each feature from the Developmental Index (BIBER et al., 2011) was calculated at each time, to find out if there was variation and its extent. A One-way Anova test was then conducted, to see if variations were statistically significant, together with a Cohen's *f* to capture the effect size of the variation. Section 3.4 in Chapter 3 explains how the procedures were conducted.

Table 4.1 below presents the mean rate of occurrences (normalized per 1,000 words) of each feature per Time with the results from One-way Anova, and Cohen's *f*. Finite clausal features are marked in italic, non-finite clausal features are underlined, and phrasal features are stylized in bold. For the sake of clarification, this three-way distinction between structures is based on Biber et al. (2011) "specific structural distinctions" (p. 20-21), which considers finite clausal features as all finite dependent clauses, non-finite clausal features as all non-finite dependent clauses, and phrasal features as all dependent phrases, including adverbs.

Table 4.1 – Mean of occurrences of each feature from the Developmental Index (BIBER et al., 2011) across time, One-way Anova results, and Cohen's *f*

Features	**Mean**			***f* value**	***p* value**	**Cohen's *f***
	TIME 1	**TIME 2**	**TIME 3**			
1 – Finite complement clauses controlled by common verbs	0.04	0.04	0.05	0.44	0.50	0.11
2a – Finite complement clauses controlled by other verbs	0.10	0.04	0.07	1.17	0.28	0.18
2b – Finite adverbial clauses	0.11	0.06	0.19	2.40	0.13	0.25
2c – Non-finite complement clauses controlled by common verbs	0.08	0.02	0.01	5.70	0.02 *	0.39
2d – Adverbs as	0.27	0.23	0.48	6.26	0.01 *	0.41

adverbials						
2e – Attributive adjectives	1.24	1.26	2.25	7.42	0.009 **	0.45
3a – Prepositional phrases as adverbials	0.75	0.73	1.28	5.55	0.02 *	0.39
3b – Finite complement clauses controlled by adjectives	0	0.01	0.02	5.10	0.02 *	0.37
3c -Non-finite complement clauses controlled by a wider set of verbs	0.08	0.06	0.12	0.73	0.39	0.14
3d – That-relative clauses	0.06	0.08	0.21	10.38	0.002 **	0.53
3e – Nouns as premodifiers	1.00	1.02	1.02	0.004	0.95	0.01
3f – *Of* phrases as post-modifiers	0.56	0.65	1.18	15.99	0.0002 ***	0.66
3g – PP with concrete/locative meanings	0.24	0.08	0.17	0.83	0.36	0.15
4a – Non-finite to complement clauses controlled by adjectives + extraposed constructions	0.05	0.06	0.08	1.06	0.30	0.17
4b – Non-finite relative clauses	0.12	0.15	0.17	0.52	0.47	0.12
4c – Relative clauses with a wh-relativizer	0.13	0.06	0.16	0.33	0.56	0.10
4d – Prepositional Phrases with abstract meanings	0.74	0.64	1.10	3.48	0.06 .	0.31
5a – Preposition + non-finite complement clauses	0.01	0.01	0.06	6.95	0.01 *	0.43
5b – Complement clauses controlled by nouns	0.04	0.005	0.03	0.29	0.58	0.09

P-value signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

According to Cohen (1988), f = 0.1 is a small effect, f = 0.25 is a medium effect, and f = 0.4 is a large effect.

Source: Prepared by the author, 2022.

According to Table 4.1 above, almost half of the features showed a statistically significant difference ($p < .05$) over time. Stage 2 features, *non-finite complement clauses (NFC) controlled by other verbs* (Excerpt 4.1), presented a statistically significant decrease,

from Time 1 to Time 3, with a variation of medium effect size over time. *Adverbs as adverbials (AD)* and *attributive adjectives (AA)* (Excerpt 4.2) had a statistically significant increase over time, with a large effect size variation. However, the increase in the AD did not take place in a row, as it happened exclusively from Time 1 to Time 3, and Time 2 to Time 3. From Time 1 to Time 2, this feature decreased in frequency. This variation will be further discussed in the qualitative analysis below.

Excerpt 4.1 – NFC controlled by other verbs written by student 525 at Time 1

The objective [of] this search project **is** <u>to analyze some textbooks</u> [that] have been adopted by elementary and high school[s] of public schools.

[] – added by the author.

Excerpt 4.2 – ADs and AAs written by student 579 at Time 3

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 [to] a *full electric* age in [the] *automotive* field.

* Wrong or badly positioned word

Stage 3 features, such as the *prepositional phrases (PP) as adverbials* (Excerpt 4.3), showed a statistically significant difference over time, with a medium to almost large effect size in the variation, but, similar to the adverbs as adverbials, the PP as adverbials did not increase from Time 1 to Time 2, only from Time 1 to Time 3, and Time 2 to Time 3. *Finite complement clauses (FCC) controlled by adjectives + extraposed constructions* (Excerpt 4.4) had a statistically significant increase over time, with a medium effect size in the variation. *That relative clauses* (TRC) (Excerpt 4.5) and *of phrases (OP) as post-modifiers* (Excerpt 4.6) also had an increase from Times 1, 2, and 3, with a large effect size in the variation.

Excerpt 4.3 – PP as adverbials written by student 774 at Time 3

(...) as most lessons are charged, more attention and dedication are supposed* <u>for them</u> and consequently learning are* better, finally classroom attendance is preferred <u>by many students</u>.

*Wrong or badly positioned word.

Excerpt 4.4 – Extraposed construction written by student 973 at Time 3

(...) **it is expected** that the return provided by the fund could cover the cost of investment.

Excerpt 4.5 – TRC written by student 516 at Time 3

This way, these studies will be able to distant* from **the common place** that comprehends and construes the emergency of the Modern Sciences only over some nations, personalities or a specific period of time.

*Wrong or badly positioned word.

Excerpt 4.6 – OP as post-modifiers written by student 972 at Time 3

Among them, there are companies that work with **maintenance** of airplane or parts, aircraft assembly, research and **development** of aerospace technology and companies that work with design and **manufacture** of airplane's parts.

As for stage 4 features, only the *PP with abstract meanings* (Excerpt 4.7) presented a statistically significant difference in their mean over time. From Time 1 to Time 2, this feature decreased in frequency, but from Time 1 to Time 3, and from Time 2 to Time 3 it increased in frequency. Stage 5 feature, the *preposition + non-finite complement clauses* (Excerpt 4.8) showed a significant increase from Time 1 to Time 3, with a large effect size in the variation across time. Stage 1 feature did not present statistically significant variation across time.

Excerpt 4.7 – PP with abstract meanings written by student 579 at Time 3

For example, *the internal combustion vehicles represent 14% of the world[’s] total emission of greenhouse gases, which can represent an important **incentive** for the* electric cars production.

[] – added by the author.
*Wrong or badly positioned word.

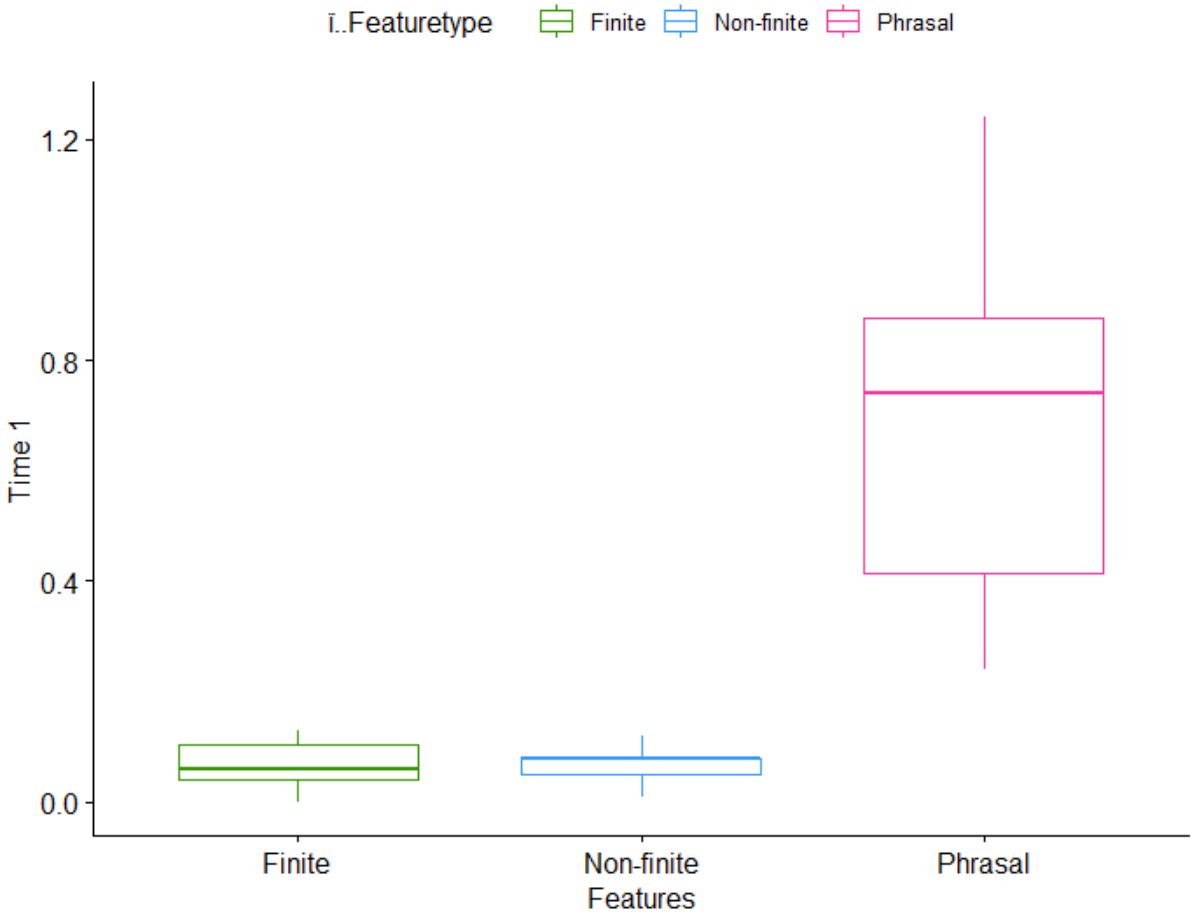
Excerpt 4.8 – Preposition + NFC written by student 774 at Time 3

Neither learning on-line exclude[s] learning in class, nor learning in class excludes the first one, for we are able to have regular class, and in advance improve our **training** by having online class and use the existent* facilities to help learning more in apps or internet sites.

[] – added by the author.
*Wrong or badly positioned word.

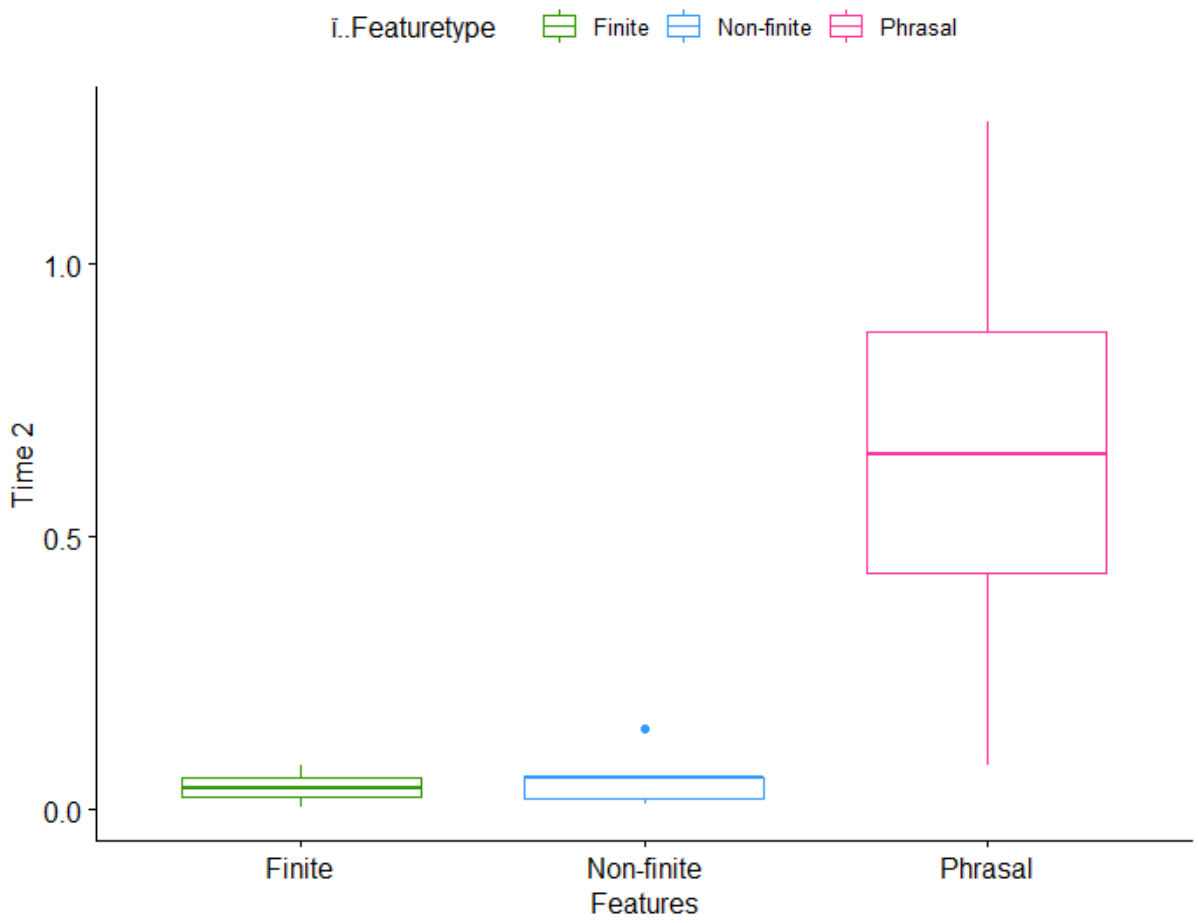
According to these results, variations in phrasal features were more statistically significant than finite and non-finite clausal features, as statistically significant variations were found in five phrasal features against two finite clausal features and two non-finite clausal features. Furthermore, considering the total variation, from Time 1 to Time 3, all features ranged as expected in the framework by Biber et al. (2011), and research hypotheses 1 and 2, as from the third stage on, all statistically significant features presented an increase, although not steadily. In all times, phrasal features were more used than finite and non-finite clausal features, as can be seen in the boxplots from Figures 4.1, 4.2, and 4.3 below.

Figure 4.1 – Boxplot of phrasal, finite, and non-finite clausal features in Time 1



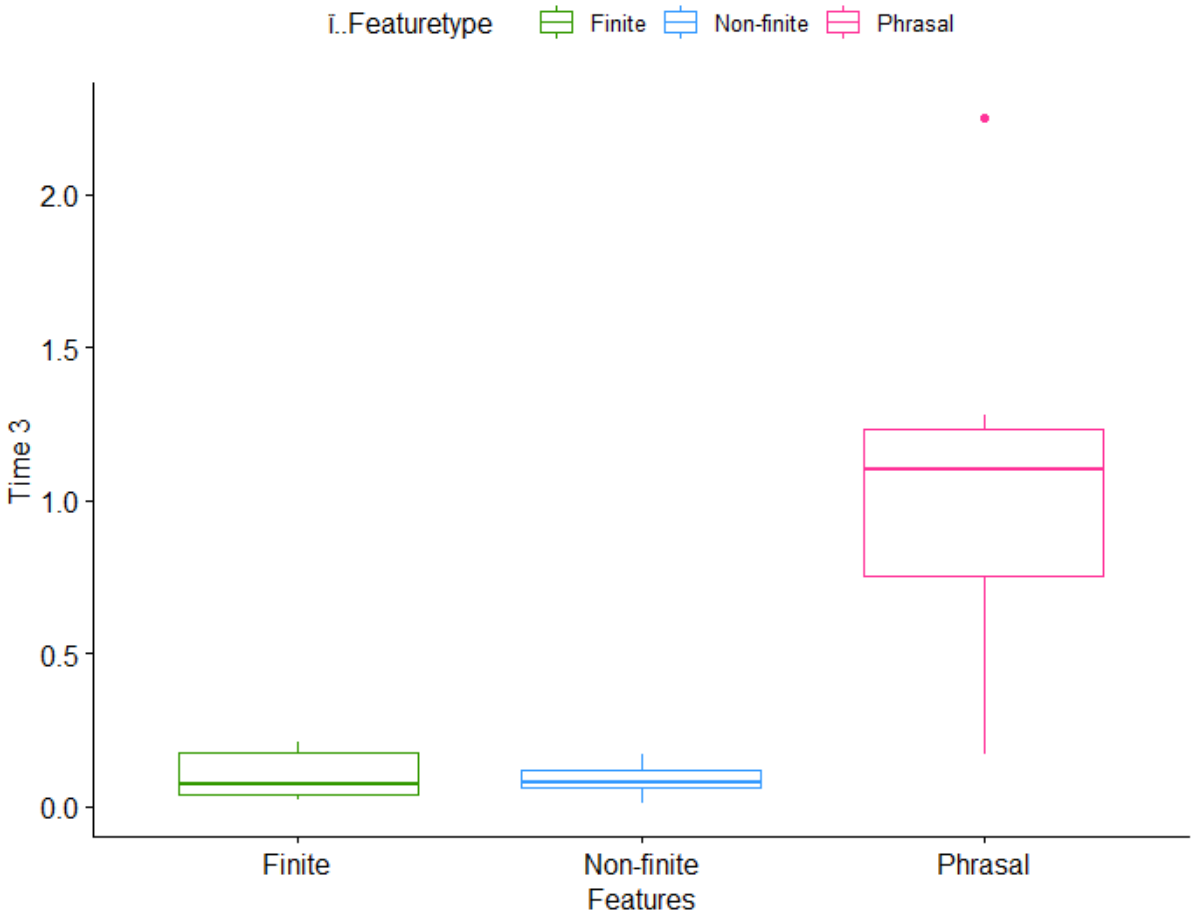
Source: Prepared by the author, 2022.

Figure 4.2 – Boxplot of phrasal, finite, and non-finite clausal features in Time 2



Source: Prepared by the author, 2022.

Figure 4.3 – Boxplot of phrasal, finite, and non-finite clausal features in Time 3



Source: Prepared by the author, 2022.

An analysis of the boxplots above shows that students rely greatly on phrasal features and seldom on clausal features since Time 1. Both the median (represented by the lines inside the squares) and the IQR[20] (represented by the squares) of phrasal features are much higher than the finite and non-finite clausal features at all times. One of the possible explanations is that they started the course already at level B1, which is an intermediate level, so they were probably able to use a variety of phrasal features in contrast to finite and non-finite clausal features.

Moreover, at all times, the values of phrasal features (represented by the extent of the plot) are more spread than the clausal features, which means that phrasal features have more diverse means, whereas clausal features have more similar means. Outliers (represented by the

[20]Interquartile range is the length of the box. The first part of the box, before the first edge, is the Q1, which indicates that 25% of all data points fall below this value. The final edge of the box is the Q3, which indicates that 75% fall below the related value. The whole box is the Q2, which is the difference between the Q1 and the Q3 (WINTER, 2020).

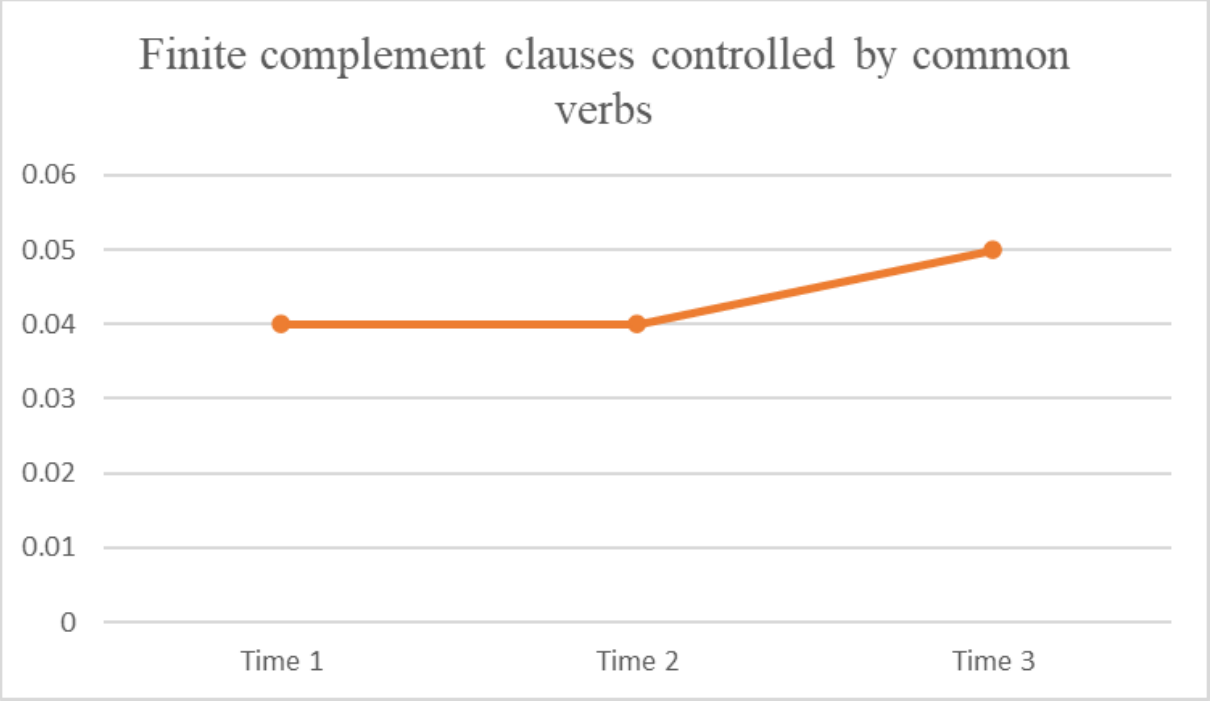
little dots) in Times 2 and 3, in Figures 4.2 and 4.3, represent the highest means in both times, and their distance to the IQR. In Time 2, there is an outlier in the plot of non-finite clausal features, which belongs to the *non-finite relative clauses*, and in Time 3, there is an outlier in the plot of phrasal features, belonging to the *attributive adjectives*, which is also the most frequent feature in the subcorpus. The fact that the corpus comparison between the mean for stage features shows more use of phrasal features, may be related to the fact that all collected texts were written by B1 students or higher.

For a better understanding of the use of features within each stage over time, a discourse qualitative analysis seems suitable to make a thorough evaluation and to capture students' variations adequately. The discussion proceeds below, divided by each complexity stage, with the focus on features that increased over time, whether statistically significant or not, from the most frequent to the least frequent.

4.1.1 Stage 1

Graph 4.1 below presents the mean rates of occurrence for stage 1 only feature, the *finite complement clauses controlled by common verbs* (Excerpts 4.9, 4.10, and 4.11 below), at Times 1, 2, and 3.

Graph 4.1 – Mean rates of occurrence of stage 1



Source: Prepared by the author, 2022

The mean of finite complement clauses (FCC) controlled by common verbs, already presented in Table 4.1, was the same in Times 1 and 2, but there was a slight increase from Time 1 to Time 3, albeit not statistically significant according to the aforementioned discussion. Even so, this result does not confirm research hypothesis H1 nor the one postulated by the Hypothesized Developmental Index (BIBER et al., 2011), as that is a feature from stage 1, generally acquired at early proficiency levels, and, therefore, expected to decrease over time as the students' proficiency increases.

The common verbs controlling FCC in Time 1 are *think, know, believe,* and *show*; in Time 2, *suggest, see,* and *show*; in Time 3, *believe, say, think, see,* and *show*. Only the communication verb *show* is similar at all times. The other common verbs can be divided into the semantic domains of mental verbs: *think, believe, know,* and *see;* speech act verbs: *say*; communication verbs: *suggest*. Although these verbs are not frequent in academic prose, they are the most frequent verbs controlling *that*-clauses (BIBER et al., 1999). Excerpts 4.9, 4.10, and 4.11 below demonstrate variations of the use of this feature over time by the same student. The common verb is marked in bold, and the finite complement clause is underlined.

Excerpt 4.9 – Time 1 FCC controlled by common verbs written by student 713

I **think** <u>that I am indicate* for this opportunity</u>.

*Wrong or badly positioned word.

Excerpt 4.10 – Time 2 FCC controlled by common verbs written by student 713

Our results **show** <u>that [the] ethanol group had a higher mortality associated with higher weight loss</u> compared to mice from the CG after A. fumigatus infection.

[] - added by the author.

