UNIVERSIDADE FEDRAL DE MINAS GERAIS – UFMG

Faculdade de Letras – FALE

Programa de Pós-Graduação em Estudos Linguísticos - POSLIN

Edilson Rosa da Rocha

ESTUDO DIRECIONADO POR CORPORA:
Estruturas Lexicais em um corpus especializado.

Belo Horizonte

2024

Edilson Rosa da Rocha

ESTUDO DIRECIONADO POR CORPORA:
Estruturas Lexicais em um corpus especializado.

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

Área de Concentração: Linguística Aplicada

Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras.

Orientadora: Profa. Dra. Deise Prina Dutra

Belo Horizonte
2024

R672e

Rocha, Edilson Rosa da.
Estudo direcionado por corpora [manuscrito] : estruturas lexicais em um corpus especializado / Edilson Rosa da Rocha. – 2024.
1 recurso online (110 f.: il., tabs., color., p&b.) : pdf.

Orientadora: Deise Prina Dutra.

Área de concentração: Linguística Aplicada.

Linha de pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras.

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

Bibliografia: f. 95-99.

Apêndices: f. 100-110.

Exigências do sistema: Adobe Acrobat Reader.

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

CDD: 420.7

Ficha catalográfica elaborada pela Bibliotecária Priscila Oliveira da Mata - CRB/6- 2706
Biblioteca Professor Rubens Costa Romanelli - FALE/UFMG

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

FOLHA DE APROVAÇÃO

UM ESTUDO DIRECIONADO POR CORPORA: Estruturas Lexicais em um corpus especializado

EDILSON ROSA DA ROCHA

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 APLICADA, linha de pesquisa Ensino/Aprendizagem de Línguas Estrangeiras.

Aprovada em 02 de agosto de 2024, pela banca constituída pelos membros:

Prof(a). Deise Prina Dutra - Orientadora

UFMG

Prof(a). Ana Eliza Pereira Bocorny

UFRGS - Instituto de Letras

Prof(a). Heliana Ribeiro de Mello

UFMG

Belo Horizonte, 02 de agosto de 2024.

Documento assinado eletronicamente por **Heliana Ribeiro de Mello**, **Professora do Magistério Superior**, em 05/08/2024, às 14:19, 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 05/08/2024, às 18:45, 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 07/08/2024, às 13:33, 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 **3436183** e o código CRC **4986AA69**.

Referência: Processo nº 23072.237231/2024-12 SEI nº 3436183

AGRADECIMENTOS

A Deus, meu bem maior.

À minha mãe (in memoriam), por sempre dizer que minha vida seria diferente.

À amiga mais chegada que uma irmã, Dênia Caetano. Muito obrigado por sustentar-me com suas orações.

À minha orientadora, Profª Drª Deise Prina Dutra, que me orientou de maneira competente e amigável. Obrigado por apresentar-me à Linguística de Corpus e Computacional, e a Estatística aplicada aos dados linguísticos. Minhas novas paixões!

À Profª Drª Heliana Ribeiro de Mello, que desde a graduação, nunca se recusou a receber-me em seu gabinete e orientar-me em minhas dúvidas acadêmicas.

Aos colegas do grupo GECEA (Grupo de Corpora Especializados e de Aprendizagem), pelas reflexões frutíferas e valiosas durante as reuniões.

À banca que avalia este trabalho.

A todos, minha sincera gratidão!