Excerpt 4.11 – Time 3 FCC controlled by common verbs written by student 713

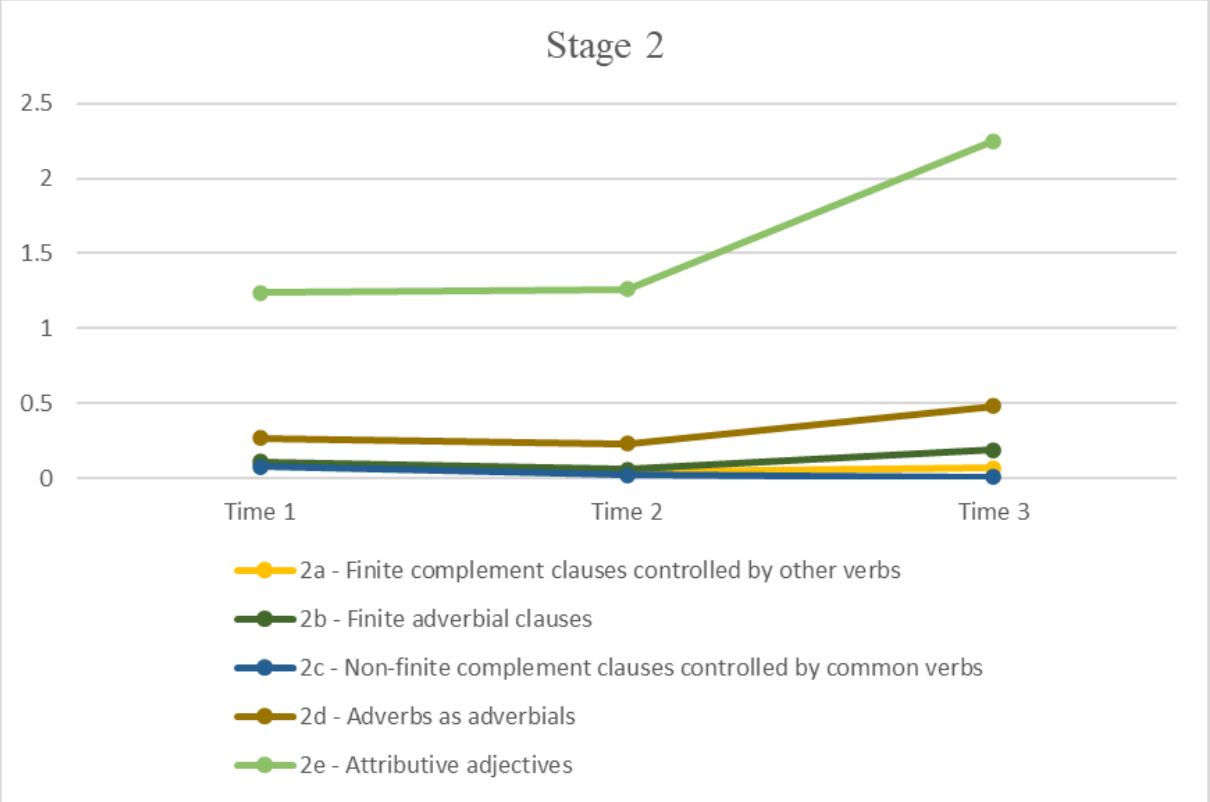
Geneticists **believe** <u>that the methods and techniques of genetics are applicable throughout the spectrum of biological activity,</u> such as cloning, genetic manipulations in embryos, plants, mammals, products, and foods.

4.1.2 Stage 2

As for stage 2 features, there was an increase in the use of *finite adverbial clauses (AC), adverbs as adverbials (AD),* and *attributive adjectives (AA),* and a decrease in the use of *finite*

complement clauses (FCC) controlled by other verbs and *non-finite complement clauses (NFC) controlled by common verbs* (Excerpts for the features are below). Graph 4.2 below shows the mean rates of occurrence of each feature inside stage 2.

Graph 4.2 – Mean rates of occurrence of stage 2



Source: Prepared by the author, 2022

Inside stage 2, only NFC controlled by common verbs, adverbs as adverbials, and attributive adjectives presented a statistically significant difference over time. Although it was a statistically significant variation, the AD feature did not follow the progression in a row, as from Time 1 to Time 2, its frequency decreased, whereas from Time 1 to 3, it increased. As for the other two features, the frequency of AA increased at each time, whereas NFC decreased.

The features that did not show statistically significant results were the FCC controlled by other verbs and the AC. Both features did not follow the progression (of decrease or increase) steadily. In the case of the FCC, the mean rate increased from Time 2 to Time 3, although having decreased from Time 1 to Time 3. On the other hand, AC's mean rate decreased from Time 1 to Time 2 but increased from Time 1 to Time 3.

The attributive adjective (AA) was the most frequent feature of stage 2 and the framework at all times. Despite increasing significantly over time, students already knew how

to use such devices since Time 1. According to the framework, AAs are acquired very early, so the fact that students highly rely on AAs, in contrast to nouns as premodifiers, indicates that they may still be in the first stages of development. This can be concluded because previous studies such as Parkinson and Musgrave (2014) have found that nouns as premodifiers are preferred by more proficient writers and are more often used in published academic prose, in contrast to AAs, which were more frequent in texts written by less proficient writers.

The occurrences of attributive adjectives can be divided into two types: AAs with a single nominal premodifier (Excerpt 4.12), and AAs with more than one premodifier (Excerpt 4.13) whether it is an adjective, noun, or genitive noun. In all times, AAs with a single premodifier were the most frequent, with occurrence rates of 72.09%, 63.95%, and 73.98% respectively, from Times 1 to 3. Excerpt 4.12 below presents an example of an AA with a single premodifier written in Time 1. AAs are marked in bold.

Excerpt 4.12 – Time 1 AA with a single premodifier written by student 974

Since the first semester I got involved in **academic** activities as a teacher of monitoring, university extension projects and **scientific** initiations.

Time 2 had the most occurrences of attributive adjectives with more than one premodifier, with 36.05%, in contrast to Time 1 (27.91%), and Time 3 (26.02%). This can be an influence of the type of register in Time 2, as Time 2 exclusively contains abstracts, which can be filled with arrangements or technical terms borrowed from the source text (whether it is a student's text or not). More about register variation will be discussed in Section 4.2 below. Excerpt 4.13 below presents an example of an attributive adjective with more than one premodifier written in Time 2. AAs are marked in bold.

Excerpt 4.13 – Time 2 AA with multiple premodifiers written by student 521

Studies with the penaeid shrimp Litopenaeus vannamei reported the ability of the species to obtain a **complete compensatory** growth after short feeding periods (...).

Excerpt 4.13 above contains two adjectives *(complete* and *compensatory)*, modifying the noun *growth*. Both adjectives are classifiers, which are the most common type of AA employed in academic prose (BIBER et al., 1999); the first one is relational and the second one is topical. As previously stated, this is a technical term or subject that may have been borrowed

from the source paper by the student, as he or she explains that the subject of the attributive adjective construction had already been reported before in other studies.

In addition, the preference for attributive adjectives shifted over time, as can be seen in Chart 4.1 below:

Chart 4.1 – Five most frequent AAs in Times 1, 2, and 3, and their relative semantic domains

	Time 1		**Time 2**		**Time 3**	
	Attributive adjective	**Semantic domain**	**Attributive adjective**	**Semantic domain**	**Attributive adjective**	**Semantic domain**
1	*electrical*	topical	*electrical*	topical	*electric*	topical
2	*new*	time	*real*	evaluative	*quality*	evaluative
3	*best*	evaluative	*equivalent*	relational	*introverted*	topical
4	*high*	size	*different*	relational	*electrical*	topical
5	*American*	affiliative	*high*	size	*mutual*	topical

Source: Prepared by the author, 2022.

According to Chart 4.1 above, the five most frequent attributive adjectives across time are extremely mixed, as there is only one AA in common in Times 1, 2, and 3, which is the topical adjective *electrical*. This may have happened because most students of the subcorpus come from the academic division of Physical Sciences and Engineering (PSE). The most frequent attributive adjective *electric* in Time 3 also illustrates this external influence on students' texts, which will be approached further in section 4.2 below. The AA of size *high* was frequent in Times 1 and 2 but not in Time 3.

The semantic domains of the five most frequent attributive adjectives are significantly diverse in Time 1, as it features five different semantic domains. Over time, the semantic domains become more stable, with a higher reliance on topical attributive adjectives. This is in line with what was expected, as topical adjectives are extremely common in academic prose, alongside relational ones (BIBER et al., 1999). Topical and relational classifiers were already found to be the most frequent AAs in Chemistry and Applied Linguistics articles according to Dutra et al. (2020), which means that our students have developed well over time, in terms of this specific feature of academic writing.

The second most employed feature of stage 2 that students increased in use over time was adverbs as adverbials (ADs). This is yet another positive and at the same time expected

outcome, as adverbials are “a relatively common feature in English” (BIBER et al., 1999, p. 766), and adverbs are a relatively common syntactic form of adverbials, only ranking behind prepositional phrases (BIBER et al., 1999). Chart 4.2 below presents the five most frequent ADs in Times 1, 2, and 3.

Chart 4.2 – Five most frequent ADs in Times 1, 2, and 3, and their relative classes

	Time 1		**Time 2**		**Time 3**	
	Adverb as adverbial	**Class**	**Adverb as adverbial**	**Class**	**Adverb as adverbial**	**Class**
1	*Also*	Circumstance adverbial	*Also*	Circumstance adverbial	*Also*	Circumstance adverbial
2	*Always*	Circumstance adverbial	*Only*	Circumstance adverbial	*However*	Linking adverbial
3	*Too*	Circumstance adverbial	*Therefore*	Linking adverbial	*Only*	Circumstance adverbial
4	*Now*	Circumstance adverbial	*Thus*	Linking adverbial	*Even*	Circumstance adverbial
5	*First*	Linking adverbial	*Even*	Circumstance adverbial	*Therefore*	Linking adverbial

Source: Prepared by the author, 2022.

By far, circumstance adverbials (Excerpts 4.14, 4.15, and 4.16, in bold) are the most frequent class of adverbials at all times, especially in Time 1, as the top four most frequent adverbs are circumstance adverbials. In Times 2 and 3, some linking adverbials (Excerpt 4.16, in italic) occupy higher positions, such as *therefore* and *thus* in Time 2, and *however*, in Time 3. In fact, the most frequent linking adverbials in academic prose are *however*, *thus*, and *therefore*, according to Longman Grammar (BIBER et al., 1999, p. 887).

Another interesting fact is that the circumstance additive adverb *also* is the most frequent at all times, which may be due to its semantic category, as it “serves to mark information being added to previous information” (BIBER et al., 1999, p. 800). This and other circumstance additive adverbs express similar relationships to linking adverbials, as they sometimes serve to contribute to cohesion in a text (BIBER et al., 1999, p. 780), which can be one of the reasons why this adverb was the most frequent.

As for linking adverbials, a thorough analysis showed that they steadily increased in frequency from Time 1 to Time 3, and also became lexically more varied. In Time 1, there are only occurrences of the linking adverbials *first*, *furthermore*, *so*, *then*, and *therefore*. In Time

2, the linking adverbials are: *finally, first, hence, however, second, then, therefore, third,* and *thus.* In Time 3 the linking adverbials are: *finally, first, furthermore, however, nevertheless, so, then, therefore,* and *thus.*

Similar to linking adverbials, stance adverbials (Excerpts 4.14 and 4.15, underlined) steadily increased in frequency from Time 1 to Time 3, and also became lexically varied over time. For example, in Time 1, there are occurrences of the stance adverbs *especially/specially, mainly, specifically, experimentally, historically, gradually, directly, subsequently,* and *probably.* In Time 2, besides presenting occurrences of some of the same stance adverbs from Time 1, there are different lexical varieties, such as *actively, certainly, highly, analytically, intensely, totally, particularly, approximately, slightly, intranasally, statistically,* and *constantly*. Different stance adverbs in Time 3 were *entirely, greatly, strongly, closely, consequently, fortunately, initially, notably, similarly, considerably, currently, essentially, generally, environmentally, heavily,* and many more.

Excerpts 4.14, 4.15, and 4.16 below demonstrate the uses of ADs by only one student over time.

Excerpt 4.14 – Time 1 ADs written by student 774

Nowadays I am graduating in English language and studying French as [a] new language, both at UFMG (Universidade Federal de Minas Gerais), Belo Horizonte, Brazil, I am easy to work with different people, and like pets a lot. I have **also** a hobby that is operating in the financial markets, <u>especially</u> in the* stock marketing.

[] – added by the author.
*Wrong or badly positioned word.

Excerpt 4.15 – Time 2 ADs written by student 774

This research was <u>slightly</u> impaired when using **only** one reader. If there were more readers, the results would <u>certainly</u> be better, but **even** with this limitation the hypothesis was valid with 70% of analysis where the creative power of the reader was proven.

Excerpt 4.16 – Time 3 ADs written by student 774

Furthermore, if we are applied* as regular students of a public university and some privates [ones] **also**, several English online course[s] may be available. Using [the] internet **also** can provide not only additional training in foreign universes, but also additional free material to help learning. In conclusion, there is no choice to be done* because the best of each method

can be used together.

[] – added by the author.
*Wrong or badly positioned word.

In stage 2, the third feature that increased in frequency over time was the finite adverbial clause (AC). This feature increased from Time 1 to Time 3, but the difference was not statistically significant. According to Longman Grammar (BIBER et al., 1999, p. 818), adverbial clauses "are used to realize time, place, manner, and contingency semantic categories.". Additionally, these semantic categories are typically marked by the presence of a subordinator, such as *when* in adverbial clauses of time.

In all times, finite adverbial clauses of time were the most frequent adverbs, with the use of subordinators *when* and *since*. It is an interesting discovery, as adverbial clauses of time are not so frequent in academic prose, but highly frequent in fiction and news (BIBER et al., 1999). In Times 1 and 2, ACs of reason were the second most frequent, marked by the subordinator *because*, followed by condition clauses, marked by the subordinator *if*, and concessive clauses, marked by *though* in Time 1, and *while* in Time 2. Time 3 revealed more frequent use of adverbial clauses, as the second most frequent type was condition clauses, marked by *if*, reason clauses, marked by *because*, and concessive clauses, marked by *although* and *though*. Excerpt 4.17 below demonstrates an example of a AC. The AC is underlined.

Excerpt 4.17 – Time 3 ACs written by student 774

For example advantages of learning online <u>when lessons can be taken anywhere</u>, lots of software and smart-phones apps available, and the communication with tutors is very easy online or by email, so distance is not an obstacle.

In general, condition and concessive clauses are more frequent in academic prose, because they "are important contributors to the development of arguments" (BIBER et al., 1999, p. 825). However, in Times 1 and 2 they were the least frequent type of adverbial clause. One of the reasons that can be influencing this result, other than the students' intermediate proficiency level, may be the registers of the subcorpus. A more in-depth discussion about register variation is presented in Section 4.2. In Time 3, condition clauses were the second most frequent, which indicates that perhaps students were starting to acquire adverbial clauses that are more frequent in academic writing.

The frequency of both non-finite complement clauses (NFC) controlled by common verbs (Excerpt 4.18) and finite complement clauses (FCC) controlled by other verbs (Excerpt 4.19) decreased over time. Nevertheless, only NFC controlled by common verbs presented a statistically significant decrease. Although the frequency of FCC controlled by other verbs decreased from Time 1 to Time 3, it increased from Time 2 to Time 3. The excerpts below show examples of these features in Time 1, which was the time with the highest number of occurrences. The clauses are underlined, and the controlling verb is in bold.

Excerpt 4.18 – Time 1 NFC controlled by common verb written by student 974

That's why your university's Evolution, Systems and Genomics program **seems** to be perfect for my academic goals.

Excerpt 4.19 – Time 1 FCC controlled by other verbs written by student 331

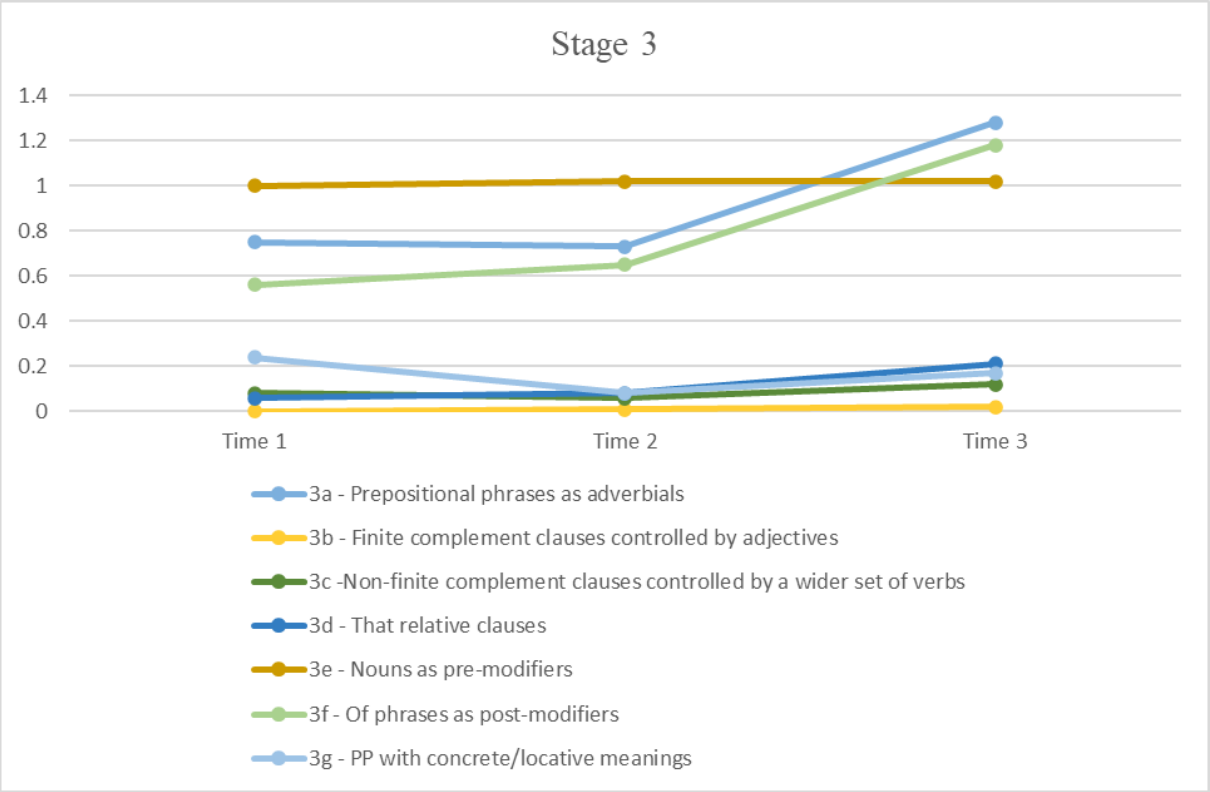
According to [the] graphs above, we can **notice** that long years ago, between 1940 - 1960, both women and men have divorced in major ages, about 50 - 55 years old.

[] – added by the author.

4.1.3 Stage 3

The frequency of stage 3 features was very diverse across each time, as can be seen in Graph 4.3 below.

Graph 4.3 – Mean rates of occurrence of stage 3



Source: Prepared by the author, 2022.

The phrasal features that presented an increase from Time 1 to Time 3 were *prepositional phrases (PP) as adverbials* (e.g., *expected by American people*), *nouns as premodifiers* (e.g., *aquaculture possibilities*), and *of phrases (OP) as post-modifiers* (e.g., *process of development*). Except for the nouns as premodifiers, PP as adverbials and OPs showed an increase not only in the mean but a statistically significant one, according to Table 4.1. Nevertheless, the increase in the frequency of PP as adverbials was not steady, as it decreased from Time 1 to Time 2. Finally, *PP with concrete/locative meanings* (e.g., *farms in Brazil*) was the only phrasal feature whose frequency decreased from Time 1 to Time 3. Also, although it increased from Time 2 to Time 3, the variation was not statistically significant.

As for the finite clausal features of stage 3, *finite complement clauses (FCC) controlled by adjectives + extraposed* (e.g., *it is clear that…*) showed a statistically significant increase from Time 1 to Time 3, along with *that-relative clauses* (e.g., *the team that was responsible for the development*). The non-finite clausal feature, *Non-finite complement clauses (NFC) controlled by a wider set of verbs* (e.g., *I've decided to try*) showed a non-statistically significant variation, which increased from Time 1 to Time 3 but decreased from Time 1 to Time 2.

The first phrasal feature of this stage (PP as adverbials) was the second most frequent feature of stage 3 in Times 1 and 2, and the most frequently used in Time 3. However, as stated earlier, this feature decreased from Time 1 to Time 2, although it increased from Time 1 to Time 3. Not only are PPs the most frequent type of post-modifiers but "prepositional phrases are the most common syntactic realization of adverbials" (BIBER et al., 1999, p. 768). This may be one of the key factors for the high frequency of this feature, even though it showed a somewhat unexpected decrease from Time 1 to Time 2.

In addition, circumstance adverbials were the most frequent type of adverbial realized by PPs at all times, followed by linking adverbials and stance adverbials. Examples of such types can be seen below in Excerpt 4.20. Circumstance adverbials are underlined, linking adverbials are in bold, and stance adverbials are in italic.

Excerpt 4.20 – Time 3 PP as adverbials written by student 525

The beginning of the XX century was characterized, specially <u>in western society</u>, <u>for a strong changing</u>* in terms of personal and social interaction. *From this context*, interactions became over stimulated*. **As a result**, a large part of [the] world's population-introversion people-has been seen <u>in a new perspective</u>. In this essay, I will discuss, *in a brief way*, the introversion phenomenon and how important [it] is to know <u>about it</u>.

[] – added by the author.
*Wrong or badly positioned word.

Excerpt 4.20 was selected due to students' heavy reliance on PP as adverbials in such a short paragraph. The text of the excerpt displayed the greatest number of occurrences of PP as adverbials in Time 3. This student displays a wide range of uses of distinct types of PP as adverbials, totaling four circumstance adverbials, two stance adverbials, and one linking adverbial. Contrary to the results found in the complete subcorpus, where linking adverbials proved to be more frequent than stance adverbials, this student, in particular, employed more stance adverbials than linking adverbials, with a similar purpose; that is, to connect different parts of the text. This is the case with the stance adverbials *from this context*, and *in a brief way* of Excerpt 4.20 above.

Nouns as premodifiers were the stage 3 feature with the highest number of occurrences in Times 1 and 2, but ranked third in Time 3, behind PP as adverbials and OPs. Besides, this feature increased from Time 1 to Time 2, but its mean was the same in Time 2 and Time 3. This result indicates that students may have yet to become proficient in this feature. Indeed,

according to the framework (BIBER et al., 2011), and Parkinson and Musgrave (2014), less proficient students tend to rely more frequently on attributive adjectives, whereas more proficient students employ more nouns as premodifiers and nouns, and its pre- and post-modifiers "express a wide range of meaning-relationships in a succinct form" (BIBER et al., 1999, p. 589). Therefore, it is more frequent in specialized academic writing.

A careful analysis of nouns as premodifiers in the students' texts shows many repetitive combinations at all times, especially Time 3, which shows the highest frequency of repetitive combinations. This signals that many structures are influenced by the discipline or topic, regardless of proficiency level, as they occur in the same text or texts written by the same student. For example, some combinations of nouns as premodifiers with over two occurrences in Time 1 were *computer science, exchange programs, graduate program, transmission lines,* and *wind tunnel*; in Time 2, they were *energy conservation, ethanol group, fossil fuels,* and *manager characteristics.*

Time 3 showed structures such as *biofloc technology, contract phase, customer requirements, food restriction, ICE vehicles, labor market, penaeid shrimp, production costs, quality manager, quotas system, voltage collapse,* and *voltage instability.* Moreover, repetitive combinations (over two occurrences) of nouns with more than one premodifier were found only in Times 2 (e.g., *digestive enzymes activities, voltage stability indexes*) and 3 (e.g., *internal combustion engine, internal combustion vehicles,* and *combustion engine vehicles*). The noun phrase that has a noun as a premodifier compresses information in such a way that by reading it, we can assume what topic or field of knowledge is being discussed. This is addressed in more detail in Section 4.2 below.

The third phrasal feature from stage 3, OPs as post-modifiers, showed a statistically significant increase from Time 1 to Time 3. Perhaps one of the reasons why the frequency of OPs as post-modifiers increased is because "prepositional phrases are by far the most common type of post-modification; (...) relatively rare in conversation [and] extremely common in academic prose" (BIBER et al., 1999, p. 606). Moreover, OPs are the most common type of PP as a post-modifier, due to its wide range of functions, particularly the expression of "a close semantic relationship between the head noun and the following noun phrase" (BIBER et al., 1999, p. 636). Excerpts 4.20, 4.21, and 4.22 below show the development in the use of OPs as post-modifiers by the same student over the years. Nouns are underlined and OPs are in bold. Some nouns inside OPs are underlined and in bold.

Excerpt 4.21 – Time 1 OPs as post-modifiers written by student 726

Why Cambridge? I verified that Cambridge is ranked **of top*** 5 of best Universities **of the* world.** The amount **of researches*** and the results show the quality and a successful* **of this challenge**.

*Wrong or badly positioned word.

Excerpt 4.22 – Time 2 OPs as post-modifiers written by student 726

This paper presents the most recent developments **of the authors' research center team** regarding power system voltage stability analysis. It focus[es] on the conception **of electrical network equivalents** [,] especially those related to voltage stability indexes. The work aims to identify their differences and similarities in terms **of mathematical basis** and specific applications within electric system activities.

[] – added by the author.

Excerpt 4.23 – Time 3 OPs as post-modifiers written by student 726

The use **of this kind of curve** permits the comprehension **of a lot** [of] different operation conditions with successive load increments. (...) Monitoring the distance between the operational point (Po) to the critical point (Pc – point **of maximum transfer of power**) it is possible [to] work in a stable part **of the curve** and to do an assessment **of the electrical system operation**.

[] – added by the author.

The analysis of those excerpts reveals that this student had already engaged in the use of OPs since Time 1. However, out of the four types of OPs in Excerpt 4.21, only the last two are correct, as the first two should be the preposition *in* instead of *of*. In Time 2, Excerpt 4.22, there is an increase in the use of OPs, and they are all correct. Time 3, illustrated by Excerpt 4.23, shows that the student increased considerably in the use of OPs. This substantial increase may be a characteristic of Brazilian learners, already attested by Queiroz (2019) in a study focused on noun phrase complexity, which also relied on a subcorpus from CorIFA. In her study, the author found that PPs are the most favored type of syntactic post-modification among Brazilian learners, contrary to noun phrases as post-modifiers.

The first finite clausal feature from stage 3, the FCC controlled by adjectives, showed a statistically significant increase over time. This feature includes both types of constructions, *that* complement clauses controlled by adjectives of stage 3 and the extraposed complement clauses originally placed in stage 4 of the index (BIBER et al., 2011). Both features are placed

together here as they are tagged together by the complexity tagger (GRAY et al., 2019), due to the low frequency of simple adjective complements. Further information can be found in Chapter 3, section 3.3.2.

In Time 1, there was no instance of FCC controlled by adjectives. In Times 2 and 3, the vast majority of occurrences were of extraposed complement clauses controlled by adjectives. The adjectival predicates of extraposed *that*-clauses were *possible* and *clear* in Time 2; and *necessary, clear,* and *expected* in Time 3. All these adjectival predicates are from the semantic domain of certainty, except for *necessary*, which is an adjectival predicate from the semantic domain of importance. Excerpt 4.24 below shows an example of an extraposed *that* complement clause. The extraposed construction is in bold, and the finite complement clause is underlined.

Excerpt 4.24 – Time 3 FCC controlled by adjectives written by student 974

It is increasingly necessary <u>that the teacher seeks new methodologies to capture the attention of these students</u>;

The last finite clausal feature from stage 3 (*that*-relative clauses) showed a statistically significant increase over time. In this analysis, all types of *that* relative clauses are considered, unlike the original feature 3e of the Developmental Index (BIBER et al., 2011), which primarily considers relative clauses with animate head nouns. Furthermore, the vast majority of occurrences of this feature in all times were restrictive clauses. The relativizer *that* is more common in conversation than in academic prose, contrary to the relativizer *which,* which is more common in academic prose (BIBER et al., 2011). Relative clauses with a *wh* relativizer will be further discussed in section 4.1.4 below, as it is a Stage 4 feature. Excerpts 4.25, 4.26, and 4.27 below are examples of TRC written by a student whose use of this feature increased over time. The nouns are highlighted in bold, and the relative clauses are underlined.

Excerpt 4.25 – Time 1 TRC written by student 973

I'm learning about computational **programs** <u>that help in data management</u> and to find Empirical evidences* in research.

*Wrong or badly positioned word.

Excerpt 4.26 – Time 2 TRC written by student 973

Hence, this research aims to identify the managers* **characteristics** that affects* the portfolio turnover, and the consequences from these characteristics in the performance of stocks investments funds in Brazil.

*Wrong or badly positioned word.

Excerpt 4.27 – Time 3 TRC written by student 973

The manager of a mutual fund is an **expert** that have* the knowledge and **skills** that are necessary to make the best investments* decisions.

*Wrong or badly positioned word.

The frequency of NFC controlled by other verbs decreased from Time 1 to Time 2 and increased from Time 1 to Time 3, but its variation was not statistically significant. In addition, this feature includes both types of constructions, the *to*-complement clauses and the *ing*-complement clauses. *To*-complement clauses were the most frequent at all times. Besides showing an insignificant number of occurrences in times 1 and 3, *ing*-complement clauses had no occurrences in Time 2. Excerpt 4.28 below demonstrates examples of both clauses controlled by the same verb. The verb is marked in bold, and the complement clauses are underlined.

Excerpt 4.28 – Time 1 *ing*-complement clause written by student 563

I **remember** helping my grandfather to fix his old cars, decorated a lot models of vehicles, and I had a collection of small cars too.

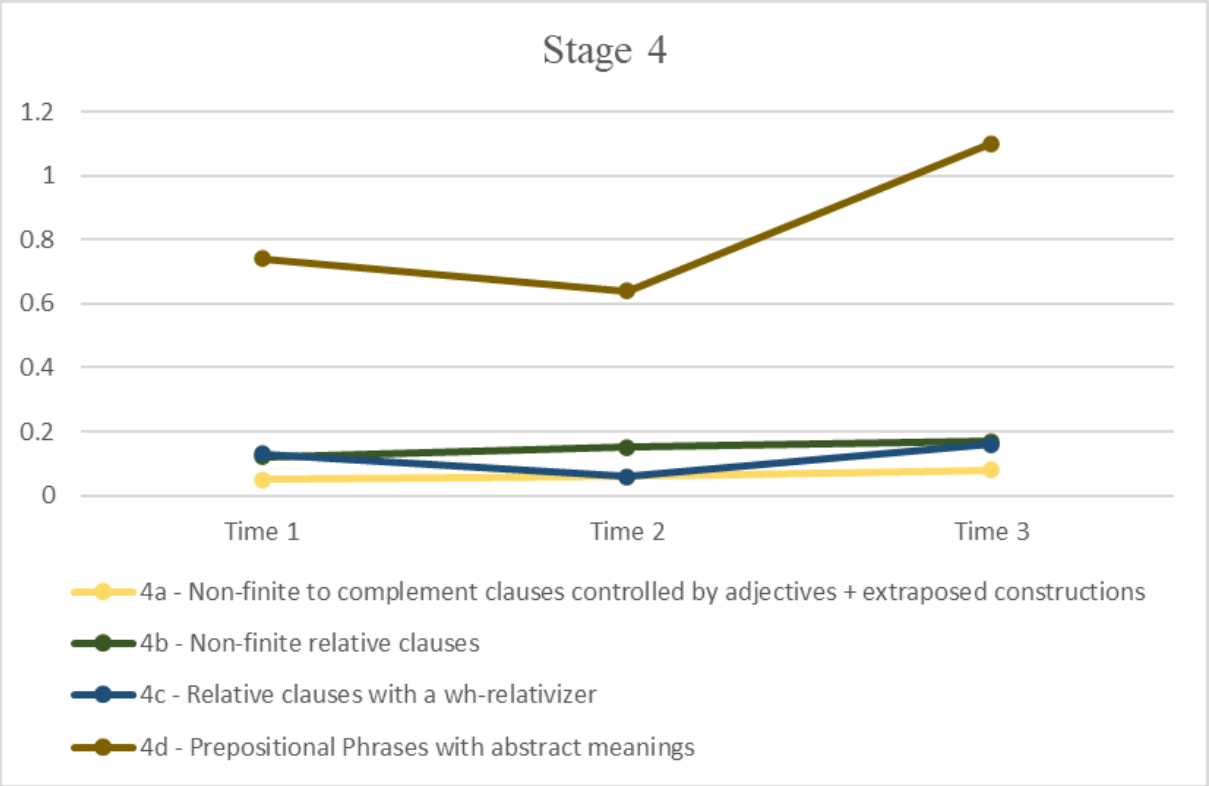
Finally, the only feature that decreased over time in stage 3 (although not steadily and not significantly) was PP with concrete/locative meanings. This feature decreased from Time 1 to Time 3 but increased from Time 2 to Time 3.

4.1.4 Stage 4

Stage 4 features include the *non-finite to complement clauses (NFTC) controlled by adjectives + extraposed constructions* (e.g., *the importance to have*), *non-finite relative clauses* (e.g., *students considering only academic merit*), *relative clauses with a wh relativizer (WRC)* (e.g., *disorders which cause dependence*), and *prepositional phrases (PP) with abstract meanings* (e.g., *program for electric vehicles*). All these features increased from Time 1 to Time 3; however, the PPs with abstract meanings were the only feature showing a statistically

significant variation, according to Table 4.1. Graph 4.4 below demonstrates the mean rates of occurrence of each feature per time.

Graph 4.4 – Mean rates of occurrence of stage 4



Source: Prepared by the author, 2022.

Besides being the only feature to show a statistically significant variation in stage 4, PP with abstract meanings was predominantly the most frequent feature of stage 4 in all times. The second most frequent feature in Time 1 was the WRC, whereas in Times 2 and 3, it was non-finite relative clauses. NFTC + extraposed constructions were the least adopted feature in all times. Perhaps such an acute difference between the frequency of PPs and RCs is because although PPs as post-modifiers can be transformed into relative clauses, PPs as post-modifiers are much more frequent than relative clauses in writing (BIBER et al., 1999).

This type of PP was the second most frequent in the subcorpus, ranking behind only PP as adverbials. Furthermore, a thorough analysis of PP with abstract meanings reveals that *in* was the most common preposition found at all times. In Longman Grammar (BIBER et al., 1999), *in* is the second most frequent preposition in PPs as post-modifiers, whereas the first most frequent is *of*. Other frequent prepositions found in all times were *for, on, to,* and *with*. In Times 2 and 3, there were a few prepositions different from Time 1, such as *regarding* in Time

2, and *over,* and *until* in Time 3. This indicates the increased frequency not only of PPs with abstract meanings over time but also their prepositions' repertoire. Excerpts 4.29, 4.30, and 4.31 demonstrate a student's development. The nouns are marked in bold, the PPs are underlined, and the prepositions are in italic.

Excerpt 4.29 – Time 1 PPs with abstract meanings written by student 516

Raffone's goals* is to become the disease known and join up supporters and **donors** *for* his JAR of Hope Foundation.

*Wrong or badly positioned word.

Excerpt 4.30 – Time 2 PPs with abstract meanings written by student 516

Finally, [it] will be analyzed* the readings that the map has undergone the following years, the way it was used as a source of geographic **information** *about* Portuguese America *along* the whole eighteenth century.

[] – added by the author.
*Wrong or badly positioned word.

Excerpt 4.31 - Time 3 PPs with abstract meanings written by student 516

In 1720, the French scientific advance was capable to fix* with more precision the astronomical positions of meridians and with this to reduce the gross **inaccuracies** *in* determining of* longitudes, in **vigour*** *until* then.

*Wrong or badly positioned word.

The second most frequent feature in Times 2 and 3 was non-finite relative clauses. Two possible patterns are considered in this analysis, namely -*ing* clauses and -*ed* clauses. In academic prose, -*ed* clauses are remarkably more common than -*ing* clauses (BIBER et al., 1999). However, in Times 1 and 3, -*ing* clauses were more common than -*ed* clauses, especially in Time 1, as *ing* clauses accounted for 80% of all non-finite relative clauses. In Time 3, -*ing* clauses accounted for 57%. Finally, Time 2 showed opposite results compared to Time 3, as -*ed* clauses accounted for 57% of the non-finite relative clauses.

Among the most frequent verbs in these constructions, described by Longman Grammar (BIBER et al., 1999), the verb *involve* appeared in -*ing* clauses, and solely in Time 1. The verb *cause* appeared only in -*ed* clauses from Time 1, and the verb *obtain* appeared in Times 2 and 3. Excerpts 4.32 and 4.33 below present examples of non-finite relative clauses written by the

same student in Times 2 and 3. Nouns are marked in bold, clauses are underlined, and verbs are in italic.

Excerpt 4.32 – Time 2 non-finite relative clause written by student 726

The authors have intensely worked on the determination of system equivalents using **measurements** *provided* by Synchronized Phasor Measurement Systems, and academic and practical experiences.

Excerpt 4.33 – Time 3 non-finite relative clause written by student 726

This last study is very useful because it is possible to compare the **results** obtained against other **studies** *done* by other researchers using the same electrical circuit.

Wh-relative clauses were the second most frequent feature in Time 1, and the third most frequent in Times 2 and 3. As previously addressed in subsection 4.1.3 above in the discussion about *that* RCs, the relativizer *which* is the most frequent in academic prose. In fact, it was the most frequent in *wh*- relative clauses at all times. *Who* was also found in all times, whereas *where* appeared exclusively in Time 1, and *why* exclusively in Time 3. Nevertheless, That-relative clauses are still more frequent in Time 3, in comparison to WRCs. Excerpt 4.34 below demonstrates this feature in use.

Excerpt 4.34 – Time 3 *wh*- relative clause written by student 331

By the way, there is no merit for a **candidate** who had the privilege of being born in a family with good financial conditions and could study in the best private schools to rank in [a] more advantageous position than **candidates** who had no choice and studied in public schools of poor quality.

[] – added by the author.

The last and least frequent feature of stage 4 that increased in frequency was the NFTC controlled by adjectives + extraposed constructions. Similar to feature 3b of stage 3, this feature includes both patterns, *to*-clauses controlled by adjectives and the extraposed *to*-clauses. In Time 1, *to*-clauses controlled by adjectives were more frequent than extraposed *to*-clauses. On the contrary, in Times 2 and 3, extraposed *to*-clauses were more frequent than *to*-clauses.

Furthermore, the choice of adjectives divided by semantic domains controlling *to* and

extraposed clauses was rather diverse at all times. In Time 1, adjectives controlling *to* clauses and extraposed can be divided into the semantic domains of evaluation (33.33%), ability or willingness and necessity (each with 22.22%), and ease (11.11%). In Time 2, adjectives were divided into necessity or importance (45.45%), ease (36.36%), personal affective stance, and ability or willingness (each with 9.09%). Finally, in Time 3, they were divided into necessity or importance (42.86%), ability or willingness (35.71%), ease (14.29%), and personal affective stance (7.14%). Excerpt 4.35 below shows an example of a non-finite *to* complement clause controlled by an adjective in Time 3. The adjective is marked in bold, and the clause is underlined.

Excerpt 4.35 – Time 3 *to*-clause controlled by adjective written by student 713

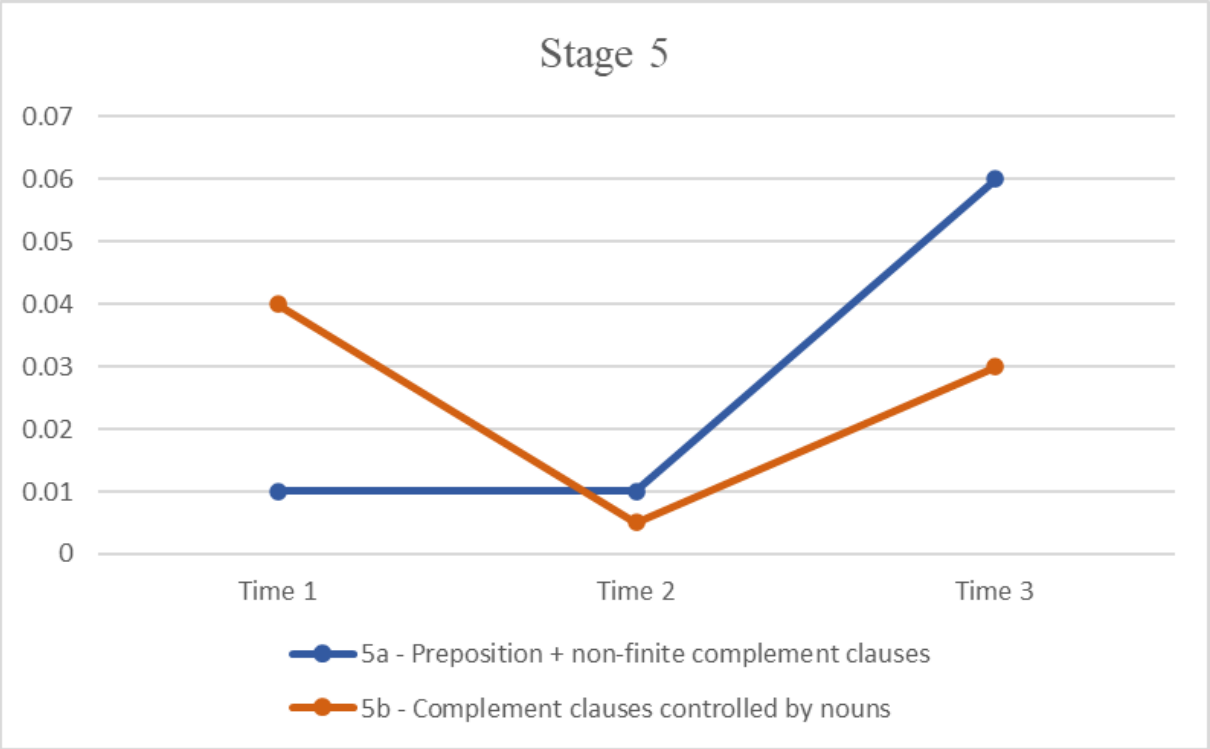
Understanding Genetics helps people being **able** <u>to give opinions</u> in* political and society topics that appear in media news and involve ethics and political issues.

*Wrong or badly positioned word.

4.1.5 Stage 5

Stage 5 from the original framework designed by Biber et al. (2011) has four distinct features. However, only two features are taken into consideration in this analysis, the *preposition + non-finite complement clauses (NFC)* (e.g., *positions for <u>studying</u>*) and the *complement clauses controlled by nouns* (e.g., *a consensus among researchers <u>that no one is introverted</u>*), due to complexity tagger limitations. Chapter 3 in section 3.3.2 better describes the tagging process. Graph 4.5 below demonstrates the frequency shifts of both features over time.

Graph 4.5 – Mean rates of occurrence of stage 5



Source: Prepared by the author, 2022

The *preposition + NFC*, which is the first feature of this stage, showed a statistically significant increase from Time 1 to Time 3. In turn, the second feature, *complement clauses controlled by nouns* showed no statistically significant variation over time. Furthermore, the frequency of this feature decreased from Time 1 to Time 3, although it increased from Time 2 to Time 3.

Results on the structure of preposition + NFCs were divided into two patterns:

noun + preposition + non-finite comp. clause

any word + preposition + non-finite comp. clause

Time 3 was the only one that presented occurrences of both types of structures, whereas Time 1 there only showed occurrences of the first structure with a noun before the preposition, and Time 2 only showed occurrences of the second structure, without a noun. Excerpts 4.36 and 4.37 present occurrences of both patterns written by the same student at two separate times. The preceding word is in bold, the preposition is in italic, and the complement clause is underlined.

Excerpt 4.36 – Time 2 any word + preposition + non-finite comp. clause written by student 774

Studying the reader behavior applied in reconstruction of texts is the theme of this article. Reconstruction which [was] **made** *by* using their creative power.

[] – added by the author.

Excerpt 4.37 – Time 3 noun + preposition + non-finite comp. clause written by student 774

The **choice** *between* learning English online or in the classroom have* not to be made, because blending learning is the best new choice.

*Wrong or badly positioned word.

In summary, this section presented the results of the development of grammatical complexity by students over time. The results of the statistical analysis presented in Table 4.1 corroborate hypotheses H1 and H2, since some clausal features from stages 1 and 2 – such as the FCC controlled by common verbs of stage 1 – increased in frequency, albeit this increase was not statistically significant. Furthermore, phrasal features with statistically significant variations all increased in frequency, from the second stage on, such as the attributive adjectives. The non-finite clausal feature (preposition + NFC) from stage 5, also showed a statistically significant increase over time. In addition, since time 1, students relied much more frequently on phrasal than finite and non-finite clausal features.

The discourse qualitative analysis showed that not only did students display an increase in the mean rates of features' occurrence over time but a lexical variation of some features as well. This can be seen in the discussions of attributive adjectives and adverbs as adverbials, especially regarding linking and stance adverbials (Subsection 4.1.2), which may suggest a repertoire development. Students also presented similar characteristics already pointed out in Parkinson and Musgrave (2014), relative to the high reliance on attributive adjectives, instead of relying on nouns as premodifiers, a feature that was already confirmed to be more frequent in proficient students' texts and specialized writing. This may indicate that the time frame chosen for the analysis was not able to capture such a development, primarily because there is no way to attest whether students' proficiency level increased (from B1 to B2) in the timespan of this analysis.

Nonetheless, to better understand the extent of the development of grammatical complexity in texts written by Brazilian learners of English, we must investigate the variables that can interfere in the lexical and feature selections, namely register, and academic divisions. Indeed, these have been asserted to be of major relevance in grammatical complexity analysis by Staples et al. (2016); therefore, section 4.2 below will address four specific features (two phrasal and two clausal) in the scope of a quasi-longitudinal analysis across registers and academic divisions.

4.2 Quasi-longitudinal analyses

For the quasi-longitudinal analyses, in which each text is treated as an observation, only four features were selected:

1. finite adverbial clauses
2. finite relative clauses (*that* and *wh*)
3. attributive adjectives, and
4. nouns as premodifiers.

These features were chosen based on the longitudinal study by Biber et al. (2020). According to the authors, these four features represent two structural types – namely dependent clause types (i.e., finite adverbial clauses and finite relative clauses), and phrases (i.e., attributive adjectives and nouns as premodifiers) – and are crucial for assessing different levels in academic writing. These features also have different syntactic functions, as the finite adverbial clauses (AC) functions as clause modifiers, whereas relative clauses (RC), attributive adjectives (AA), and nouns as premodifiers all modify a head noun. Section 3.4 in Chapter 3 introduced the selection of features. Findings and discussions are divided according to register and academic divisions and a qualitative discussion will be performed along with the quantitative analysis to make the variation of the features within variables clearer.

In the quantitative analysis, we tested the linear mixed-effects model per feature, to properly check whether variables or interactions between variables were predictors of variance. Features were treated as dependent variables, whereas semester (time), register, and academic division were treated as independent variables. Also, students were nested as random effects, whereas variables were considered fixed effects. This process is described in further detail in Chapter 3, Section 3.4. Nevertheless, no feature presented a statistically significant effect on the dependent variables. Tables 4.2, 4.3, 4.4, and 4.5 demonstrate the results per each feature.

4.2 – Regression model predicting the use of finite adverbial clauses

Analysis of Deviance Table (Type II Wald chi-square tests)					
	Chisqu	**Df**	**Pr(>Chisq)**		
Semester	3.42	2	0.18		
Register	3.58	2	0.16		
Academic division	1.51	3	0.67		
Random Effects					
Groups	**Name**	**Variance**	**Std. Dev.**		
ï..Student	(Intercept)	0.04	0.20		
Residual		0.21	0.45		
Number of obs: 39, groups: ï..Student, 13					
Fixed Effects					
	Est.	**S.E.**	***t val.***	**d.f.**	***p*-value**
(Intercept)	0.27	0.44	0.61	30.99	0.55
Semester: Time 2	-0.02	0.40	-0.05	25.97	0.96
Semester: Time 3	-0.94	0.54	-1.76	29.82	0.09
Register: Argumentative Essay	1.03	0.66	1.58	29.02	0.13
Register: Statement of Purpose	0.07	0.42	0.16	27.12	0.87
Academic division: Humanities and Arts	0.19	0.29	0.67	10.09	0.52
Academic division: Physical Sciences and Engineering	0.26	0.24	1.09	8.59	0.31
Academic division: Social Sciences and Education	0.37	0.39	0.96	8.39	0.36

Significance level (alpha value): $p < 0.05$
Significance code: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Prepared by the author, 2022.

Table 4.3 – Regression model predicting the use of finite relative clauses

Analysis of Deviance Table (Type II Wald chi-square tests)			
	Chisqu	**Df**	**Pr(>Chisq)**
Semester	3.59	2	0.16
Register	3.30	2	0.19
Academic division	2.00	3	0.57
Random Effects			
Groups	**Name**	**Variance**	**Std. Dev.**
ï..Student	(Intercept)	0.02	0.17
Residual		0.39	0.62
Number of obs: 39, groups: ï..Student, 13			

Fixed Effects					
	Est.	**S.E.**	***t val.***	**d.f.**	***p*-value**
(Intercept)	0.23	0.58	0.39	31.00	0.70
Semester: Time 2	0.30	0.54	0.55	26.94	0.59
Semester: Time 3	1.21	0.71	1.69	30.67	0.10
Register: Argumentative Essay	-0.51	0.88	-0.58	30.11	0.56
Register: Statement of Purpose	0.57	0.57	1.00	28.30	0.33
Academic division: Humanities and Arts	0.42	0.35	1.19	10.31	0.26
Academic division: Physical Sciences and Engineering	0.30	0.29	1.04	8.49	0.33
Academic division: Social Sciences and Education	0.49	0.46	1.06	8.27	0.32

Significance level (alpha value): p < 0.05
Significance code: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Prepared by the author, 2022

Table 4.4 – Regression model predicting the use of attributive adjectives

Analysis of Deviance Table (Type II Wald chi-square tests)			
	Chisqu	**Df**	**Pr(>Chisq)**
Semester	1.71	2	0.42
Register	1.04	2	0.59
Academic division	1.31	3	0.72
Random Effects			
Groups	**Name**	**Variance**	**Std. Dev.**
ï..Student	(Intercept)	1.19	1.09
Residual		5.76	2.40
Number of obs: 39, groups: ï..Student, 13			
Fixed Effects			
	Est.	**S.E.**	***t val.***
(Intercept)	6.52	2.29	2.85
Semester: Time 2	-0.34	2.09	-0.16
Semester: Time 3	3.49	2.80	1.24
Register: Argumentative Essay	-3.21	3.43	-0.94
Register: Statement of Purpose	-0.67	2.22	-0.30
Academic division: Humanities and Arts	0.33	1.52	0.22
Academic division: Physical Sciences and	-0.05	1.26	-0.04

Engineering					
Academic division: Social Sciences and Education	-2.00	2.04	-0.98	8.40	0.35

Significance level (alpha value): $p < 0.05$
Significance code: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Prepared by the author, 2022

Table 4.5 – Regression model predicting the use of nouns as pre-modifiers

Analysis of Deviance Table (Type II Wald chi-square tests)					
	Chisqu		**Df**		**Pr(>Chisq)**
Semester	0.62		2		0.73
Register	1.99		2		0.36
Academic division	2.90		3		0.40
Random Effects					
Groups	**Name**		**Variance**		**Std. Dev.**
ï..Student	(Intercept)		3.31		1.81
Residual			4.31		2.07
Number of obs: 39, groups: ï..Student, 13					
Fixed Effects					
	Est.	**S.E.**	***t val.***	**d.f.**	***p*-value**
(Intercept)	6.54	2.19	2.98	29.88	0.01
Semester: Time 2	-0.81	1.83	-0.45	24.03	0.66
Semester: Time 3	1.62	2.51	0.65	26.73	0.52
Register: Argumentative Essay	-4.11	3.05	-1.35	26.00	0.19
Register: Statement of Purpose	-1.92	1.95	-0.99	24.65	0.33
Academic division: Humanities and Arts	-2.79	1.83	-1.53	9.61	0.16
Academic division: Physical Sciences and Engineering	-0.53	1.55	-0.34	8.78	0.74
Academic division: Social Sciences and Education	0.06	2.52	0.02	8.66	0.98

Significance level (alpha value): $p < 0.05$
Significance code: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Source: Prepared by the author, 2022.