RESUMO

Estudos linguísticos relacionados à fraseologia têm ganhado credibilidade, principalmente quanto à formação e análise da linguagem formulaica (Hunston e Francis, 1999; Wray, 2002; Biber, 2009). Nesta pesquisa, buscamos identificar, analisar e classificar diretamente as Estruturas Lexicais (ELexs), sob a perspectiva metodológica direcionada por corpus (Biber, 2012). Consideramos a hipótese de identificação das ELexs independentemente dos Pacotes Lexicais (PLs), analisando as estruturas internas das unidades lexicais (ULs). Utilizamos o corpus especializado *Corpus of Articles of Applied Linguistics* (CorAAL), compilado de 6 revistas de alto impacto da área de Linguística Aplicada em língua inglesa, totalizando 973.844 palavras de 150 artigos, publicados entre 2014 e 2018. O AntConc (Anthony, 2022), através do *N-gram,* gerou a lista de ULs. Investigamos sequências lexicais de 5-palavras com uma lacuna variável, com frequência mínima de 20 vezes por milhão de palavras, com dispersão mínima de 10 vezes, resultando em uma lista final de 66 ULs. Identificamos 11 ULs que não estão associadas aos PLs do estudo de Biber *et al*. (1999), mas sua ausência nesse estudo não as classifica automaticamente como ELexs. Para tal análise foram integrados ao critério de frequência os parâmetros de variabilidade e previsibilidade (Tan e Römer 2022). Utilizamos agrupamento hierárquico aglomerativo e *scripts* em R para comparar frequência, variabilidade e entropias internas das ULs, constatando baixa variabilidade (0.02 - 0.05) e previsibilidade (0.0 - 0.0). Por exemplo, as unidades lexicais ((*at, in) the + of the* [*end, beginning, time*]), (*english as a + language* [*foreign, second*]) e (*it is + to note* [*important*]) exibem características de ELexs por apresentarem descontinuidade em sua unidade lexical e flexibilidade quanto ao preenchimento das lacunas com palavras funcionais e de conteúdo. Assim, ao identificar ELexs apenas a partir de ULs contínuas, excluímos as de baixa variabilidade, conforme destacado na análise. Além do mais, o preenchimento dos espaços internos das 11 ULs identificadas (1*345, 12*45, 123*5) são com palavras de conteúdo de base nominal (bN), base verbal (bV) e base adjetival (bA), como exemplificado pela expressão: *the + of the* [*purpose*(*s*), *validity, teaching, use, majority*, etc.]. Esses *clusters* demonstram níveis elevados de variabilidade (de .11 a .74) e previsibilidade (de .58 a .97) internamente, sendo divididos em subgrupos de acordo com a redução da similaridade dos *clusters* que estão sendo fundidos. O segundo agrupamento apresenta subdivisões distintas no dendrograma. Os resultados mostram que conforme a variabilidade interna aumenta, as ELexs, preenchidas com palavras de conteúdo e diferentes entre si,

tendem a formar grupos distintos. Assim, a análise estatística usando variabilidade e entropia interna permitiu identificar ELexs não derivadas de PLs.

Palavras chave: Direcionado por corpus; Estruturas Lexicais; Pacotes Lexicais; Análise Multivariada de Dados; Clusters

ABSTRACT

Linguistic studies related to phraseology have been gaining credibility, especially regarding the formation and analysis of formulaic language (Hunston & Francis, 1999; Wray, 2002; Biber, 2009). In this research, we aimed to directly identify, analyze, and classify Phrasal frames (P-frames) from a corpus-driven perspective (Biber, 2012). We considered the hypothesis of identifying p-frames independently of Lexical Bundles (LBs), analyzing the internal structures of lexical units (ULs). We utilized the specialized Corpus of Articles of Applied Linguistics (CorAAL), compiled from 6 high-impact journals in the field of Applied Linguistics in English language, totaling 973,844 words from 150 articles published between 2014 and 2018. AntConc (Anthony, 2022), through N-gram tool, generated the list of ULs. We investigated lexical sequences of 5-words with a variable gap, with a minimum frequency of 20 times per million words and a minimum dispersion of 10 times, resulting in a final list of 66 ULs. We identified 11 ULs that are not associated with the LBs in the study by Biber et al. (1999), but their absence in this study does not automatically classify them as p-frames. For such analysis, the parameters of variability and predictability (Tan & Römer, 2022) were integrated into the frequency criterion. We employed agglomerative hierarchical clustering and R scripts to compare the frequency, variability, and internal entropies of the ULs, observing low variability (0.02 - 0.05) and predictability (0.0 - 0.0). For instance, the lexical units ((*at, in*) *the + of the* [*end, beginning, time*]), (*english as a + language* [*foreign, second*]), and (*it is + to note* [*important*]) exhibit characteristics of p-frames by displaying discontinuity in their lexical units and flexibility regarding the filling of gaps with functional and content words. Thus, by identifying p-frames only from continuous ULs, we exclude those with low variability, as highlighted in the analysis. Furthermore, the filling of internal spaces in the 11 identified ULs (1*345, 12*45, 123*5), consists of content words with nominal base (Nb), verbal base (Vb), and adjectival base (Ab), as exemplified by the expression: *the + of the* [*purpose*(*s*), *validity*, *teaching*, *use*, *majority*, etc.]. These clusters demonstrate high levels of internal variability (from .11 to .74) and predictability (from .58 to .97), being divided into subgroups according to the reduction of similarity of the clusters being merged. The second grouping presents distinct subdivisions in the dendrogram. The results show that as internal variability increases, p-frames filled with content words, and different from each other, tend to form distinct groups. Thus, statistical analysis using internal variability and entropy allowed the identification of p-frames not derived from LBs.