Perhaps the lack of significance across features can be explained by the fact that the sample size does not suffice for this type of model, since the corpus had more independent variables (three semesters, four registers, and four academic divisions) than texts (n = 39). It is interesting to note, however, that certain variables of some features were almost statistically

significant ($p < .05$), such as register and semesters in ACs (e.g., *avoided if early detections are implemented)* (Table 4.2), which presented *p*-values of 0.16 and 0.18, respectively. As for fixed effects, two variables were almost statistically significant and predictors of linguistic variation, namely the argumentative essay (AE compared to the abstract (AB), since it had a 0.13 *p*-value and a positive estimate, thus indicating an increase; and semester 3 compared to semester 1, which had a 0.09 *p*-value and a negative estimate, thus indicating a decrease in the use of this feature.

As for relative clauses (e.g., *merit that brings excellence to the university*) (Table 4.3), semester and register almost predicted linguistic variation ($p < .05$) once again, showing *p*-values of 0.16 and 0.19, respectively. As for fixed effects, only semester 3, compared to semester 1, achieved an almost statistical significance ($p = 0.10$). Moreover, the positive estimate in semester 3 indicates an increase in the frequency of RCs from semester 1 to semester 3. Finally, as for attributive adjectives (e.g., *genetic information*) (Table 4.4), no variable showed an almost statistical significance.

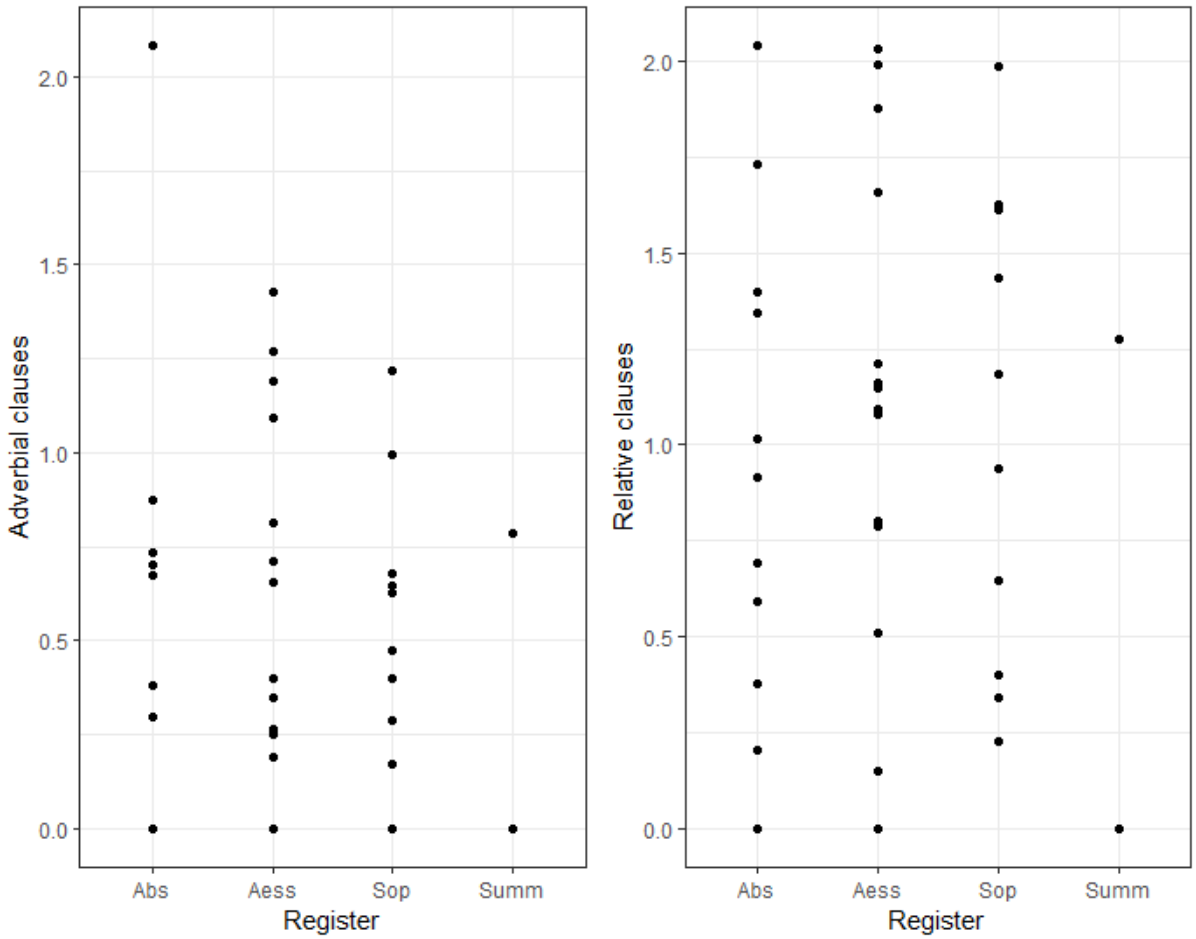
As for nouns as premodifiers (e.g., *disease carrier*) (Table 4.5), only some variables as fixed effects were almost able to predict linguistic variation ($p < .05$), namely the academic division of Humanities and Arts (HA) compared to Biologic and Health Sciences (BHS), which had a *p*-value of 0.16 and a negative estimate, thus indicating a decrease in the use of this feature; and the argumentative essay register, in contrast to abstracts, which had a *p*-value of 0.19 and a negative estimate, thus indicating a decrease on the frequency of this feature in AEs texts, as for the ABs texts.

To accurately describe the variations across registers and academic divisions, a qualitative analysis will be performed next, considering the features' normalized rates of occurrences per variable, and disregarding their statistical significance. The section was divided into two subsections, namely "register" and "academic division", as follows.

4.2.1 Register

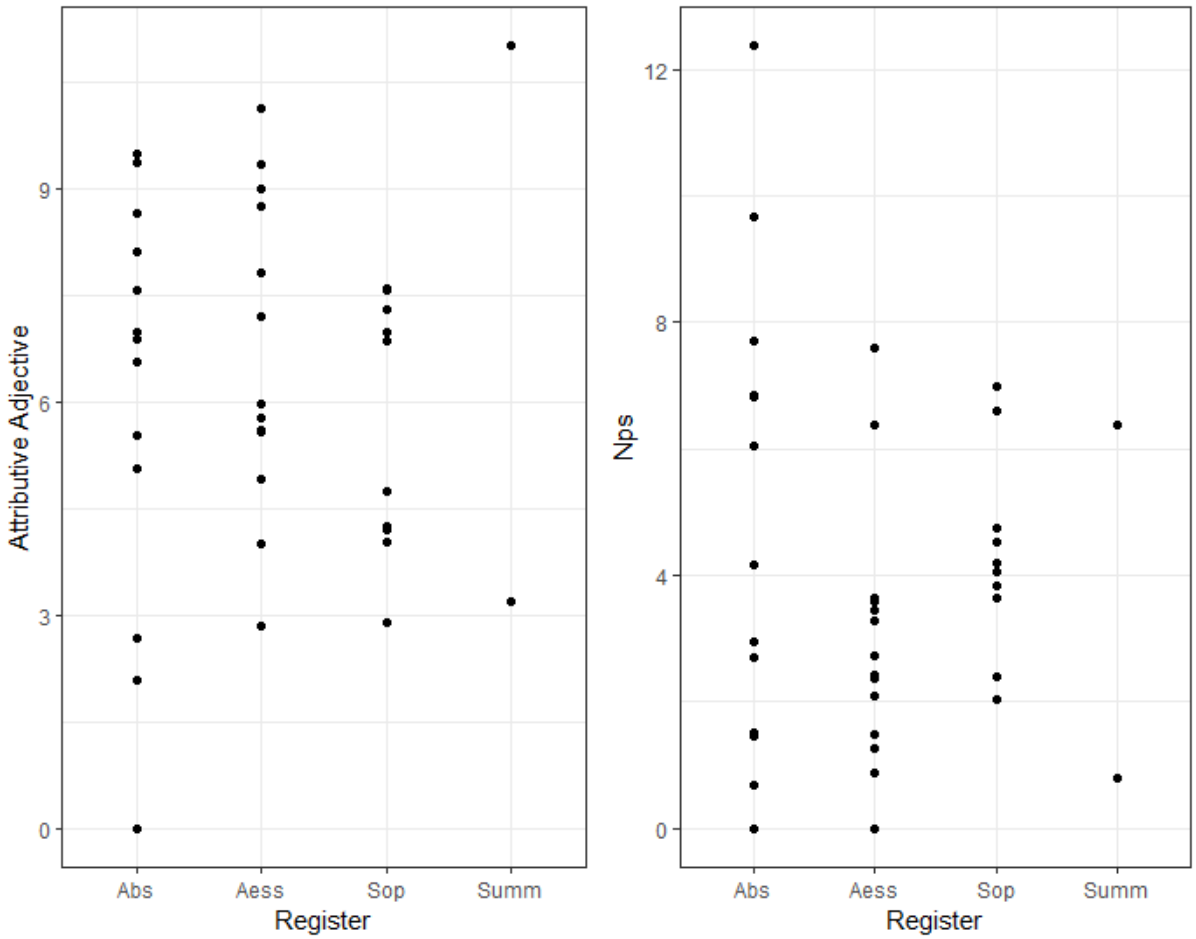
Register plays a key role in grammatical complexity analysis, as already attested by Staples et al. (2016), and is an important mediator of language development and variation. Thus, it is crucial to verify the extent of grammatical complexity variation across registers through the four features selected for the analysis (adverbial clause (AC), relative clause (RC), attributive adjective (AA), and nouns as premodifiers). Figures 4.4 and 4.5 below show the distribution plots of clausal and phrasal features per register.

Figure 4.4 – Plot of adverbial and relative clauses per register (rates of occurrence per 1,000 words)



Source: Prepared by the author, 2022.

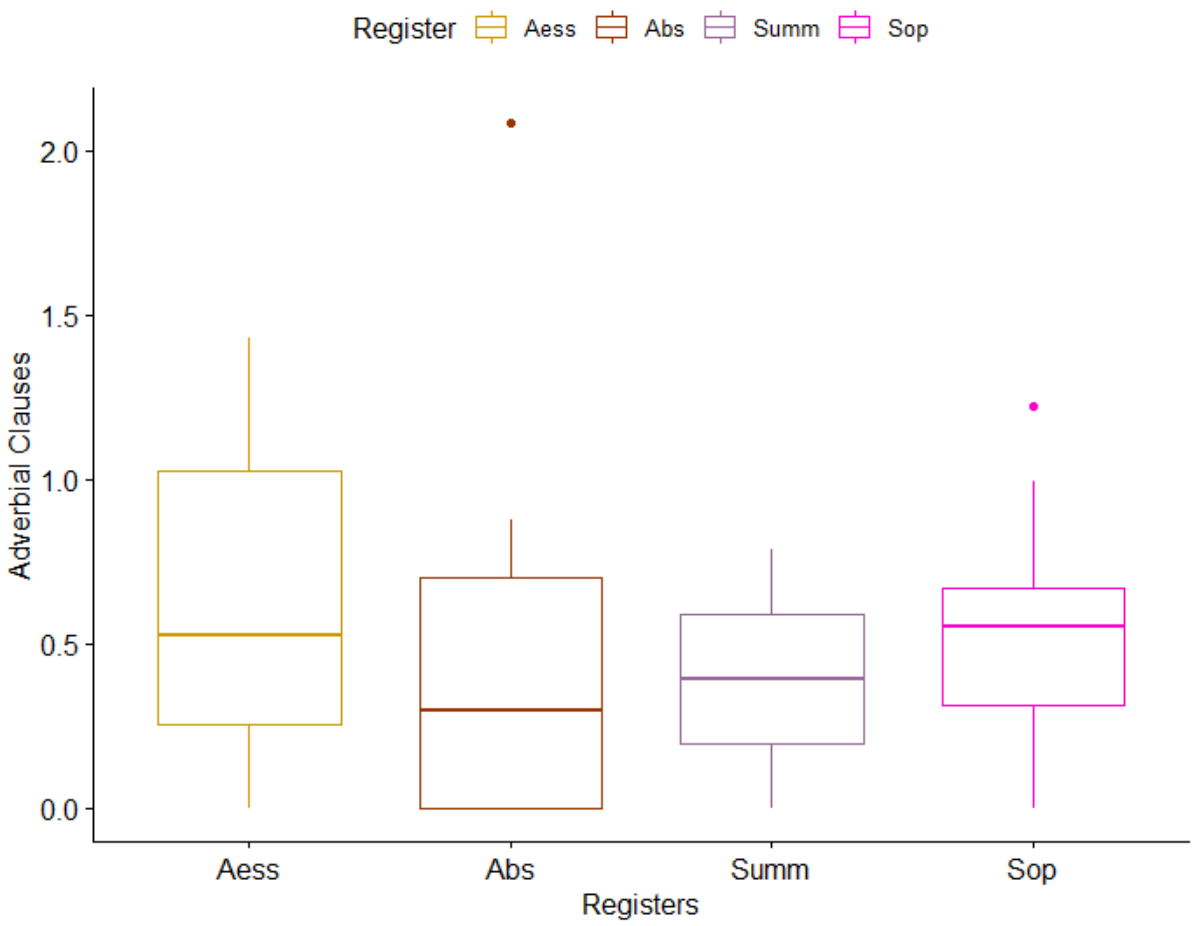
Figure 4.5 – Plot of attributive adjectives and nouns as premodifiers per register (rates of occurrence per 1,000 words)



Source: Prepared by the author, 2022.

As the plots above show, adverbial clauses (AC) and relative clauses (RC) had very few occurrences per register, whereas attributive adjectives (AA) and nouns as premodifiers had more occurrences. The plot analysis reveals that the number of samples (texts) of the summary (SUM) register is the smallest (only two), and the number of argumentative essays (AE) samples is the highest. Thus, it is important to present the mean, median, IQR, and SD of every register for each feature, to ensure an accurate comparison and an organized discussion. We begin with the register analysis in ACs. Figure 4.6 shows the boxplot of ACs per register, and Table 4.6 presents the features data per register. The information is repeated in the tables and boxplots to contribute to the comprehension of the results.

Figure 4.6 – Boxplot of ACs per register



Source: Prepared by the author, 2022.

Table 4.6 – Adverbial clause results per register

Register	Count	Mean	SD	Median	IQR
Abstract	13	0.44	0.59	0.29	0.69
Argumentative Essay	14	0.61	0.48	0.52	0.77
Statement of Purpose	10	0.54	0.36	0.54	0.35
Summary	2	0.39	0.55	0.39	0.39

Source: Prepared by the author, 2022.

Adverbial clauses (AC) were the least frequent per text among all registers, according to Figure 4.4, the boxplot in Figure 4.6, and the mean rates of occurrence presented in Table 4.6. In fact, argumentative essays (Figure 4.6 and Excerpt 4.38 below) were the register with the highest mean. Although the AEs had the highest mean score for this feature, the text that had the highest frequency among all ACs stemmed from the abstract register (Excerpt 4.39

below), as displayed in the outlier in Figure 4.6. Indeed, in the excerpts below, ACs are underlined, and the subordinator is underlined and in bold.

Excerpt 4.38 – AC of time in AEs 984

And energy means quality of life or even life **when** considered critical situations as a natural disaster, for example.

Excerpt 4.39 – AC of reason in ABs 1398

It is possible to reach more people **if** the consumers keep "good habits", not wasting energy.

One of the reasons that may explain why argumentative essays had the highest number of occurrences among adverbial clauses is the communicative purposes of these registers. AEs are typically argumentative, so students are expected to try and convince readers of their ideas and may, therefore, build their arguments by relying on finite adverbial clauses. For example, in Excerpt 4.38 the student is reinforcing his or her point of view by adding a hypothetical time using the subordinator *when* to assign more importance to his or her statement.

Furthermore, a thorough analysis of adverbial clauses in essays shows the following semantic categories, from the most to the least frequent: ACs of time (i.e., with the subordinators *when* and *since*), contingency adverbial of condition (i.e., with the subordinator *if*), contingency adverbial of reason (i.e., with the subordinator *because*), and concession (i.e., with the subordinator *although*).

Besides raking as the second register with the highest mean, the Statement of Purpose (SOP), also showed occurrences of the same semantic categories, but not in the same order of frequency. Indeed, it had more occurrences of adverbial clauses of time (i.e., with the subordinators *when* and *since*), contingency adverbial of reason (i.e., with the subordinator *because*), contingency adverbial of condition (i.e., with the subordinator *if*), and concession (i.e., with the subordinators *although* and *though*).

A Statement of Purpose (SOP), also known as a personal statement, is a highly demanded register for admission processes in international universities and programs, and even job interviews (SAMRAJ; MONK, 2008). In fact, SOPs have been found to contain five moves: "*persuading the reader, introducing the applicant and letter objectives, reasons for applying,*

expressing future expectations, and final greeting and signing" (LÓPEZ-FERRERO; BACH, 2016, p. 307). This means that students must present a range of statements explaining why they fit in such positions, which can require the use of adverbial clauses and their extensive semantic purposes.

This can be attested to since almost all adverbial clauses in statements of purpose position the personal pronoun *I* as a subject first, in comparison to the argumentative essay, for instance, in which the vast majority of them referred to third-person pronouns, particularly inanimate beings (e.g., *study*). Excerpt 4.40 below is one of the few instances of inanimate subjects in ACs of SOPs. In this fragment, the student describes the university that he or she has graduated from, and to boost the university's prestige, he or she enhances the argument by specifying how renowned their field of study is, through the use of an adverbial clause.

Excerpt 4.40 – AC in SOP 1252

I graduated at Universidade Federal de Minas Gerais – Brazil, a prestigious University, ranked as one of top 10 in Latin America, mainly **if** it considers the* Exact Science.

*Wrong or badly positioned word.

Abstract (ABs) was the third-highest register with occurrences of adverbial clauses in the following semantic categories: ACs of time (i.e., with the subordinators *when* and *since*), contingency adverbial of reason (i.e., with the subordinator *because*), contingency adverbial of condition (i.e., with the subordinator *if*), and result (i.e., with the subordinator *so that*). In Excerpt 4.41 below, taken from an ABs, the student is explaining the topic being discussed, presenting a cause/consequence (i.e., *preventing the public / so that it is…)* relation with an AC of the result.

Excerpt 4.41 – AC in ABs 1702

Large car manufacturing companies have their own research and development departments, secreting important information regarding the performance of the car, preventing the public from having access to this data so that it is not possible to copy the cars, and losing [the] market.

[] – added by the author.

The register with the lowest mean rate of occurrences of adverbial clauses was the summary (SUM). Moreover, SUM only showed occurrences of adverbial clauses of reason

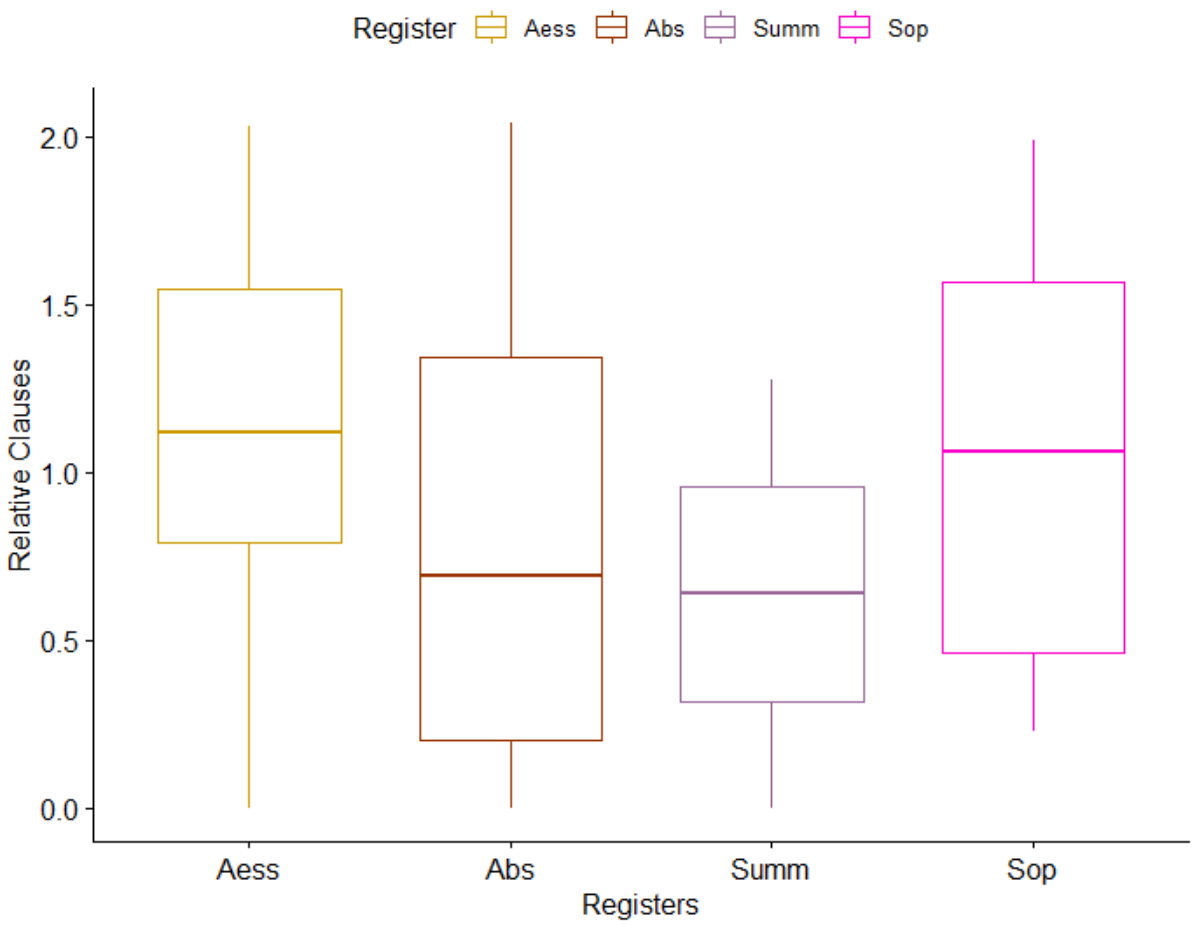
using the subordinator *because*. In Excerpt 4.42, which was taken from a SUM, the student presents an AC of the reason for the statement he or she had previously made. The AC is underlined, and the subordinator is in bold and underlined.

Excerpt 4.42 – AC in SUM 1115

The Great Depression wasn't expected by American People **because** in the period before it, America was the richest country in the world.

The second feature analyzed per register (relative clauses) is a clausal feature similar to the ACs, but different in their syntactic functions because RCs are noun-modifying features. Along with the plot demonstrating data distribution of this feature across registers, presented in Figure 4.4 above, Figure 4.7 below shows the boxplot of RCs per register, similar to the information available in Table 4.7 below.

Figure 4.7 – Boxplot of RCs per register



Source: Prepared by the author, 2022.

Table 4.7 – Relative clause results per register

Register	Count	Mean	SD	Median	IQR
Abstract	13	0.79	0.68	0.69	1.14
Argumentative Essay	14	1.11	0.63	1.12	0.75
Statement of Purpose	10	1.04	0.62	1.06	1.11
Summary	2	0.63	0.90	0.63	0.63

Source: Prepared by the author, 2022.

According to the mean rates of occurrences, relative clauses were more frequent than adverbial clauses. However, interestingly, the order of frequency across registers remained the same, as the argumentative essays presented the highest mean, followed by statement of purposes, abstracts, and summaries. This outcome can be a further indicator that some registers

require the use of clausal features, such as AEs and SOPs, contrary to other ones such as ABs and SUMs. In this part of the analysis, both *that* and *wh*-relative clauses are included.

That-relative clauses were more frequent than *wh*-clauses in AEs and ABs, whereas *wh*-clauses were more frequent in SOPs and SUMs. In fact, no occurrence of *that*-relative clauses was identified for the SUM register. Furthermore, not all occurrences of *wh*-clauses started with the relativizer *which*, which is the most frequent relativizer in academic prose (BIBER et al., 1999); yet, there were instances of *who* in all registers, whereas *where* only occurred in ABs, and *why* only in AEs

One of the reasons why *which* is preferred over *that*-clauses is due to the belief that this relativizer is more formal, whereas *that* is regarded as more colloquial (BIBER et al., 1999). Moreover, the head noun can also influence the relativizer's choice, as it can be animate or inanimate. When animate, it is usually followed by the relativizer *who*, although it can also be followed by *that*; and when inanimate, it can be followed by both *that* or *which*, also considering restrictive and non-restrictive clause differences[21].

Nonetheless, some L2 studies already attested a high reliance on *that*-clauses by learners, in comparison to *wh*-clauses, such as Roberts (2017), who analyzed Swedish L2 learners of English. They stated that such a high degree of can be explained by the unmarkedness characteristic of the relativizer *that*, especially among L1 students, since *that* can be regarded as easier than *which* because it accepts both animate and inanimate head nouns. This can be one of the reasons why *that*-clauses were more frequent than *wh*-clauses in AEs and ABs.

As for AEs, almost all instances of *that*-clauses were controlled by an inanimate head noun, such as *universities, technique, topics, cities*, among others. Nevertheless, there were still some occurrences of animate head nouns controlling *that* relative clauses, such as *population,* and *society*. In ABs, there is also the occurrence of the words *population* and *animals*. As for SOPs, there was the occurrence of the word *teams*.

We searched these words in the collocates option in the academic section of COCA[22], and we found that *that* is the most frequent relativizer for such antecedent nouns, although they

[21] According to Longman Grammar "Restrictive relative clauses are used to establish the reference of the antecedent, while non-restrictive relatives give additional information which is not required for identification" (BIBER et al., 1999, p. 195).

[22] The Contemporary Corpus of American English (COCA) is the

are animate nouns. This can be because, even though animate, these words refer to groups, rather than only one animate being, which can be a determining factor for such choices of relativizer, aside from the above-mentioned reasons already presented. Figure 4.8 below demonstrates the frequency of determiners after the word *population* in COCA.

Figure 4.8 – Frequency of relativizers *that* and *who* following the word *population* in the academic section of COCA

SEC 1 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	THAT	498	498	4.2	4.2	1.0

SEC 1 (ACADEMIC): 119,790,456 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	WHO	100	100	0.8	0.8	1.0

Source: Prepared by the author, 2022.

Almost all instances of *who* in all registers were correctly employed, as they all were controlled by animate head nouns, such as *boys*, *person*, *people*, *student*, *candidates*, *investors*, etc. There was only a single instance of the word *companies* controlling a relative clause with the relativizer *who* in an AEs. In COCA, the pronoun *who* is not frequently placed after this word, as opposed to *that*, which is extremely frequent. This can be because this pronoun "is used almost exclusively with an animate (human) head" (BIBER et al., 1999, p. 612). Excerpts 4.43 and 4.44 below demonstrate examples of relative clauses with *that* and *who* in the AEs. Nouns are marked in bold, relativizers are in italic, and relative clauses are underlined.

Excerpt 4.43 – *That*-clause in AEs 2106

> Genes have **information** *that* give instructions to tell your body how to make all the proteins it needs to survive and grow.

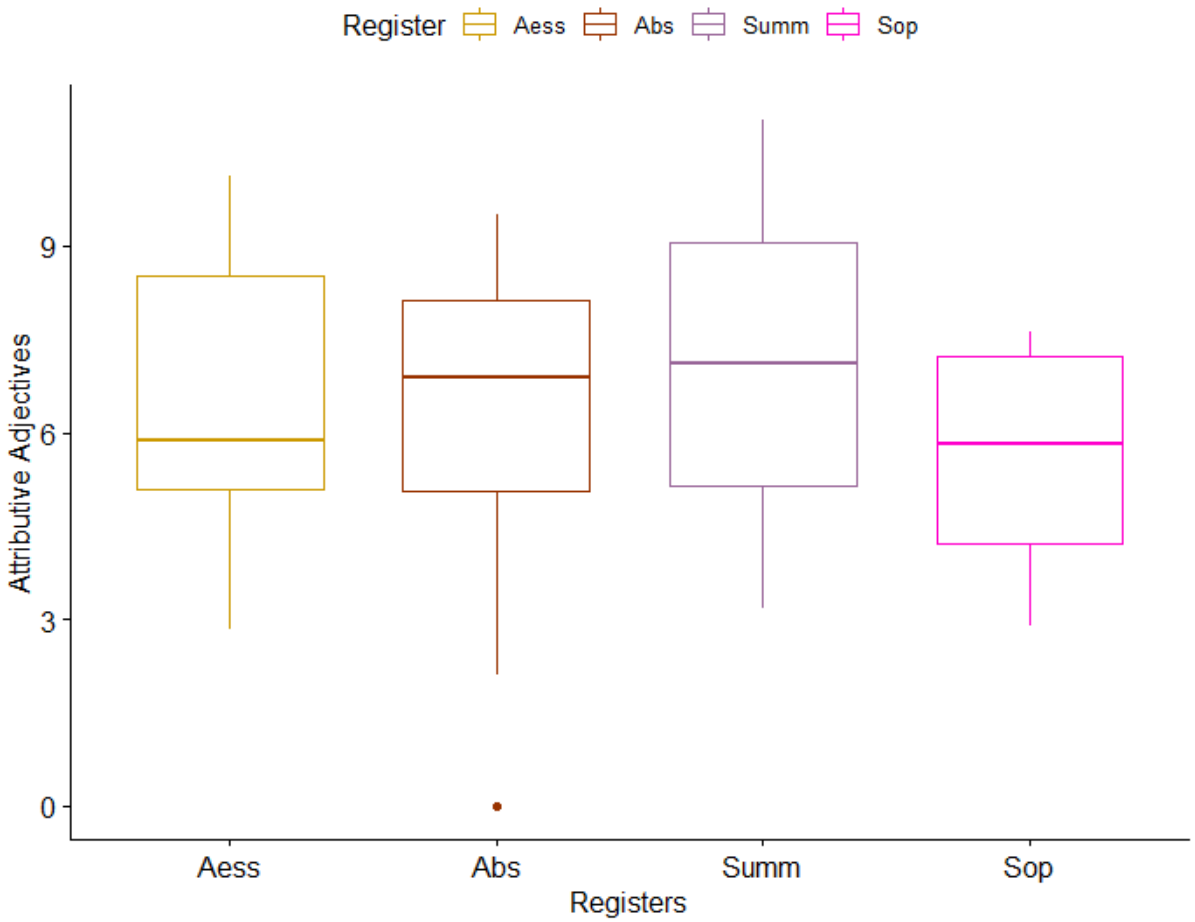
Excerpt 4.44 – *Wh*-clause in AEs 2110

> (...) that are **investors** *who* do not have the experience and knowledge to invest in the financial market, (...).

https://www.english-corpora.org/coca/

The first phrasal feature to be discussed concerning variation across registers is attributive adjectives (AA). As already presented in Figure 4.5 above, data distribution was very similar per register, which can be seen in Figure 4.9 and Table 4.8 below. The outlier in the ABs, in Figure 4.8, concerns the text with the small number of occurrences of AAs in the whole subcorpus.

Figure 4.9 – Boxplot of AAs per register



Source: Prepared by the author, 2022.

Table 4.8 – Attributive adjective results per register

Register	Count	Mean	SD	Median	IQR
Abstract	13	6.08	2.93	6.90	3.05
Argumentative Essay	14	6.42	2.36	5.88	3.44
Statement of Purpose	10	5.65	1.79	5.81	3.03
Summary	2	7.10	5.54	7.10	3.92

Source: Prepared by the author, 2022.

Attributive adjectives (AA) were the most frequent feature in all registers, and SUMs were the register with the highest mean, followed by AEs, ABs, and SOP. Perhaps the reason why SUM had the highest mean is that only two texts in the subcorpus were of this type, and one of the texts had the highest normalized frequency of attributive adjectives in the whole subcorpus, with a frequency of 11.02.

Regarded as "one of the primary mechanisms used to pack additional information into noun phrases" (BIBER et al., 1999, p. 506), it is no surprise that this feature was the most preferred by learners in all registers. Moreover, this feature can also be divided by semantic categories, and the relational and topical categories are the most recurrent in academic prose. In all registers, the semantic categories of topical, relational, evaluative, and affiliative adjectives were the most employed, albeit not in this exact order. Some examples per register are described below:

- As for SUMs, instances can be divided into topical (e.g., *economic, muscular, productive*), evaluative (e.g., *relevant, richest, tragic, disastrous, great*), affiliative (e.g., *American*), relational (e.g., *main, whole*), and descriptor of time (e.g., *old*).
- As for AEs, there were instances of all semantic categories, such as affiliative (e.g., *American, Brazilian, Chinese, Portuguese*), topical (e.g., *academic, aerospace, electrical, electric, public, introverted, genetic, federal*), relational (e.g., *main, mutual, major, final, different, common*), evaluative (e.g., *good, great, best, important*), descriptor of size (e.g., *big, high, short, small*), descriptor of color (e.g., *black, white*), and descriptor of time (e.g., *new, early*).
- As for ABs, the following semantic categories were found: affiliative (e.g., *American, Brazilian, European*), topical (e.g., *academic, auriferous, biological, comic, historical*), relational (e.g., *common, different, following, general, main, similar*), evaluative (e.g., *best, ideal, great*), descriptor of size (e.g., *big, high, large*), and descriptor of time (e.g., *recent*).
- As for SOPs, there were instances of affiliative (e.g., *Brazilian, French*), topical (e.g., *academic, aerodynamic, biological, democratic, federal, financial*), relational (e.g., *mutual, whole, different, single*), evaluative (e.g., *important, good, great*), descriptor of size (e.g., *big, high, huge, large, small*), and descriptor of time (e.g., *new, old)*

In addition to the preferences over semantic categories, some adjectives in attributive adjectives constructions were quite frequent according to each register. Appendix B shows the

most frequent AAs of all registers (more than five occurrences). According to the instances presented above, although all registers had similar preferences over semantic categories, lexical preferences over adjectives varied – except for the evaluative *great*, which appeared in all registers. Excerpts 4.45 and 4.46 below present examples of AAs in the registers with the highest means: SUM and AEs. The whole construction is underlined, adjectives are in bold, and nouns are in italic.

Excerpt 4.45 – Attributive adjectives in SUM 1115

The **Great** *Depression* as it became historically known had as [the] **main** *cause* [of] overproduction and brought **many** *problems* for **American** *people*, like increasing unemployment rates and dropping in income. The **Great** *Depression* wasn´t expected by **American** *People* because in the period before it, America was the **richest** *country* in the world. The United States had in that moment made **relevant** *progress* in the* science and technology too.

*Wrong or badly positioned word.
[] – added by the author.

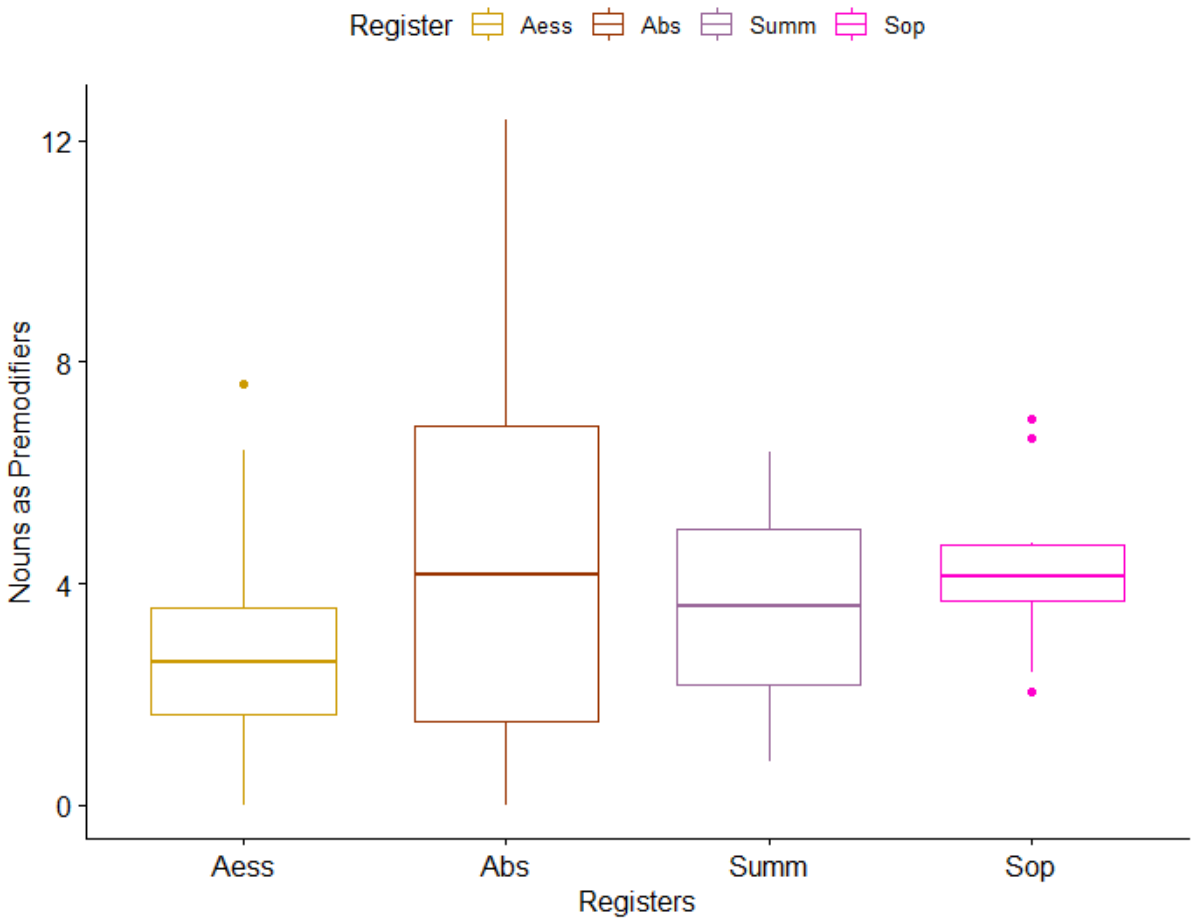
Excerpt 4.46 – Attributive adjectives in AEs 984

The voltage instability process can be a **short** *duration event* (ms to s), like a **short** *circuit*, or a **long** *duration event* (30s to 30min), like a on-load-tap-changer operation. In a heavily **loaded electrical** *system* where **large** *amount* of **active** and **reactive** *power* are transmitted over **long** *transmission lines* and in the absence of **reactive** power at the **receiving** *end*, a contingency like a line or generator outage, can lead to voltage collapse, resulted from a voltage instability condition in **weak** *areas* or **non secure** *buses*.

In the case of the SUM excerpt, attributive adjectives seem to have been clearly borrowed from the source text, such as *Great Depression* and *American people*. Swales (1994) proposes several guidelines concerning SUM writing, and one of them encourages students to "write down the key support points for the main topic (...)" (p. 106), which are generally constructions of attributive adjectives or nouns as premodifiers. As for AEs, in addition to having more occurrences of topic-related constructions (e.g., *reactive power*), AAs with more than one premodifier were found, such as *loaded electrical system*. This difference may be related to the registers' communicative purposes, besides the possibility of being influenced by students' proficiency level, as SUMs are typically written in IFA I, whereas AEs are written in IFA III.

Finally, nouns as premodifiers were the last feature analyzed per registers. Figure 4.10 and Table 4.9 below demonstrate the mean rates of occurrence for nouns as premodifiers by each register.

Figure 4.10 – Boxplot of nouns as premodifiers per register



Source: Prepared by the author, 2022.

Table 4.9 – Nouns as premodifiers results per register

Register	Count	Mean	SD	Median	IQR
Abstract	13	4.84	3.76	4.17	5.33
Argumentative Essay	14	2.95	2.04	2,59	1.93
Statement of Purpose	10	4.30	1.57	4.13	0.99
Summary	2	3.58	3.95	3.58	2.79

Source: Prepared by the author, 2022.

The register with the highest mean of nouns as premodifiers was the abstract, followed by the statement of purpose, summary, and argumentative essay. Nouns as premodifiers in this

analysis include both genitive nouns and nouns as premodifiers, with one modifier or more. In fact, the registers with more occurrences of genitive nouns were the ABs and SOPs. In contrast, the SUM register did not present genitive nouns.

Moreover, further analysis of genitive nouns indicates influences regarding the register's communicative purposes. Abstracts and argumentative essays presented genitive nouns related to random subjects, such as *fish's mortality*, and *world's population*, whereas statements of purpose presented genitive nouns related specifically to this register purpose of students' application, such as *master's program*, and *universities' evolution*, for example.

The same applies for nouns as premodifiers, as most of these in SOPs are influenced by the register's purposes, such as *graduate program, undergraduate program, research group, engineering field, Science Department*, and many more, whereas occurrences in ABs, SUMs and AEs refer to a variety of different specific topics, as *enzymes activities, mineral deposits, muscle degeneration, body health, power system*, etc. A similar result was already found by Queiroz (2019), concerning topic-specific nouns as premodifiers in essays.

Excerpts 4.47 and 4.48 show fragments of two texts of different registers with the highest occurrences of nouns as premodifiers. Curiously, the two texts with the highest frequency of nouns as premodifiers were written by the same student (726). Constructions are underlined, the head noun is in bold, and the premodifier noun is in italic.

Excerpt 4.47 – Nouns as premodifiers in ABs 1385

This paper presents the most recent developments of the authors' research center team regarding *power system voltage stability* **analysis**. It focus[es] on the conception of electrical *network* **equivalents** [,] especially those related to *voltage stability* **indexes**. The work aims to identify their differences and similarities in terms of mathematical basis and specific applications within electric *system* **activities**.

[] – added by the author.

Excerpt 4.48 – Nouns as premodifiers in AEs 984

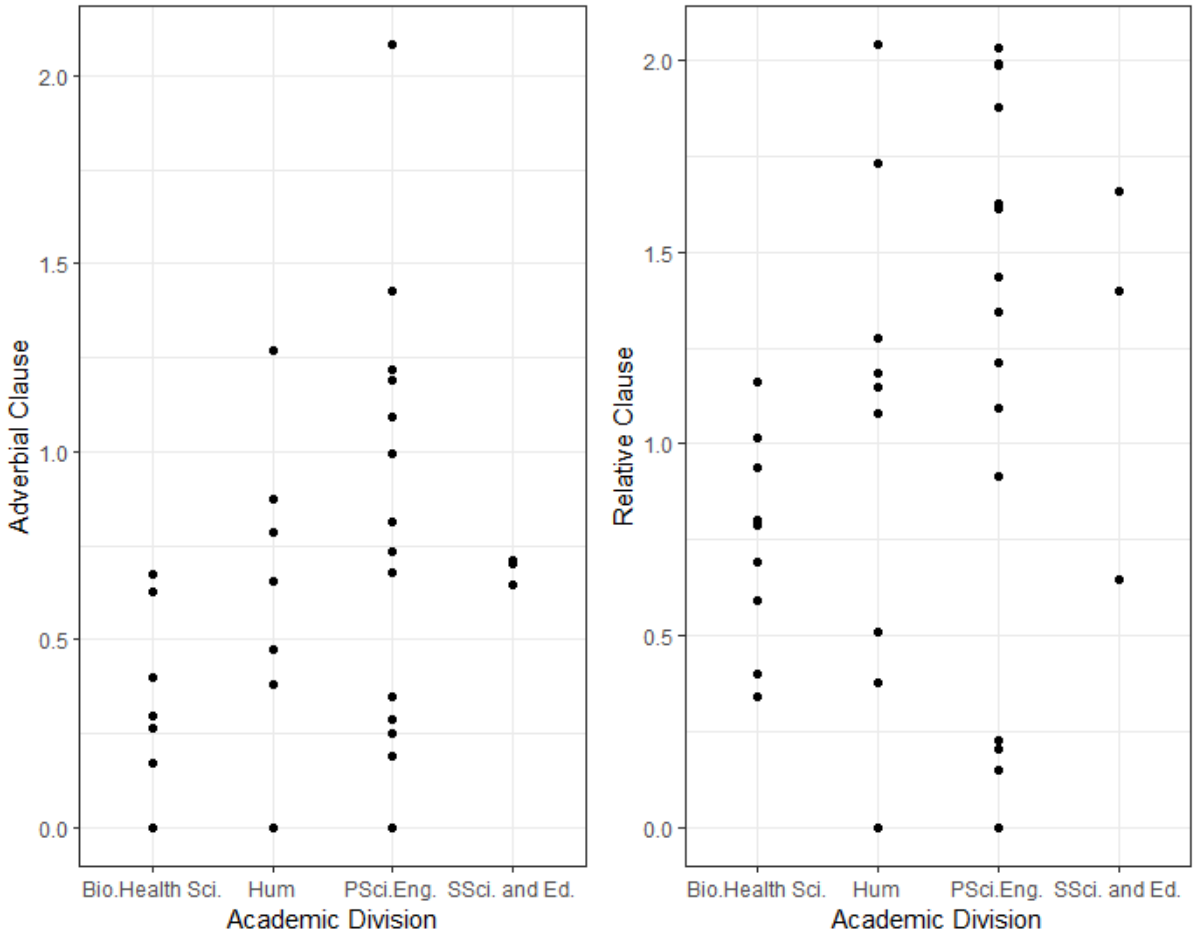
The *Electricity* **Agency** (ANEEL) defines that margin must be greater than 4% (*Stability* **Margin** – SM) in active *power* **transmission**, to keep the system in a safety* condition, with spare to supply power* in case of some occurrence. To reduce this risk of *voltage* **instability** and blackout, a lot of indices were investigated to assess the electrical system to detect weaknesses and monitor the *stability* **margins**.

*Wrong or badly positioned word.

4.2.2 Academic divisions

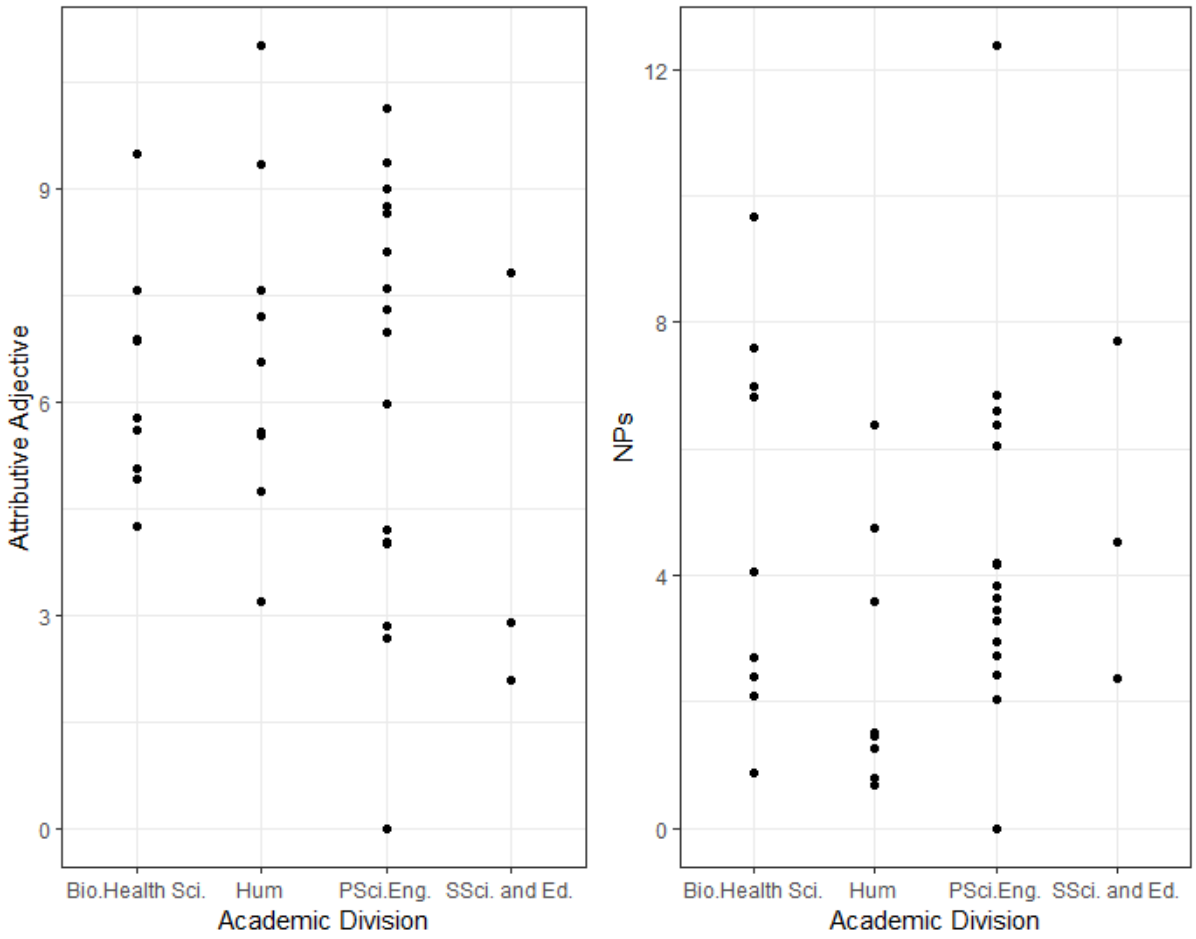
Similar to registers, academic divisions have been analyzed in grammatical complexity studies and were found to be important mediators of language development and variation, such as in Staples et al. (2016), and Biber et al. (2020). Therefore, this subsection discusses the findings regarding the four features selected for the analysis (adverbial clauses, ACs, relative clauses, RCs, attributive adjectives, AA, and nouns as premodifiers). Figures 4.11 and 4.12 below show the data distribution of clausal and phrasal features per academic division.

Figure 4.11 – Plot of adverbial clauses and relative clauses per academic division (rates of occurrence per 1,000 words)



Source: Prepared by the author, 2022.