Key Words: Corpus-Driven; Phrasal Frames; Lexical Bundles; Multivariate Data Analysis; Clusters.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

AMD – Análise Multivariada de Dados

BNC – British National Corpus

DE – Distância Euclidiana

ELexs – Estruturas Lexicais

EHI – Early Human Intervention

DHI – Delayed Human Intervention

LC – Linguística de Corpus

LA – Linguística Aplicada

LB – Lexical Bundle

PL – Pacotes Lexicais

PP – Phraseological Profile Model

TTR – Type-Token Ratio

ULs – Unidades Lexicais

SUMÁRIO

CAPÍTULO V

ANEXOS

ESTUDO DIRECIONADO POR CORPORA:

Estruturas Lexicais em um corpus especializado.

CAPÍTULO I

1 INTRODUÇÃO

Estudos linguísticos relacionados à fraseologia têm ganhado credibilidade entre os acadêmicos nas últimas décadas, principalmente quanto à formação e análise da linguagem formulaica[1]. Historicamente, as pesquisas relacionadas à linguagem formulaica teve seu início na década de 70, em que Lexicógrafos, como: Charles e Lily W. Fillmore (1977), Coulmas (1981), Pawley e Syder (1983)[2], buscavam por *multiword chunks.*

Em 1991, Sinclair (p. 109) postulou que a linguagem é um sistema probabilístico constituído de dois princípios: o de escolha aberta e o idiomático. O princípio de escolha aberta diz respeito à sequência de palavras combinadas a partir de regras gramaticais (por exemplo: *I think I might, she is a smart dresser, the effect of, the print was easy to read*), e que o preenchimento dos espaços de cada construção linguística depende das restrições gramaticais. Já o princípio idiomático refere-se à sequência de palavras que são, em parte, pré-fabricadas e adequadas a certos contextos de uso, como em: *at last, in fact, aim at, afraid that.*

A abordagem idiomática, ao invés de identificar unidades formulaicas baseadas em critérios linguísticos, utiliza uma abordagem baseada em corpus na busca de coocorrências lexicais que considera frequência e combinações de palavras[3] que não se encaixam em categorias gramaticais tradicionais.

---

[1] Wray (2002, p. 9) explica que o termo *linguagem formulaica* foi cunhado pela necessidade de um termo neutro para nomear um padrão lexical que não trouxesse uma bagagem anterior de significado. O autor segue explicando que o termo neutro *formulaic language* é comumente utilizado na literatura quando deseja-se livres associações.

[2] Sobre as pesquisas dos referidos Lexicógrafos e outros, à época dos feitos, sugiro a leitura de "Fundamentals of Formulaic Language - An Introduction", por David Wood, 2004.

[3] Villalva & Silvestre (2014, p. 76) argumentam que "palavra" é um termo multifacetado e sua "definição depende de cada campo de domínio de análise linguística (fonologia, morfologia, semântica e sintaxe). No léxico, "palavra" são unidades lexicais, ou seja, conjuntos de formas portadoras de informações fonológicas,

O pioneirismo de Sinclair nesse tipo de abordagem fraseológica ampliou a "prospecção sintagmática", pois ao contrário da clássica abordagem fraseológica que buscava distinguir diferentes categorias linguísticas e de estabelecer limites claros para a fraseologia, definiu itens fraseológicos com precedência sobre palavras isoladas (Sinclair, 2004, p. 28-29). Contudo, desde 1991, o autor já advertia quanto às dificuldades em definir linguagem formulaica em virtude da complexidade natural da linguagem e esclareceu que

> as pessoas utilizam-na para os seus próprios fins, sem normalmente terem consciência da relação entre o seu comportamento verbal e a forma como esse comportamento é caracterizado. Elas são criativas, ou convenientes, ou casuais, ou confusas; ou elas têm assuntos incomuns para colocar em palavras comuns, então elas têm que combiná-las de maneiras incomuns (Sinclair, 1991, p. 101)[4].

Diante do exposto, questões importantes emergem no cenário linguístico quanto à identificação da linguagem formulaica em textos e discursos, por exemplo, os verbos frasais (*bring up, call off, get along*) e compostos nominais (*sunflower, shoebox, raincoat*) e, em questões mais complexas, como em expressões contínuas e descontínuas com lacunas (*slots*) variáveis, ([*little, scant, less, no*] * *attention has been paid to*).