Figure 4.12 – Plot of attributive adjectives and nouns as premodifiers per academic division (rates of occurrence per 1,000 words)

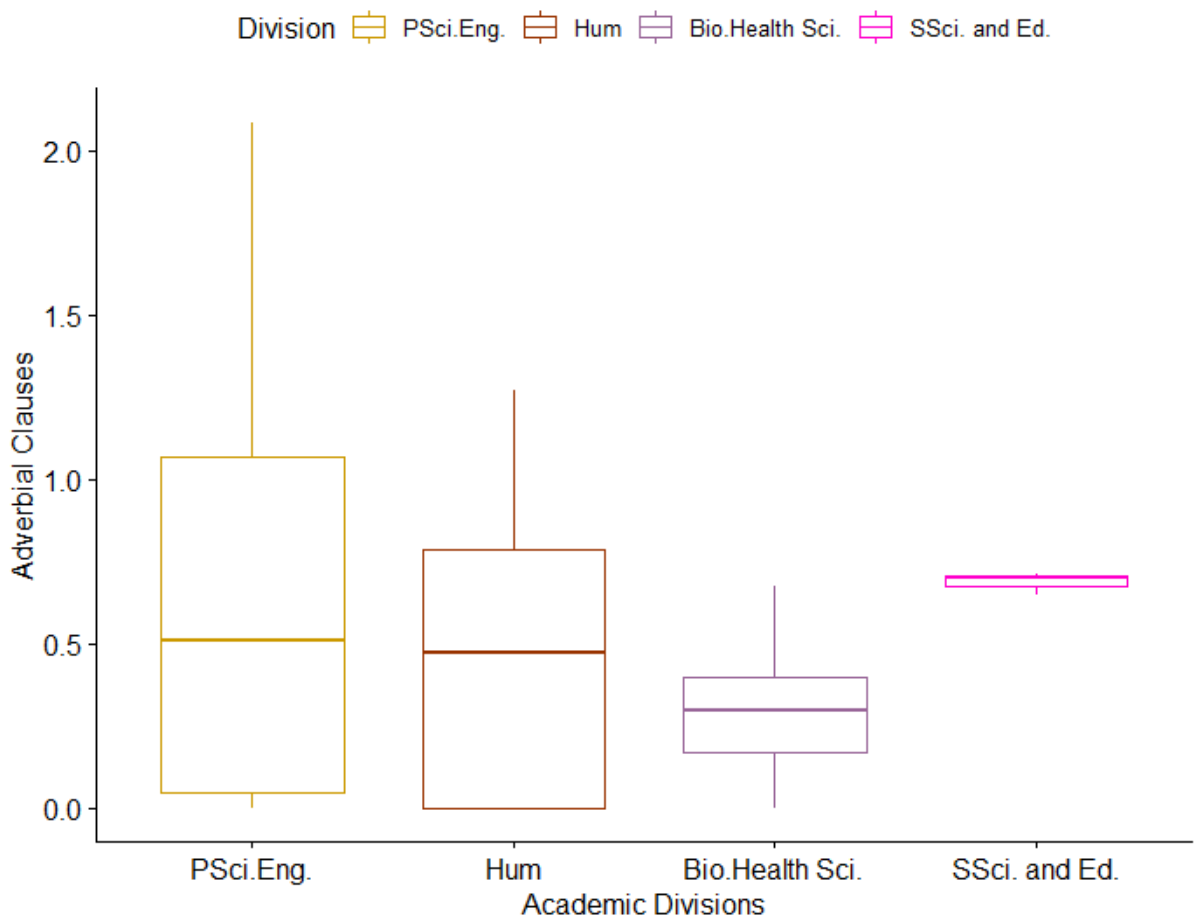


Source: Prepared by the author, 2022

According to plots in Figures 4.10 and 4.11 above, phrasal features were more frequent than clausal features in all academic divisions, although they ranged from text to text of course. Adverbial clauses showed more similar frequencies, with only one text differing from the others, and relative clauses had more spread frequencies across texts. Furthermore, attributive adjectives and nouns as premodifiers had also spread frequencies across texts. For a proper comparison, the mean, median, and IQR of each academic division will be considered, as samples' sizes vary largely from one academic division to the other.

Following the same organization from Subsection 4.2.1 above, we will start with adverbial clauses. Figure 4.13 presents the boxplot of ACs per academic division and Table 4.10, demonstrates the results of ACs per academic division.

Figure 4.13 – Boxplot of ACs per academic division



Source: Prepared by the author, 2022.

Table 4.10 – Adverbial clause results per academic division

Division	Count	Mean	SD	Median	IQR
Biological and Health Sciences	9	0.31	0.24	0.29	0.23
Humanities and Arts	9	0.49	0.44	0.47	0.78
Physical Sciences and Engineering	18	0.62	0.61	0.51	1.02
Social Sciences and Education	3	0.68	0.03	0.69	0.03

Source: Prepared by the author, 2022.

Finite adverbial clauses (ACs) showed similar means in all academic divisions, and the area of Social Sciences and Education (SSE) had the highest mean, followed by Physical Sciences and Engineering (PSE), Humanities and Arts (HA), and Biological and Health Sciences (BHS). The choice for subordinators varied across fields of study, which also influenced choices in the most used semantic categories of ACs. In SSE, the most frequent semantic category was the adverbial clause of time, with the subordinators *when* and *since*.

There was also one occurrence of a contingency adverbial of reason with the subordinator *because*, and one occurrence of a contingency adverbial of condition with the subordinator *if*. Excerpt 4.49 below presents an instance of an AC written by an SSE student.

Excerpt 4.49 – AC in ABs 1573 from SSE

Particularly, **when** you invest in a mutual fund you could access markets that you may not access investing individually.

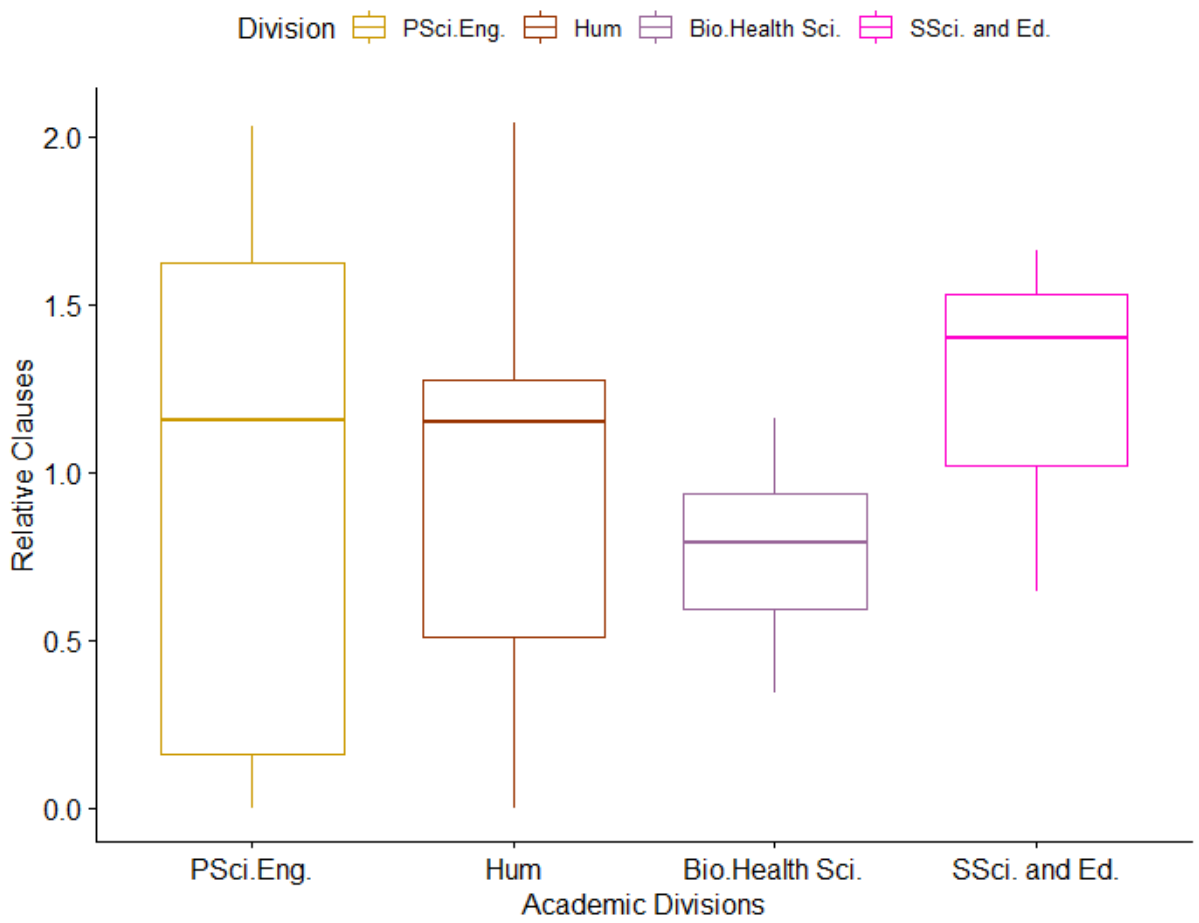
In Physical Sciences and Engineering, the three semantic categories with the highest number of occurrences were the contingency adverbial of condition, realized by the subordinator *if*; adverbial clauses of time, with the subordinators *when* and *since*, and ACs of concession, through the subordinators *although* and *though*. There were also occurrences of ACs of reason, with the subordinator *because*, and result, with the subordinator *so that*. In Humanities and Arts, the most frequent semantic categories of ACs were adverbials of reason, with the subordinator *because*, conditional, with the subordinator *if*, and time, with subordinators *since* and *when*.

Biological and Health Sciences was the division with the smallest mean in adverbial clauses; indeed, it was actually lower than half of the academic division with the highest mean. The most frequent semantic category was the adverbial clause of time, with the subordinators *when*, *since*, and *while*. There were also occurrences of adverbials of reason, with the subordinator *because*, and one occurrence of a contingency adverbial of condition, realized by the subordinator *if*. In addition to having presented the smallest rate of occurrences for the two clausal features (the relative clause discussion is as follows), BHS texts presented a high rate of occurrence for the two phrasal features analyzed, which is addressed below.

Finally, although some divisions showed more occurrences of adverbial clauses than others, the overall means were all very low, indicating a low reliance on this feature across academic divisions. Even so, the most frequent semantic categories of ACs across divisions, except for PSE, are not completely in accordance with Longman Grammar (BIBER et al., 1999). Indeed, it states that purpose and condition clauses are the most frequent in academic prose, and ACs of time were the most frequent in our subcorpus.

The second clausal feature analyzed was the relative clauses. Figure 4.14 and Table 4.11 below show the rates of occurrence of RCs per academic division.

Figure 4.14 – Boxplot of RCs per academic division



Source: Prepared by the author, 2022

Table 4.11 – Relative clause results per academic division

Division	Count	Mean	SD	Median	IQR
Biological and Health Sciences	9	0.74	0.27	0.78	0.34
Humanities and Arts	9	1.04	0.64	1.15	0.76
Physical Sciences and Engineering	18	0.98	0.80	1.15	1.46
Social Sciences and Education	3	1.23	0.52	1.40	0.50

Source: Prepared by the author, 2022.

Relative clauses (RC) were the most frequent clausal feature employed in all divisions. In this analysis, we consider both types of relative clauses, namely *that* and *wh*. The division with the highest RC mean was Social Sciences and Education, followed by Humanities and Arts, Physical Sciences and Engineering, and Biological and Health Sciences. In SSE, *that*-relative clauses were more frequent than *wh*-relative clauses. Moreover, all instances of relative

clauses with the relativizers *that* and *which* referred to inanimate head nouns, while all instances of *who* referred to animate head nouns.

In PSE, *wh*-relative clauses were more frequent than *that*-relative clauses. In *wh*-clauses, the relativizer *which* was the most frequent, followed by *who* and one instance of *why*. All instances of *who* were referring to animate head nouns, except for one instance with the antecedent inanimate *companies*. This case was already discussed in subsection 4.2.1 above, in the discussion on RCs across registers. There were other occurrences with the antecedent *companies* in PSE texts, and all the other RCs started with the relativizer *that*.

In HA, *that*-relative clauses were more frequent than *wh*-relative clauses, and in *wh*-clauses, the relativizer *which* was the most employed, followed by *who*. All instances of *who* referred to an animate antecedent, even though not always being the closest word. Excerpt 4.50 below demonstrates this occurrence. The RC is underlined, the antecedent is in bold and the noun before the RC is in italic.

Excerpt 4.50 – RC in ABs 1573 from HA

Second, it is based on **Barthes** *theory* who not only created the dialogism between the texts, but also, predicted the death of the author in favor of an independence of the symbols.

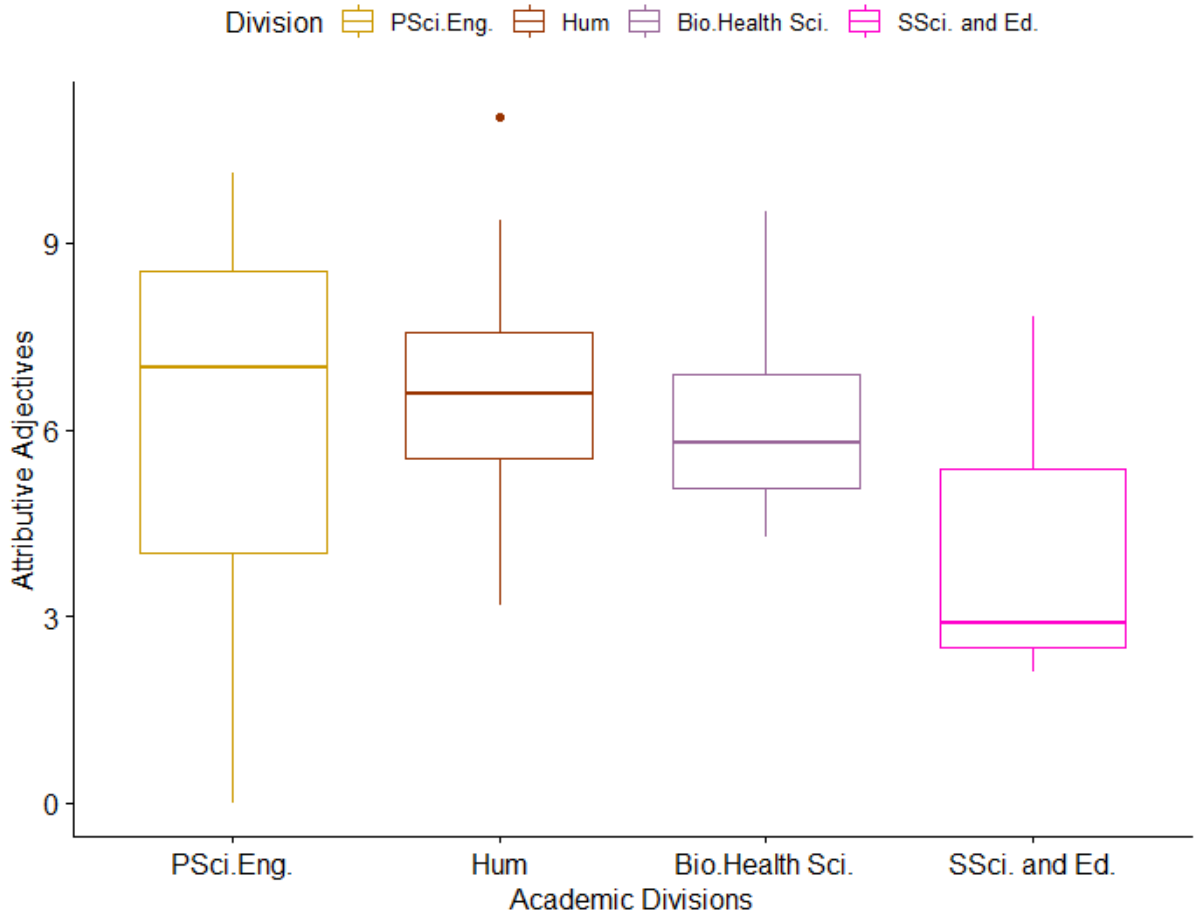
BHS was the division with the smallest number of RC occurrences, according to the mean displayed in Table 4.11. Moreover, it had more instances of *wh*-relative clauses than *that*-relative clauses. The relativizers of *wh*-clauses were *which, who,* and *where*. Although having presented few occurrences of this feature, all relativizers were correctly employed, as all instances of *who* referred to animate head nouns, all instances of *where* referred to a place, and all instances of *which* and *that* were properly employed according to student's choice (as they are both used interchangeably, with some stylistic preferences already mentioned in subsection 4.2.1 above).

Regarding the phrasal features, the attributive adjective was the feature with the highest means in almost all academic divisions, second only to nouns as premodifiers in SSE. This is an interesting outcome since AAs were the most frequent feature in all variables (i.e., time and register) in the longitudinal analysis and the quasi-longitudinal analysis of registers. Therefore, the preference for nouns as premodifiers in the academic division of SSE indicates that these

students rely on features more found in specialized academic writing and proficient texts (PARKINSON; MUSGRAVE, 2014).

Furthermore, HA was the division with the highest mean of attributive adjectives, and the smallest means of nouns as premodifiers (as it will be presented below), which may indicate that HA students have a strong preference for AAs or are still in the first stages of complexity development. Figure 4.15 and Table 4.12 below illustrate this statement.

Figure 4.15 – Boxplot of AAs per academic division



Source: Prepared by the author, 2022.

Table 4.12 - Attributive adjective results per academic division

Division	Count	Mean	SD	Median	IQR
Biological and Health Sciences	9	6.28	1.61	5.79	1.83
Humanities and Arts	9	6.75	2.39	6.57	2.04
Physical Sciences and Engineering	18	6.09	2.88	6.99	4.52
Social Sciences and Education	3	4.27	3.10	2.90	2.86

Source: Prepared by the author, 2022

The frequency of attributive adjectives in HA with a single modifier account for 83.07% of occurrences, compared to merely 16.93% occurrences with more than one premodifier. Besides presenting AAs of various semantic domains, some occurrences were recurrent (more than three occurrences), such as the affiliative *American people,* and the relational *introverted people.* The repletion of these constructions indicates a strong topic influence, which, of course, is related to academic division. The most frequent semantic domains found in HA texts were topical (e.g., *astronomical, auriferous, scientific, professional*), relational (e.g., *whole, introverted, similar, different, specific),* affiliative (e.g., *American, Brazilian, Portuguese, European*), descriptors of size (e.g., *short,*), time (e.g., *modern, new, recent, early*), and evaluative (e.g., *nice, great, ideal, effective*).

The second highest mean of attributive adjectives was found in Biological and Health Sciences texts. Of all the instances of AAs, 62.37% accounts for AAs with a single modifier, whereas 37.63% accounts for AAs with multiple modifiers. Topical AAs were the most frequent, such as *academic, alcoholic, aquatic, biological, computational, acute, airborne, alkaline,* and many more. There was also a strong reliance on relational AAs, such as *single, similar, higher, different, basic, whole,* and *common*. Descriptors were very frequent, such as evaluative AAs, *best, good, great, key, severe,* and *super,* size/amount, *bigger, lower, high, large,* and *short,* and time, *new*.

From all occurrences of attributive adjectives in the PSE division, AAs with only one premodifier had the third-highest mean (67.62%), whereas 32.38% of AAs had more than one premodifier. In addition, there were extremely repetitive AAs, mostly from the following semantic domains: topical (i.e., more than two occurrences: *electric, electrical, federal, mechanical, operational, professional, social, aerospace, academic,* and many others). Besides the heavy reliance on topical AAs, there were many instances of descriptors, such as size/amount (e.g., *big, great, high, huge*), *evaluative* (e.g., *best, good, important),* time (e.g., *new, early*), and *color* (e.g., *green, white*). Some relational AAs were also found, as *different, main, specific, various,* and more.

SSE was the division with the smallest mean in attributive adjectives. Within AA occurrences, 82.22% accounts for AAs with a single modifier, whereas 17.78% accounts for AAs with more than one modifier. Topical AAs were not as frequent as relational and evaluative AAs, as there were only a few topical constructions, but used with more diverse adjectives,

such as *academic, computational, empirical, financial, popular, professional,* and *transactional*. There were many occurrences of relational AAs with the adjective *mutual*, which is probably related to the topic of the texts since they were all followed by the noun *fund*, indicating an influence of jargon used in the discipline in question.

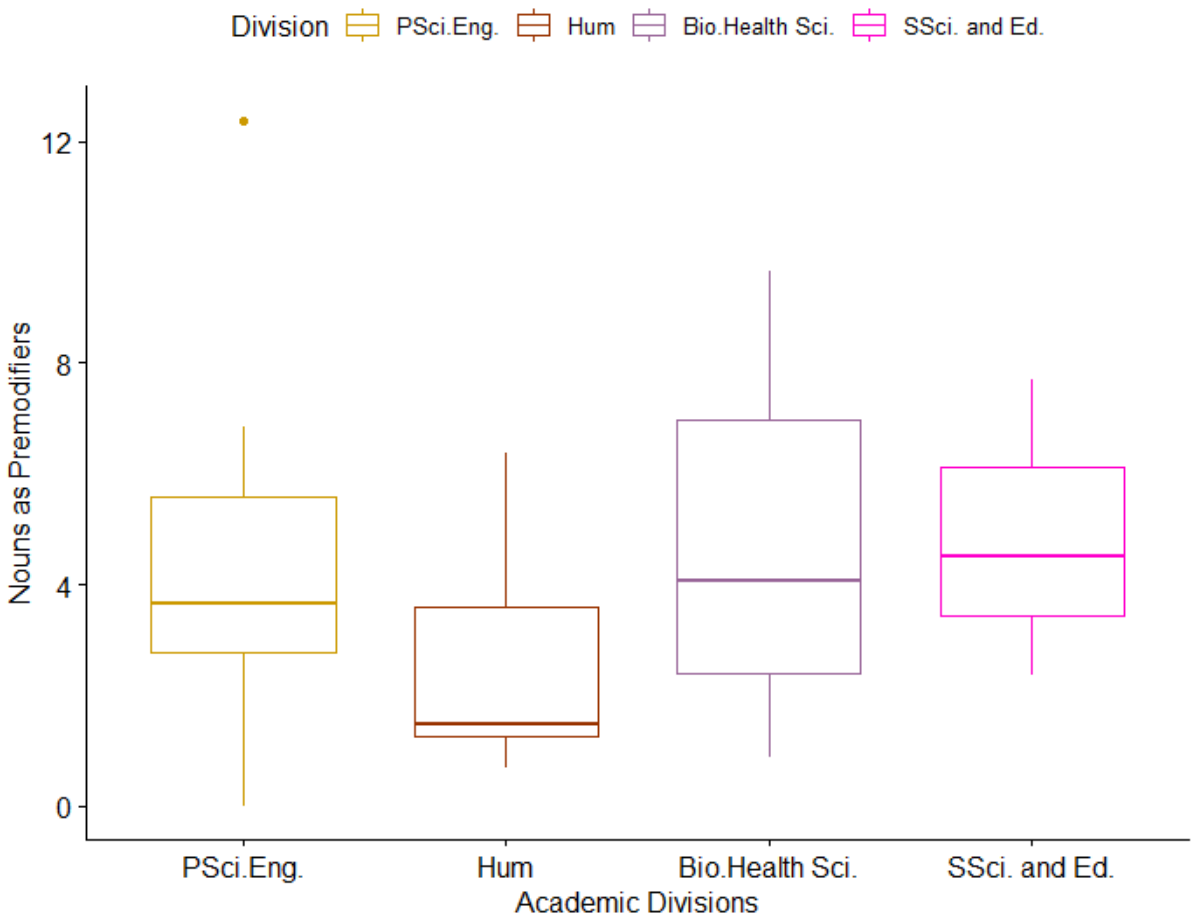
Other relational attributive adjectives were written with the adjectives *main, individual, higher,* and *different*. There were also some evaluative adjectives, such as *good* and *best*, and descriptors of size, such as *big*. Excerpt 4.51 below presents the part of a text with a high occurrence of AAs. The whole AA construction is underlined, and the adjectives are in bold.

Excerpt 4.51 – AAs in ABs 1385 from PSE

In a heavily loaded **electrical** system where **large** amount of **active** and **reactive** power are transmitted over **long** transmission lines and in the absence of **reactive** power at the **receiving** end, a contingency like a line or generator outage, can lead to voltage collapse, resulted from a voltage instability condition in **weak** areas or **non secure** buses.

Finally, the last feature analyzed for variations in academic divisions was nouns as premodifiers. Figure 4.16 and Table 4.13 below show the mean rates of occurrence per academic division.

Figure 4.16 – Boxplot of nouns as premodifiers per academic division



Source: Prepared by the author, 2022.

Table 4.13 – Nouns as premodifiers results per academic division

Division	Count	Mean	SD	Median	IQR
Biological and Health Sciences	9	4.80	3.04	4.06	4.59
Humanities and Arts	9	2.43	2.00	1.48	2.33
Physical Sciences and Engineering	18	4.15	2.84	3.64	2.79
Social Sciences and Education	3	4.86	2.68	4.52	2.66

Source: Prepared by the author, 2022.

Nouns as premodifiers were the second most frequent feature in almost all divisions, and the most frequent feature in the Social Sciences and Education division, as it was introduced in the discussion above, about attributive adjectives. Interestingly, the academic division that had the smallest mean of nouns as premodifiers also had the highest mean in attributive adjectives (Humanities and Arts). As previously stated, this may be caused by academic divisions particularities but can also mean that the HA students are less proficient in academic

writing than Biological and Health Sciences students, corroborating the findings by Parkinson and Musgrave (2014).

In this analysis, both types of nouns as premodifiers are considered, namely genitive nouns and nouns as premodifiers (i.e., with single and multiple modifiers). Although SSE, presented the highest mean in nouns as premodifiers, it did not show any occurrences of a genitive noun as a premodifier. Furthermore, nouns with more than one premodifier were almost as frequent as nouns with only one premodifier, as it presented 48.57% against only 51.43% of nouns with a single nominal premodifier.

Moreover, the most frequent (more than three occurrences) nouns were *investment* (e.g., *portfolio, funds), manager (*e.g., *characteristics),* and *portfolio (*e.g., *impact, turnover).* As can be seen, all nouns as premodifiers are clearly related to the SSE academic division, as they all refer to topics of this field such as *data management, finance statements, bank account, administration fees*, and many more.

The second highest mean was found in BHS texts. Contrary to the first highest mean in SSE, BHS had six occurrences of genitive nouns as premodifiers, such as *fish's growth*, and *survival's tax*. Nouns with a single modifier account for 58.04% of all occurrences of premodifying nouns, whereas nouns with multiple modifiers account for 41.96%. In addition, the most frequent nouns as premodifiers (rate over three occurrences) found in BHS texts were: *ethanol (*e.g., *group, treatment), feed (*e.g., *conversion), fishing (*e.g., *company, farms, industry),* and *penaeid (*e.g., *shrimp).* Aside from these, there were not as many repetitive nouns as premodifiers nor repetitive constructions.

In PSE, there were only two instances of genitive nouns as premodifiers: *master's program,* and *researcher's interest*. The number of nouns with more than one premodifier was extremely close to the number of nouns with only one premodifier, 49.83%, against 50.17%. The most frequent nouns as premodifiers (rate over three occurrences) found in PSE texts were *computer (*e.g., *science, programming), energy (*e.g., *conservation, source), graduate (*e.g., *program, students), greenhouse (*e.g., *effect, gas), life (*e.g., *span), measurement (*e.g., *data, errors), quality (*e.g., *manager),* and *voltage (*e.g., *collapse, instability).* In fact, some of the constructions were highly repetitive, which indicates the influence of topics across different texts.

In HA, the division with the smallest mean in this feature, there were three genitive nouns as premodifiers, two had a single nominal premodifier, *French's method*, and *world's population,* and the other one had multiple premodifiers, namely *reader's collaborative role.* Furthermore, this academic division presented fewer nouns with multiple premodifiers (31.75%), in comparison to nouns as single premodifiers (68.25%).

There were not many repetitive combinations, nor many repetitive premodifying nouns. Some of the most frequent premodifying nouns (rate over two occurrences) in HA texts were: *computer (*e.g., *science, development), labor (*e.g., *market), reader (*e.g., *analysis, behavior), search (*e.g., *project), telecommunications (*e.g., *engineering, plants)*. According to these examples, nouns as premodifiers in the HA division seem to be more topic-related than discipline-related, since some constructions, such as the ones presented above, do not relate to the student's field (e.g., *computer science*) (Chapter 3 presents a broad explanation of this grouping per academic divisions).

Since nouns as premodifiers are rather compressed structures containing a heavy load of information, the majority of nouns as premodifiers found in all academic divisions' texts are closely related to the student's field, as in the case of *computer science, penaeid shrimp,* and *greenhouse effect,* for instance. The same happens with attributive adjectives since many names of disciplines, topics, or places are created from these structures. Excerpt 4.52 below presents the part of a text with a high frequency of nouns as premodifiers. Constructions are underlined, head nouns are in bold, and adjectives are in italic.

Excerpt 4.52 – NPs in ABs 1385 from PSE

This paper presents the most recent developments of the **authors'** research center team regarding power system voltage stability **analysis**. It focus[es] on the conception of *electrical* network **equivalents** [,] especially those related to voltage stability **indexes**. The work aims to identify their differences and similarities in terms of mathematical basis and specific applications within electric system **activities**.

[] – added by the author.

4.3 Comparison with a native corpus

To understand the extent of grammatical complexity variation between learners and natives, and to answer research hypothesis H5, results found in our learner subcorpus were compared to results found by Staples et al. (2016), who analyzed the L1 subset of BAWE.

It is important to point out some differences between Staples et al. and this master's thesis. First, their study is cross-sectional, not longitudinal, which means that they gathered groups of students at the same point in time and compared these groups across different levels. Second, their cross-sectional analysis divided levels by years of study, such as "first-year undergraduate, second-year undergraduate, final-year undergraduate, and graduate" (STAPLES et al., 2016, p. 155), whereas, in our research, we divided the time frame into six-month increments (Time 1, Time 2, and Time 3).

Third, they analyze texts from a genre perspective, instead of a register one, as is the case herein. This type of analysis is similar to the register perspective, as they both include "description of the purposes and situational context of a text variety." However, a genre analysis focuses more on the "conventional structures used to construct a complete text within the variety" (BIBER; CONRAD, 2009, p. 2). Besides, the genres in their analysis are quite different from the registers of the subcorpus analyzed in this thesis, as the authors included *critiques, case studies,* and *explanations* as well. Thus, the comparison between genres and registers will only consider essays, as they are the only genre/register in common between the two studies.

Finally, their subcorpus is much larger than ours (1,103 texts). Thus, statistical results will not be compared, as our subcorpus has only 39 texts. Instead, we will focus on the discussion of the findings.

On the other hand, disciplines in their study are grouped into primary divisions, similarly to this thesis, as *Arts and Humanities (AH), Social Sciences (SS), Life and Physical Sciences (LPS).* In our study, we divided between Biological and Health Sciences (BHS), Humanities and Arts (HA), Physical Sciences and Engineering (PSE), and Social Sciences and Education (SSE). Therefore, the comparison across academic divisions will be more accurate.

Moreover, they analyzed all 23 features from the Developmental Index, divided by phrasal, clausal, and intermediate features (linking adverbials, relative clauses, and non-finite clauses), which enables a complete comparison with the 18 features analyzed here, divided per phrasal (which includes linking adverbials), non-finite (includes non-finite clauses) and finite clausal features (include both finite and relative clauses). They hypothesized that

> for L1 writers (...), phrasal complexity develops most noticeably during university years, much later than researchers have normally considered (p. 154). (...) [and] the overall hypothesized trend was for writers to show movement away from finite

> dependent clauses toward nonfinite dependent clauses and then to dependent phrases (STAPLES et al., 2016, p. 162).

According to their findings, phrasal features, such as nouns as premodifiers, attributive adjectives (AA), and prepositional phrases (PP), all increased from Level 1 to Level 4. Nouns as premodifiers presented a steady increase from Level 1 to Level 4, while AAs also increased from Level 1 to Level 4, but more radically from the third to the fourth level. *Of* phrases (OP) as post-modifiers presented a small decrease from Level 1 to Level 2, but then steadily increased until Level 4. In our subcorpus, all phrasal features increased from Time 1 to Time 3 as well, but some PPs, such as PPs as adverbials, PP with concrete/locative meanings, and PPs with abstract meanings slightly decreased from Time 1 to Time 2. In contrast, OPs steadily increased from Time 1 to Time 3.

Clausal features in their native subcorpus, such as finite adverbial clauses (AC), verbs + *that*-complement clauses, and verbs + *wh*-complement clauses all decreased from Level 1 to Level 4. ACs presented a steadily decline, while *that* and *wh*-complement clauses had a huge decrease from Level 3 to Level 4, attesting their research hypothesis, that "development may occur at particular points during university study, or may accelerate at specific times" (STAPLES et al., 2016, p. 164).

In our learner subcorpus, the frequency of adverbial clauses increased from Time 1 to Time 3 but decreased from Time 1 to Time 2. *That* and *wh*-complement clauses were very diverse, as the complement clauses controlled by common verbs presented an increase from Time 2 to Time 3, and complement clauses controlled by other verbs presented a decrease from Time 1 to Time 3 and an increase from Time 2 to Time 3. This variation in our findings follows the research hypothesis of Staples et al. (2016), as the frequency of some features, such as the ones mentioned in this paragraph, did not follow a progression line, but shifted over time.

Intermediate features in their subcorpus, such as linking adverbials, and RC (*that* and *wh*), all presented a decrease, most specifically from Level 3 to Level 4. The use of *wh*-relative clauses increased from Level 1 to Level 2, and linking adverbials increased steadily from Level 1 to Level 3. Other intermediate features, such as non-finite complement clauses controlled by adjectives and nouns showed no variation across levels.

In our subcorpus, the use of linking adverbials by learners steadily increased from Time 1 to Time 3, besides varying in the use of such devices, which signals a certain degree of lexical development. The frequency of t*hat* and *wh*-clauses also increased from Time 1 to Time 3, but

wh-clauses decreased from Time 1 to Time 2. This may have happened because *that* is usually the most preferred relativizer by learners, due to its unmarked characteristic. All non-finite clauses increased in frequency over time, except for the non-finite complement clauses controlled by common verbs, which decreased over time.

Across genres, their results show that essays relied heavily on clausal features, such as finite adverbial clauses, and *wh*-relative clauses, in the scope of other genres. In our register analysis, essays also presented a higher frequency of ACs and *wh*-relative clauses, in contrast to the other registers. This result may indicate that essays accept a higher frequency of clausal features compared to other registers; this, in turn, means that this can be a characteristic of this register, as a great frequency of this type of feature was found in both Staples et al. (2016) examination of L1 English and our study about Brazilian L2 learners of English.

Across disciplines, their results show variations in the use of features in specific disciplines. In the LPS texts, there were more occurrences of nouns as premodifiers, whereas in SS and AH, attributive adjectives were more frequent. In our subcorpus, the only academic division that heavily relied on nouns as premodifiers in contrast to AAs was the SSE. Similarly, in the results found by Staples et al. (2016), the HA division had the highest frequency of AAs and the smallest frequency of nouns as premodifiers, in contrast to the other academic divisions. This interesting outcome strongly points out to the hypothesis that AAs are preferred by HA students, differently from the previous thought expressed in Subsection 4.2.2, that perhaps HA students were less proficient writers than the other academic divisions, in compliance with Parkinson and Musgrave (2014). However, to attest to this hypothesis, more studies on this subject are necessary.

Finally, in their subcorpus, the clausal features, such as finite clauses, were more frequent in AH, and less frequent in LPS. *Wh* relative clauses were also more frequent in AH. In our subcorpus, ACs were more frequent in SSE, and *wh*-relative clauses in PSE, although relative clauses in general (*that* and *wh*) were more frequent in SSE.

4.4 Overall Discussion of Findings and Hypotheses

This chapter discussed the subcorpus findings from two main perspectives: longitudinal and quasi-longitudinal. After that, we proceeded to compare the results found with the results of the grammatical complexity analysis of native L1 English by Staples et al. (2016). These steps were taken to confirm research hypotheses numbers 1 to 5, as follows:

Hypotheses:

1. Over time, students will increase the use of phrasal features from the second stage onwards and decrease the use of finite and non-finite clausal features from the first and second stages onwards.
2. Over time, students will follow the hypothesized developmental index, thus showing an increase in the use of features from the later stages.
3. There will be variations in the use of certain features across registers.
4. There will be variations in the use of certain features across academic divisions.
5. There will be differences between the texts written by Brazilian and British university students in the scope of the development of certain complexity features.

The statistical results of the longitudinal analysis confirm Hypothesis H1 and H2 entirely, since all features with statistical significance ($p < .05$) variation either increased or decreased according to what was expected. However, the discourse qualitative analysis often showed an increase in clausal features of stages 1 and 2, partly confirming Hypothesis H1. Furthermore, the discourse qualitative analysis partially supports Hypothesis H2, as the frequency of almost all features, whether phrasal, finite, or non-finite from the third stage increased over time, although sometimes not steadily (e.g., PP with abstract meanings).

Moreover, together with the quantitative increase of some phrasal features, our students showed a degree of lexical development in relation to these features, such as attributive adjectives, adverbs as adverbials (linking and stance adverbials), and nouns as premodifiers over time. Frequencies of these constructions were quite diverse across points of time, especially in Time 3, which presented a lot of first occurrences of these constructions (e.g., the occurrence of the linking adverbials *nevertheless* and *thus* only in Time 3).

The quasi-longitudinal analysis confirms Hypotheses 3 and 4, as there were variations in the use of some features across registers and academic divisions. Although the variables register and academic divisions did not present statistically significant ($p < .05$) variation, a discourse analysis demonstrated that different academic registers and different academic divisions do not have the same preference for features (e.g., summary had the highest frequency of attributive adjectives, while abstract had the highest frequency of nouns as premodifiers).

Hypothesis H5 was also confirmed by our analysis, as learners and natives present differences in grammatical complexity development, such as a decrease in the use of clausal features by natives over time, whereas learners increased the frequency of some clausal features over time. Learners also relied more on linking adverbials over time compared to natives, who showed a decrease in the use of this feature over time. Nevertheless, there were very interesting similarities that can contribute to future research. The first one involves the heavy reliance on clausal features in essays in both L1 and L2 students' texts. The second refers to the preference of attributive adjectives in texts of students from the Humanities and Arts division of both L1 and L2 subcorpora.

5 – CONCLUSION

This thesis analyzed grammatical complexity development in an EAP corpus of Brazilian learners of English texts, through a longitudinal perspective. The analysis was grounded on the Developmental Index theory proposed by Biber et al. (2011), which states that students progressively acquire grammatical complexity features over five different stages in academic writing. Therefore, this process can be very long, as some features are acquired relatively late (BIBER et al., 2011; STAPLES et al., 2016).

To this end, analyses were divided into two perspectives, namely longitudinal and quasi-longitudinal. The primary purpose of the longitudinal analysis was to evaluate the development of grammatical complexity by the same students (n = 13) over a three-time period. All the students attended the same EAP courses (IFA courses) for one year and a half (not necessarily together but in different semesters and years). During this period, each student wrote three different texts of different registers six months apart. Therefore, we proceeded to perform a quasi-longitudinal analysis as well, to check whether registers and/or academic divisions were predictors of linguistic variation.

Overall findings for the longitudinal analysis, from Time 1 to Time 3, indicate students' development, especially in relation to phrasal features, such as the attributive adjectives. However, some clausal features that were expected to decrease over time – such as the finite complement clauses controlled by common verbs and the finite adverbial clauses – increased in frequency, even though not statistically. Nevertheless, students relied much more on phrasal features than clausal features since Time 1, indicating that they were already familiar with phrasal devices in academic writing.

In addition, some features did not present a steady increase or decrease, as Time 2 frequently showed the smallest means across all times. This finding corroborates the statement by Biber et al. (2011) that development is a slow process, and the conclusion by Staples et al. (2016) that development may occur at specific points in time. The qualitative analysis has revealed that, besides the increase in the use of some features over time, particular features varied as well, indicating lexical development, most specifically in the scope of attributive adjectives, linking adverbials, nouns as premodifiers, adjectives in extraposed constructions, and PP as postmodifiers, since Times 2 and 3 often presented more varied types than Time 1.

This corroborates the statement by Gray et al. (2019) regarding the importance of a discourse qualitative analysis in a grammatical complexity study.

The results of the quasi-longitudinal analysis demonstrate variation in the preference of some features per registers and academic divisions. Clausal features were the least ones used in all registers, besides presenting no variation in the order of the mean rates of occurrences per register, since the order was the same in both adverbial clauses and relative clauses, with the argumentative essay ranking in the first place. In turn, academic divisions presented variations in the mean rates of occurrence of clausal features, except for the fields of Biological and Health Sciences, which had the smallest mean for both features.

Phrasal features were the most frequent across registers, especially attributive adjectives, which were the most frequent feature in all registers. Discourse analysis also demonstrated a variation in the lexical preference for attributive adjectives across registers, and a high reliance on topical, relational, evaluative, and affiliative adjectives. This variation can be influenced both by the registers' communicative purposes (e.g., ***American*** *people* in summaries) and text topics (e.g., ***reactive*** *power* in argumentative essays).

Nouns as premodifiers were very frequent, although they were not as frequent as attributive adjectives across registers. Interestingly, essays were the register with the smallest rate of occurrence for this feature but with a high reliance on attributive adjectives. A discourse analysis pointed out variations in the lexical preference of nouns as premodifiers across registers, similar to the attributive adjectives. Lexical variations are influenced both by registers' communicative purposes (e.g., ***graduate*** *program* in statements of purpose), and text topics (e.g., ***muscle*** *degeneration* in argumentative essays).

Furthermore, in academic divisions, clausal features were not as frequent as phrasal features. Once again, the attributive adjective stands out, since it was the most frequent feature in almost all academic divisions. The only academic division that relied more heavily on nouns as premodifiers than attributive adjectives was Social Sciences and Education. Furthermore, this academic division also showed the smallest mean in attributive adjectives. For the Humanities and Arts division, there was the opposite finding, as attributive adjectives were highly preferred, in contrast to nouns as premodifiers, in which this division had the smallest means.

Besides demonstrating variation in the development of L1 and L2 students, comparisons to a grammatical complexity study of a native corpus also shed light on some of our findings. The most interesting ones were that L1 essays also presented a great frequency of clausal features, in comparison to other registers, which may suggest a register's preference; and the academic division of Humanities and Arts in L1 texts also relied on attributive adjectives, in contrast to nouns as premodifiers, which may suggest that this feature is typical of this academic division and does not necessarily imply students lack of proficiency.

LIMITATIONS OF THIS STUDY

This study has some limitations regarding its methodology; therefore, we recommend that results be interpreted with caution and not generalized. The first limitation involves the number of participants (n = 13) and data (n = 39), which may not suffice for the statistical analysis comparison. The second one relates to the timeframe chosen for the longitudinal analysis, since the six-month increment may be too short to adequately capture the students' progress since the features often increased only from Time 1 to Time 3, which adds to a one-year difference.

Nevertheless, we still hope that our research will contribute to the body of literature in the field, particularly the development of grammatical complexity, longitudinal corpus-based studies, academic writing of Brazilian L2 learners of English, and, on a general basis, academic writing in EAP university students. We also hope that our study shed light on future studies, especially concerning the students' development of grammatical complexity over a longer timeframe, register specificities, such as the preference for clausal features in argumentative essays, and disciplinary specificities, such as the high reliance on attributive adjectives in Humanities and Arts.

REFERENCES

ANSARIFAR, A.; SHAHRIARI, H.; PISHGHADAM, R. 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, L. **Antconc** (version 3.5.8) [software]. Waseda University, 2019.

BARRON, A. Using corpus-linguistic methods to track longitudinal development: Routine apologies in the study abroad context. **Journal of Pragmatics**, v. 146, p. 87-105, 2019.

BEERS, S.; NAGY, W. Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?. **Reading and Writing**, v. 22, n. 2, p. 185-200, 2009.

BESTGEN, Y.; GRANGER, S. Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In: HOFFMANN, S.; SAND, A.; ARNDT-LAPPE, S.; DILLMANN, L. (eds.). **Corpora and lexis**, v. 81, n. 81. Leiden, The Netherlands: Brill, 2018. p. 277-301.

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. **Linguistics,** v. 23, n. 2, p. 337-360, 1985.

BIBER, D. Spoken and written textual dimensions in English: Resolving the contradictory findings. **Language**, v. 62, n. 2, p. 384-414, 1986.

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1991.

BIBER, D. Register as a predictor of linguistic variation. **Corpus linguistics and linguistic theory**, v. 8, n. 1, p. 9-37, 2012.

BIBER, D.; CONRAD, S. **Register, genre, and style**. Cambridge: Cambridge University Press, 2019.

BIBER, D.; GRAY, B. Discourse characteristics of writing and speaking task types on the TOEFL ibt® test: A lexico-grammatical analysis. **ETS Research Report Series**, v. 2013, n. 1, p. i-128, 2013.

BIBER, D.; GRAY, B. **Grammatical complexity in academic English**: Linguistic change in writing. Cambridge: Cambridge University Press, 2016.

BIBER, D.; GRAY, B.; POONPON, K. 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, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. **Longman grammar of spoken and written English**. Harlow: Longman, 1999.

BIBER, D.; REPPEN, R.; STAPLES, S.; EGBERT, J. Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. **International Journal of Learner Corpus Research**, v. 6, n. 1, p. 38-71, 2020.

BULTÉ, B.; HOUSEN, A. Defining and operationalising L2 complexity. In: HOUSEN, A.; KUIKEN, F.; VEDDER, I. **Dimensions of L2 performance and proficiency**: Complexity, accuracy and fluency in SLA. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2012, p. 23-46.

CHAU, M. **From language learners to dynamic meaning makers**: a longitudinal investigation of Malaysian secondary school students' development of English from text and corpus perspectives. 2015. Tese (Doutorado) - Department of English Language and Applied Linguistics, School of English, Drama, and American and Canadian Studies, College of Arts and Law, University of Birmingham, Birmingham, 2015.

COUNCIL OF EUROPE. COUNCIL FOR CULTURAL CO-OPERATION. EDUCATION COMMITTEE. MODERN LANGUAGES DIVISION. **Common European framework of reference for languages**: Learning, teaching, assessment. Cambridge: Cambridge University Press, 2001.

CRESWELL, J. Revisiting mixed methods and advancing scientific practices. In: HESSE-BIBER, S.; JOHNSON, R. **The Oxford handbook of multimethod and mixed methods research inquiry**, 2015.

CROSTHWAITE, P. Does EAP writing instruction reduce L2 errors? Evidence from a longitudinal corpus of L2 EAP essays and reports. **International Review of Applied Linguistics in Language Teaching**, v. 56, n. 3, p. 315-343, 2018.

DUTRA, D.; NUNES, L.; ORFANÒ, B.; ARRUDA, C. Institutional internationalisation through academic literacies in English: teaching and learning written genres in the Brazilian higher education context. **The Especialist**, v. 40, n. 2, 2019.

DUTRA, D.; ORFANÓ, B.; ALMEIDA, V. Result linking adverbials in learner corpora. **Domínios de Lingu@ gem**, v. 13, n. 1, p. 400-431, 2019.

DUTRA, D.; ORFANÓ, B.; GUEDES, A.; ALVES, J.; FEKETE, J. The learner corpus path: a worthwhile methodological challenge. **SciELO Preprints.** 2022. Disponível em: <https://doi.org/10.1590/1678-460x202149731>. Acesso em: 5 fev. 2022.

DUTRA, D.; QUEIROZ, J.; MACEDO, L.; COSTA, D.; MATTOS, E. Adjectives as nominal pre-modifiers in chemistry and applied linguistics research articles. In: RÖMER, U.; CORTES, V.; FRIGINAL, E. (org.). **Advances in Corpus-based Research on Academic Writing:** Effects of discipline, register, and writer expertise. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2020.

DUTRA, D.; QUEIROZ, J.; ALVES, J. Adding information in argumentative texts: a learner corpus-based study of additive linking adverbials. **Revista de Estudos Anglo-Americanos**, v. 46, n. 1, p. 9-32, 2017.

GÖTZ, S.; MUKHERJEE, J. Investigating the effect of the study abroad variable on learner output: A pseudo-longitudinal study on spoken German learner English. In: BREZINA, V.; FLOWERDEW, L. (eds.). **Learner Corpus Research**: New Perspectives and Applications. New York: Bloomsbury Academic, 2017, p. 47.

GRANGER, S. The computer learner corpus: a versatile new source of data for SLA research. In: GRANGER, S. (ed.). **Learner English on Computer**. London & New York: Addison Wesley Longman, 1998, p. 3-18.

GRANGER, S. The contribution of learner corpora to second language acquisition and foreign language teaching. **Corpora and language teaching**, v. 33, p. 13-32, 2009.

GRANGER, S. DAGNEAUX, E.; MEUNIER, F.; PAQUOT, M. (Ed.). **International corpus of learner English**. Louvain-la-Neuve: Presses universitaires de Louvain, 2009.

GRAY, B.; GELUSO, J.; NGUYEN, P. The longitudinal development of grammatical complexity at the phrasal and clausal levels in spoken and written responses to the TOEFL iBT® test. **ETS Research Report Series**, v. 2019, n. 1, p. 1-51, 2019.

GRANGER, S.; DUPONT, M.; MEUNIER, F.; NAETS, H.; PAQUOT, M. **International Corpus of Learner English**, Version 3. Louvain: Presses Universitaires de Louvain, 2020.

GUEDES, A. **Verbos do inglês acadêmico escrito e suas colocações**: um estudo baseado em um corpus de aprendizes brasileiros de inglês. Orientador: Deise Prina Dutra. 2017. 202 f. Tese (Doutorado) - Programa de Pós-graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2017.

HO-PENG, L. Using T-unit measures to assess writing proficiency of university ESL students. **RELC Journal**, v. 14, n. 2, p. 35-43, 1983.

KNOCH, U.; ROUHSHAD, A.; OON, S.; STORCH, N. What happens to ESL students' writing after three years of study at an English medium university?. **Journal of Second Language Writing**, v. 28, p. 39-52, 2015.

KREYER, R.; SCHAUB, S. The development of phrasal complexity in German intermediate learners of English. **International Journal of Learner Corpus Research**, v. 4, n. 1, p. 82-111, 2018.

KYLE, K.; CROSSLEY, S. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. **The Modern Language Journal**, v. 102, n. 2, p. 333-349, 2018.

LAN, G.; LUCAS, K.; SUN, Y. Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. **System**, v. 85, p. 102-116, 2019.

LIMA JR, R. A longitudinal study on the acquisition of six English vowels by Brazilian learners. **Proceedings of the 19th International Congress of Phonetic Sciences**. 2019. p. 3180-3184.

LÓPEZ-FERRERO, C.; BACH, C. Discourse analysis of statements of purpose: Connecting academic and professional genres. **Discourse studies**, v. 18, n. 3, p. 286-310, 2016.

MADEN-WEINBERGER, U. "Hätte, wäre, wenn…": A pseudo-longitudinal study of subjunctives in the Corpus of Learner German (CLEG). **International Journal of Learner Corpus Research**, v. 1, n. 1, p. 25-57, 2015.

MATTOS, E. **A Corpus-Based Study of Hyphenated Premodifiers in Complex NPs in Biology Research Articles**. Orientador: Deise Prina Dutra. 2020. 161 f. Dissertação (Mestrado) - Programa de Pós-graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2020.

NUNES, L.; ORFANÒ, B. Investigating the system of TRANSITIVITY in passive that-clauses of research abstracts. In: KENNY, N.; ESCOBAR, L. (eds.). **The changing face of ESP in today's classroom and workplace**. Wilmington: Vernom Press, 2020, p. 163.

PARKINSON, J.; MUSGRAVE, J. 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.

QUEIROZ, J. **The grammatical complexity of English noun phrases in Brazilian learners' academic writing:** a corpus-based study. Orientador: Deise Prina Dutra. 2019. 137 f. Dissertação (Mestrado) - Programa de Pós-graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2019.

RÖMER, U.; CORTES, V.; FRIGINAL, E. (org.). **Advances in Corpus-based Research on Academic Writing:** Effects of discipline, register, and writer expertise. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2020.

SAMRAJ, B.; MONK, L. The statement of purpose in graduate program applications: Genre structure and disciplinary variation. **English for Specific Purposes**, v. 27, n. 2, p. 193-211, 2008.

SANTOS, M. **Descrição do uso das conjunções but e however em redações acadêmicas em língua inglesa de nível B1 com base em corpus**. [Unpublished master thesis]. Universidade Estadual Paulista Júlio de Mesquita Filho, 2008.

STAPLES, S.; EGBERT, J.; BIBER, D.; GRAY, B. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. **Written Communication**, v. 33, n. 2, p. 149-183, 2016.

SWALES, J.; FEAK, C. **Academic writing for graduate students**. Ann Arbor: University of Michigan Press, 1994.

WINTER, B. **Statistics for linguists**: An introduction using R. New York: Routledge, 2020.

WOLFE-QUINTERO, K.; INAGAKI, S.; KIM, H. **Second language development in writing**: Measures of fluency, accuracy, & complexity. Hawaii: University of Hawaii Press, 1998.

APPENDIX A

Tags and operational definitions of the Complexity Tagger

Stage	Tag	Feature and Operational Definition
1		**Finite complement clauses (that, wh-) controlled by common verbs (e.g., think, know, say)**
	vcmpth-1a	That-complement clauses Instances of that tagged as verb complement when preceded by very common verbs occurring in this structure (>100 times per million words; Biber et al., 1999, pp. 661–666). Occurring in the following patterns: V + that-clause: *believe, feel, find, guess, know, see, think, say, show, suggest* V + NP + that-clause, allowing up to three intervening words for NP: *show* V + to NP + clause, allowing up to two intervening words for NP*: say, suggest* Instances of these very common verbs tagged as containing a zero complementizer
	vcmpwh-1a	Wh-complement clauses Instances of wh-words (where, when, who, whom, which, why, whose, whatever, whoever, what, whichever, how) following very common verbs occurring in this structure (>50 times per million words; Biber et al., 1999, pp. 685–686). Occurring in the following patterns: V + wh-clause: *tell, know, wonder, see* V + NP + wh-clause, allowing up to three intervening words for NP: *tell* Instances of if preceded by very common verbs (BIBER et al., 1999, pp. 691–693): *know, see, wonder* Instances of whether preceded by very common verbs (p. 692): *know* Note: Patterns in which the wh-complementizer is followed by a to-clause (e.g., I don't know where to put this)are tagged as verb to-complement clauses (see Biber et al., 1999, p. 685)
2a		**Finite complement clauses (that, wh-) controlled by a wider set of verbs**
	vcmpth-2a	That-complement clauses Instances of that tagged as verb complement when not preceded by one of the very common verbs listed in 1A (BIBER et al., 1999, pp. 685–686). Occurring in the following patterns:

		V + that-clause V + NP + that clause V + to NP + clause Instances of verbs tagged as containing a zero complementizer (excluding the verbs from 1A)
	vcmpwh-2a	Wh-complement clauses Instances of wh-words (*where, when, who, whom, which, why, whose, whatever, whoever, what, whichever, how*) following other common verbs occurring in this structure (>20 times per million words and "other attested verbs"; Biber et al., 1999, pp. 685–686) Occurring in the following patterns (lists of verbs for each pattern available upon request): V + wh-clause V + NP + wh-clause (allowing up to three intervening words for NP) V + prep + wh-clause Instances of if preceded by other verbs (p. 693) Instances of whether preceded by other verbs (p. 692)
2b		**Finite adverbial clauses**
	fadvl-2b	Single- and multiword subordinators Lexical, tag, and/or contextual matches for common circumstance adverbial subordinators (BIBER et al., 1999, pp. 841–844) Lexical match (all occurrences): *although, because* (not followed by of), *unless, whenever, whereas, wherever* Lexical match when tagged as a subordinator: *for, once* Lexical match when (a) tagged as a subordinator and (b) not followed by an ing-form: *after, before, like, since(even/as) though, until, while, whilst* Lexical match when (a)tagged as a subordinator and(b) not followed by an ed-clause: *as* Lexical match for words when they (a) do not meet criteria for wh-complement clauses (see 1A, 2A), (b) is not followed by an ing-form, and (c) is not preceded by an adjective or a preposition: *when, where, whatever* Lexical match for common multiword subordinators when that is tagged as a subordinator (not demonstrative): except that, now that, so that, such that Instances of if when not preceded by a common verb controlling if-clauses (see 1A, 2A) or common adjectives controlling if-clauses (*sure, unsure, clear, unclear, certain, uncertain*); instances of *as if*
2c		**Non-finite (to-, ing-) complement clauses controlled by common verbs**

	vcmpto-2c	To-clauses Instances of *to* tagged as an infinitive marker when preceded by very common verbs occurring in this structure (>100 times per million words; Biber et al., 1999, pp. 699–705). Occurring in the following patterns: V + to-clause: *attempt, begin, like, seem, tend, try, want* Note: Excludes bare infinitive clauses
	vcmping-2c	Ing-clauses Ing-verb forms tagged as nonfinite when preceded by very common verbs occurring in this structure (>40 times per million words; Biber et al., 1999, pp. 740–741). Occurring in the following patterns: V + ing-clause: begin, go (around/on), keep (on), start, stop V + NP + ing-clause, allowing up to three intervening words for NP: see
2d		**Phrasal embedding in the clause: Adverbs as adverbials**
	adv-2d	Circumstance adverbials Most common single-word circumstance adverbs (Biber et al., 1999, pp. 795–798); single words typically functioning as adverbials, when tagged as an adverb: *again, already, also, always, ever, Friday, here, Monday, never, now, often, Saturday, sometimes, still, Sunday, then, there, Thursday, today, Tuesday, usually, Wednesday, yesterday*. Single words that can be adverbials or modifiers when followed by a verb (all other instances excluded): *even, just, only*. Note: Single words that can be adverbials or modifiers that are excluded: too.
	adv-2d	Stance adverbials Most common single-word stance adverbs (Biber et al., 1999, pp. 869–879); single words typically functioning as adverbials, when tagged as an adverb: *actually, certainly, definitely, generally, maybe, perhaps, probably.* Single words that can be adverbials or modifiers/other parts of speech, when followed by a verb (all other instances excluded): *really, totally*. Note: Single words that can be adverbials or modifiers that are excluded: like.
	adv-2d	Linking adverbials Most common single-word linking adverbials (Biber et al., 1999, p. 887); single words typically functioning as adverbials, when tagged as an adverb: *anyway, finally,*

		first(ly), furthermore, hence, however, nevertheless, second(ly), then, therefore, third(ly), though, thus, yet Note: Single words that can be adverbials or modifiers/other parts of speech are excluded: rather, so.
2e		**Simple phrasal embedding in the noun phrase: Attributive adjectives**
	jatrb-2e 3A	Attributive adjective as nominal premodifier Adjectives tagged "attributive" by the Biber Tagger only when: 1. the following word is not tagged as an adverbial noun (e.g., … feel very tired tomorrow) 2. there are no additional adjectives, nouns, or genitive nouns in the phrase
3b		**Finite complement clauses controlled by adjectives**
	jcmpth-3b	That-complement clauses controlled by adjectives, simple (i.e., non-extraposed) Instances of *that* tagged as an adjective complement when no extraposed patterns are found
3c		**Non-finite complement clauses controlled by a wider set of verbs**
	vcmpto-3c	To-clauses Instances of *to* tagged as an infinitive marker when preceded by other verbs occurring in this structure (>20–50 times per million words, other attested verbs; Biber et al., 1999, pp. 700–705). Occurring in the following patterns (list of verbs included for each pattern available upon request): V + to-clause V + NP + to-clause, with up to two intervening words for NP Instances of *to* tagged as an infinitive marker when preceded by a wh-complementizer (which, who, whom, whose, where, when, why, what, how, whether) Note: Excludes bare infinitive clauses
	vcmping-3c	Ing-clauses Ing-verb forms tagged as nonfinite when preceded by verbs occurring in this structure (excluding very common verbs in 2C; Biber et al., 1999, pp. 740–741). Occurring in the following patterns (list of verbs included for each pattern available upon request): V + ing-clause be + Ved + prep + ing-clause V + prep + ing-clause
3d		**Finite relative clauses**

	finrel-3d	Finite relative clauses with that All instances of that tagged as “rel” Note: Does not count instances of zero relativizer
	finrel-3d	Finite relative clauses with wh-relative pronouns and adverbs All instances of relative pronouns or determiners (who, whom, whose, which) tagged as “rel.” Instances of relative adverbs (where, when, why) when preceded by common nouns heading relative clauses with adverbial gaps (Biber et al., 1999, p. 627–628): nouns preceding relative adverb where: *area, bit, case, condition, country, example, hospital, house, place, point, room, situation, spot* nouns preceding relative adverb when: *bit, case, day, moment, occasion, period, season, time* nouns preceding relative adverb why: *reason* Note: Does not count instances of zero relativizer
3e		**Simple phrasal embedding in the noun phrase: Nouns as noun premodifiers**
	npnm-3e	Nouns as noun premodifier Word tagged as a noun followed by another word tagged as a noun, excluding 1. When N2 is tagged as an adverbial noun, e.g., *our dreams today* 2. When there are no additional adjectives, nouns, or genitive nouns in the phrase.
3f		**Possessive nouns as noun premodifiers**
	npnmgen-3f	Word tagged as a noun followed by a word tagged as a possessive marker (^$) followed by another noun (i.e., noun + possessive + noun). Excludes instances when there are additional adjectives, nouns, or genitive nouns in the phrase.
4a		**Non-Finite complement clauses controlled by adjectives (simple)**
	jcmpto-4a	To-complement clauses controlled by adjectives, simple (i.e., non-extraposed) Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that typically occur only with non-extraposed, post predicative *to*-clauses (BIBER et al., 1999, pp. 718–721). Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that can occur with either post-predicate or extraposed complements, when

		not preceded by it and a form of be/seem/become (BIBER et al., 1999, pp. 718–720; see patterns listed in 4B). Note: Does not include patterns adj + for NP+ to-clause (e.g., too difficult for them to remember)
4b		**Extraposed complement clauses**
	jcmpxtra-4b	Extraposed that-complement clauses controlled by adjectives Instances of *that* tagged as an adjective complement and preceded by the set of adjectives that typically occur only with extraposed complements (BIBER et al., 1999, p. 671–674). Instances of that tagged as an adjective complement and preceded by the set of adjectives that can occur with either post-predicate or extraposed complements when preceded by it and a form of be/seem/become.
	jcmpxtra-4b	Extraposed to-complement clauses controlled by adjectives Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that typically occur only with extraposed *to*-complements (BIBER et al., 1999, pp. 618–621). List of adjectives included available upon request. Instances of *to* tagged as an infinitive marker and preceded by the set of adjectives that can occur with either post-predicate or extraposed complements, when not preceded by it and a form of be/seem/become (BIBER et al., 1999, pp. 618–621) Note: Does not capture pattern adj + for NP+ to-clause (e.g., it is difficult for them to choose)
4c		**Nonfinite relative clauses**
	nfrel-4c	Ing- and –ed clauses as post-nominal modifiers present participle forms tagged as post-nominal modifier (tag "vwbg") past participle forms tagged as a post-nominal modifier (tag "vwbn"). Note: to-clauses as post-nominals are not included in this analysis.
4d		**More phrasal embedding in the NP: Attributive adjectives and nouns as premodifiers (multiple modifiers)** Note: The tagger works by processing all features before this feature; after all features are tagged, the tagger processes the text again to look for instances of multiple modifiers. Tags are appended with "4d" when multiple modifiers are found so that single modifiers can be counted under Features 2e, 3d, and 3f. To obtain a count for this feature (i.e., a noun phrase with multiple premodifiers), a tag is then added to the head noun so that

		there is one tag per noun phrase.
	jatrb-2e4d	Attributive adjectives occurring with other premodifiers Adjectives tagged as jatrb-2e when they are preceded or followed by another word tagged as a noun pre-modifier (i.e., jatrb-2e, npnm-3e, npnmgen-3f)
	npnm-3e4d	Nouns as noun premodifier with other premodifiers Nouns tagged as npnm-3e when they are preceded or followed by another word tagged as a noun pre-modifier (i.e., jatrb-2e, npnm-3e, npnmgen-3f)
	npnmgen-3f4d	Possessive nouns as noun premodifiers with other premodifiers Possessive nouns tagged as npnmgen-3f when preceded or followed by another word tagged as a noun premodifier (i.e., jatrb-2e, npnm-3e, npnmgen-3f)
	hn-4d	Head noun modified by multiple premodifiers Head noun of phrase (i.e., a noun not tagged as a noun premodifier npnm-3e) when preceded by a word tagged as multiple modifiers (i.e., jatrb-2e4d, npnm-3d4d, npnmgen-3f4d)
5a		**Prepositions with non-finite complement clauses**
	ppning-5a	Postnominal preposition + ing-complement clause constructions Instances of the pattern noun + preposition + ing-clause, where the function of the PP is post-nominal.
	ppxing-5a	Preposition + ing-complement clauses when not following a noun. Instances of the pattern preposition + ing-clause when not functioning as a noun postmodifier.
5b		**Complement clauses controlled by nouns**
	ncmpth-5b	That-complement clauses controlled by nouns Instances of that tagged as a noun complement.
	ncmpto-5b	To-complement clauses controlled by nouns Instances of *to* tagged as an infinitive marker and preceded by a common (>10 times per million words) and less common noun controlling to-complement clauses (Biber et al., 1999, p. 652).
	ncmping-5b	Noun + of + ing-complement clauses Instances of an ing- form tagged as non-finite and preceded by common (>5 times per million words) nouns in the following pattern (Biber et al., 1999, pp. 653–654). Pattern: noun + of + present participle

Source: Adapted from Gray et al. (2019, p. 39-44)

APPENDIX B

Operational Definitions and Coding Notes for the Manual Coding of Prepositional Phrases

Stage	Tag	Feature and Operational Definition
3a	ppadvl-3a	**Prepositional phrases as adverbials**
		Specific combinations of words in which the PP is always adverbial in nature (based on common words/phrases observed in the TOEFL iBT Longitudinal corpus during testing and program development): *for example, prepare for, in reference to, because of* Manual coding of prepositions when answering questions of where, when, how, or why an action occurred. These include the following typical contexts (not exhaustive): 1. PPs directly following a verb (but distinguished from multiword verbs) 2. PPs directly following a be verb 3. PPs indicating a causative meaning (e.g., *because of X*) 4. by-phrases indicating the agent in passives (e.g., *was taken by the student*) 5. PPs that function as linking adverbials (e.g., *on the other hand, in addition*)
3g	ppnof-3g	**Of-phrases as noun postmodifiers**
		Noun followed by preposition *of* (including instances of three-word complex prepositions in which the second preposition is *of*). Excludes instances of the following: multiword determiners (e.g., a lot of) N + *of* + *ing*-clause patterns (these are included under ppning-5a) instances of two-word complex prepositions with *of*
		Simple prepositional phrases as postmodifiers, especially with prepositions other than of when they have concrete/locative meaning

3h	ppnc-3h	Prepositional phrase occurring after a noun, in which the prepositional phrase identifies the referent of the head noun or adds descriptive information about the head noun. PPs as postnominals can often be rephrased with a relative clause. Restricted to instances in which the PP carries a concrete or locative meaning, including textual location. Examples: *something in the dust* (cf. something which is located in the dust) *everyone around you* (cf. everyone who is around you*) the correct way to their home* (cf. the correct way that leads to their home) *the theory in the passage* (cf. the theory that appears in the passage)
4e	ppna-4e	**Simple prepositional phrases as noun postmodifiers, especially with prepositions other than of when they have abstract meanings**
		Specific phrase *such as*, which is typically postnominal Three-word complex prepositions (BIBER et al., 1999, p. 75) are analyzed compositionally. The second preposition in a three-word complex preposition is automatically tagged as a noun postmodifier when the second word is a noun (e.g., *in exchange for, in return for*) The first preposition in three-word complex prepositions is tagged pp? Combinations with of as the second preposition (e.g., *in light of, by way of*) are already captured by ppnof-3g. Prepositional phrase occurring after a noun, in which the prepositional phrase identifies the referent of the head noun or adds descriptive information about the head noun. PPs as postnominals can often be rephrased with a relative clause. Restricted to instances in which the PP carries an abstract meaning. Many types of abstract meaning are possible; a few example meanings observed in the longitudinal iBT corpus include the following: Topic: three theories about bird's navigational abilities,

		information about the job Time: the air in the morning, a period in the past Other: all the responses for them, a conclusion from an experiment, landmarks like rivers and coastlines, heated discussion among the students, a dilemma between a useful job and an interested job Prepositions (especially in, on, and to) are coded as abstract if there is not a literal, locative meaning. Thus, the following examples would be coded as abstract: *your major in the university, the pull on the crystals, influence on their future job, their way to success*. Instances of for are tagged as abstract postnominals in the construction N+ for NP + to-clause: *three ways for birds to navigate*
5a	ppning-5a	**Prepositions with nonfinite complement clauses**
		Preposition occurring in the following pattern: Word tagged as noun + preposition + word tagged as nonfinite ing-form. Instances of the pattern noun + preposition + ing-clause, where the function of the PP is postnominal: *the best way for ensuring that research about choosing subjects effort in studying this subject.*
	ppxing-5a	**Preposition + ing-complement clauses when not following a noun**
		Preposition occurring in the following pattern: Word not tagged as noun + preposition + word tagged as nonfinite ing-form. Instances of the pattern preposition + ing-clause when not functioning as a noun postmodifier: *feels happy through teaching* *afraid of going into the class* *came into the human body by breathing* *the professor rebutted it by mentioning that*

Source: Adapted from Gray et al. (2019, p. 36-38)

APPENDIX C

R Scripts

True longitudinal analysis

One-way Anova

```
#Install packages
install.packages("effectsize")
install.packages("dplyr")

#Run packages
library("effectsize")
library("dplyr")

#Define folder
setwd("C:/Users/jessi/OneDrive/Área de Trabalho/Resultados estatísticos – Dissertação/True longitudinal tables")

#Open data
File1 = read.table('filename.csv', header = T, sep=";")
print(File1)
#Summary statistics by groups
group_by(File1, Time)%>%
  summarise(
    count = n(),
    mean = mean(Feature1, na.rm = TRUE),
    sd = sd(Feature1, na.rm = TRUE),
    median = median(Feature1, na.rm = TRUE),
    IQR = IQR(Feature1, na.rm = TRUE)
  )

# Compute the analysis of variance
Anova1 <– aov(Feature1 ~ Time, data = File1)
```

```
# Summary of the analysis
summary(Anova1)

#run cohen's f
cohens_f(Anova1)
```

Boxplot creation

```
#Install package
install.packages("ggpubr")

# Run plot
library("ggpubr")

#Define folder
setwd("C:/Users/jessi/OneDrive/Área de Trabalho/Resultados estatísticos – Dissertação/True longitudinal tables")

#Open data
File2 = read.table('filename2.csv', header = T, sep=";")
print(File2)
ggboxplot(File2, x = "ï..Featuretype", y = "Time1",
        color = "ï..Featuretype", palette = c("#339900", "#3399FF"),
        ylab = "Time1", xlab = "Features")
```

Quasi-longitudinal analysis

```
#Install packages
install.packages("lme4") #Mixed effects linear regression – lmer function
install.packages("car") #Type II Wald chisquare tests – Anova function
```

```
install.packages("jtools") #Summ function

install.packages("nlme")

install.packages("dplyr")


#Run packages

library("lme4", lib.loc="~/R/win-library/4.1")

library("car", lib.loc="~/R/win-library/4.1")

library("jtools", lib.loc="~/R/win-library/4.1")

library("nlme")

library("dplyr")


#Define folder

setwd("C:/Users/jessi/OneDrive/Área de Trabalho/Resultados estatísticos – Dissertação/Quasi longitudinal tables")

#Open data

File3 = read.table('filename3.csv', header = T, sep=";")

print(File3)


#Run regressions

reg1 <– lmer(File3$fadvl2b ~ factor(File3$Semester) + factor(File3$Register) +

                    factor(File3$Division) + (1 | File3$ï..Student))

summary(reg1)

summ(reg1)

anova1 <– Anova(reg1,type=2)

anova1
```

```
reg2 <– lmer(File3$frel ~ factor(File3$Semester) + factor(File3$Register) +
                    factor(File3$Division) + (1 | File3$ï..Student))
summary(reg2)
summ(reg2)
anova2 <– Anova(reg2,type=2)
anova2

reg3 <– lmer(File3$jatrbTotal ~ factor(File3$Semester) + factor(File3$Register) +
                    factor(File3$Division) + (1 | File3$ï..Student))
summary(reg3)
summ(reg3)
anova3 <– Anova(reg3,type=2)
anova3

reg4 <– lmer(File3$NPs ~ factor(File3$Semester) + factor(File3$Register) +
                    factor(File3$Division) + (1 | File3$ï..Student))
summary(reg4)
summ(reg4)
anova4 <– Anova(reg4,type=2)
anova4
```

Plot creation in R

```
#Install package
install.packages("ggpubr")

# Run plot
library("ggpubr")

#Define folder

setwd("C:/Users/jessi/OneDrive/Área de Trabalho/Resultados estatísticos – Dissertação/Quasi longitudinal tables")


#Open data

File3 = read.table('filename3.csv', header = T, sep=";")


print(File3)


#plot creation

p1 <- ggplot(File3, aes(Division, jatrbTotal)) +

  geom_point() +

  theme_bw() +

  labs(x = "Academic Division") +

  labs(y = "Attributive Adjective") +

  geom_smooth()


p2 <- ggplot(File3, aes(Division, NPs)) +

  geom_point() +

  theme_bw() +
```

```
  labs(x = "Academic Division") +
  labs(y = "NPs") +
  geom_smooth(method = "lm")

# display plots
ggpubr::ggarrange(p1, p2, ncol = 2, nrow = 1)
```

APPENDIX D

The most frequent attributive adjectives across registers (more than five occurrences)

- Abstract: **comic** (books, magazines, histories), and **public** (sector, health, policies).
- SOP: **academic** (activities, doubts, goals), **best** (solutions, universities), **big** (country, interest), **electrical** (engineering, projects, vehicles), **good** (labs, point, reputation), **professional** (area, goal, life).
- Summary: **American** (economy, nation, people).
- Essay B2: **aerospace** (area, company, technology), **better** (chances, jobs, solution), **big** (amount, hydropower, investments), **different** (field, products, reality), **electric** (cars, energy, vehicles), **electrical** (circuit, systems, variables), **federal** (university), **genetic** (causes, principle, variation), **good** (grades, performance, results), **great** (autonomy, effort, importance), **hard** (work), **important** (factor, point, roles), **introverted** (people, person, personality), **mental** (school, diseases, health), **mutual** (fund), **natural** (cycle, disaster, resources), **new** (abilities, age, criteria), **professional** (formation, management, record), and **social** (communication, groups, inequality).

APPENDIX E

CorIFA subcorpus text codes

TIME 1

corifa-ufmg-b1.int.ne.aess.2015-2.0554.0331

corifa-ufmg-b1.int.ne.sum.2016-1.1092.0516

corifa-ufmg-b1.ind.ne.sop.2016-1.1048.0521

corifa-ufmg-b1.int.ne.sum.2016-1.1115.0525

corifa-ufmg-b1.ind.ne.sop.2016-1.1052.0537

corifa-ufmg-b1.ind.ne.sop.2016-2.1231.0563

corifa-ufmg-b1.ind.ne.sop.2016-2.1226.0579

corifa-ufmg-b1.ind.ne.sop.2016-2.1239.0713

corifa-ufmg-b1.ind.ne.sop.2016-2.1252.0726

corifa-ufmg-b1.ind.ne.sop.2017-1.1340.0774

corifa-ufmg-b1.ind.ne.sop.2017-2.1538.0972

corifa-ufmg-b1.ind.ne.sop.2017-2.1539.0973

corifa-ufmg-b1.ind.ne.sop.2017-2.1540.0974

TIME 2

corifa-ufmg-b1.int.ne.abs.2016-1.0795.0331

corifa-ufmg-b1.ind.ne.abs.2016-2.0888.0516

corifa-ufmg-b1.ind.ne.abs.2016-2.0893.0521

corifa-ufmg-b1.ind.ne.abs.2016-2.0897.0525

corifa-ufmg-b1.ind.ne.abs.2016-2.0909.0537

corifa-ufmg-b1.int.ne.abs.2017-1.1398.0563

corifa-ufmg-b1.int.ne.abs.2017-1.1403.0579

corifa-ufmg-b1.int.ne.abs.2017-1.1413.0713

corifa-ufmg-b1.int.ne.abs.2017-1.1385.0726

corifa-ufmg-b1.int.ne.abs.2017-2.1573.0774

corifa-ufmg-b1.ind.ne.abs.2018-1.1702.0972

corifa-ufmg-b1.ind.ne.abs.2018-1.1705.0973

corifa-ufmg-b1.ind.ne.abs.2018-1.1693.0974

TIME 3

corifa-ufmg-b2.ind.ne.aess.2016-2.0714.0331

corifa-ufmg-b2.ind.ne.aess.2017-1.0933.0516

corifa-ufmg-b2.ind.ne.aess.2017-2.0997.0579

corifa-ufmg-b2.ind.ne.aess.2018-2.2106.0713

corifa-ufmg-b2.ind.ne.aess.2017-1.0938.0521

corifa-ufmg-b2.ind.ne.aess.2017-2.0984.0726

corifa-ufmg-b2.ind.ne.aess.2017-2.1001.0525

corifa-ufmg-b2.ind.ne.aess.2018-1.1845.0774

corifa-ufmg-b2.ind.ne.aess.2017-1.0956.0537

corifa-ufmg-b2.ind.ne.aess.2018-2.2109.0972

corifa-ufmg-b2.ind.ne.aess.2017-2.0981.0563

corifa-ufmg-b2.ind.ne.aess.2018-2.2110.0973

corifa-ufmg-b2.ind.ne.aess.2018-2.2104.0974