Segundo Wray (2002, p. 31-42) a linguagem é "formulaica" e, de certa forma, previsível. Ela estabeleceu alguns critérios para identificar essas características, como, por exemplo, as sequências normalmente começam com conjunções, artigos, pronomes, preposições. Além disso, a autora aponta que o preenchimento de lacunas pode ser fixo ou variável. E observou que, além desses critérios, há os relacionados aos aspectos fonológicos ou prosódicos da articulação em uma sequência formulaica. No entanto, admitiu que um padrão linguístico parece não "basear-se em um único critério, mas sim basear-se em um conjunto de características. Em vez disso, a *formulaicidade* pode ser governada por algum critério unificador que os nossos esforços até agora não conseguiram captar" (p. 43)[5].

Outras pesquisas seguiram demonstrado que há padronização na linguagem e as unidades lexicais podem ser fixas ou semifixas (Hunston e Francis, 1999; Biber, 2009).

---

morfológicas, sintáticas e semânticas e ainda de outras informações de várias ordens como, por exemplo, a etimológica".

[4]No original: "It may simply be that identification cannot be based on a single criterion, but rather needs to draw on a suite of features. Alternatively, formulaicity may be governed by some unifying criterion that our efforts so far have failed to capture".

Martinez e Schmitt (2012, p. 299)[6] sugerem que a linguagem padronizada (*formulaic language)* é "fundamental para a maneira como a linguagem é usada, processada e adquirida tanto na L1 quanto na L2". Tan e Römer (2022, p. 1-2) argumentam que é essencial desenvolver competência fraseológica, bem como, apropriar-se de linguagem formulaica se o desejo for proficiência comunicativa na língua alvo. Segundo Hunston e Francis (1999, p. 37),

> os padrões de uma palavra podem ser definidos como todas as palavras e estruturas que estão regularmente associadas à palavra e que contribuem para o seu significado. Um padrão pode ser identificado se uma combinação de palavras ocorrere com relativa frequência, se depender de uma escolha de palavra específica e se houver um significado claro associado a ela[7].

Os autores acrescentam (p. 270-271) que a maioria das palavras não possui significado isolado, e que o padrão e o significado das unidades formulaicas estão associada e são essenciais para fluência em língua materna ou adicional em ato comunicacional. Berber Sardinha (2004, p. 349-355) sugere que embora certas combinações de "traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência" e nem são aleatórios, mas são padronizados.

Nas últimas décadas, diferentes perspectivas metodológicas têm sido empregadas na identificação da linguagem formulaica: *clusters* (Hyland, 2008), n-gramas (Stubbs, 2007) e pacotes lexicais[8] (doravante, PLs) (Biber e Barbieri, 2007; Cortes, 2002). Todos esses termos (*clusters*, n-gramas e pacotes lexicais) são sequências contínuas de unidades lexicais.

Contudo, Gray e Biber (2013) explicam que outros estudos seguiram buscando por unidades fraseológicas descontínuas significativas em diferentes registros, por exemplo, Olofsson e Altenberg (1994) em um corpus de falantes de língua inglesa e Butler (1998) examinou sequências lexicais em um corpus em língua espanhola. Fletcher (2003/2004/2011/2021)[9] cunhou o termo "*phrase-frame*" ou "*p-frame*" a partir do British

[6] No original: "Whereas formulaic language was once considered a peripheral phenomenon (Ellis *et al.*, 2008), research has now established that it is fundamental to the way language is used, processed, and acquired in both the L1 and L2".

[7] No original: "The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it".

[8] Nesta pesquisa utilizamos o termo Pacote LexicaL (PL) como correspondente ao que em inglês é conhecido como *Lexical Bundle*.

[9] http://phrasesinenglish.org/index.html

National Corpus (BNC) para descrever as sequências lexicais semifixas, recorrentes e descontínuas com lacunas variáveis (p. ex., *it is * to* [*easy, hard, important*]).

Muitas dessas investigações têm sido possíveis graças à Linguística de Corpus que, segundo Berber Sardinha (2004, p. 3), é uma das subáreas da pesquisa linguística que coleta, trata e explora um corpus, isto é, um "conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para pesquisa de uma língua ou variedade linguística".

1.1 A Linguística de Corpus: área de pesquisa indispensável à linguagem formulaica.

Segundo Berber Sardinha (2004, p. 3) a Linguística de Corpus (doravante, LC) é "uma das áreas de pesquisa da linguagem mais ativa nos últimos tempos", e seu *status* dentro da área tem sido discutido sob diferentes pontos de vista, a cerca de ser um "método" [10] ou uma "metodologia" investigativa. Por exemplo, para O'Keeffe e McCarthy (2010, p. 3) é um *método* "constituído por inúmeras linhas de concordância e listas de palavras geradas por um programa de computador, com o objetivo de entender fenômenos que ocorrem em textos grandes ou em compilações de textos pequenos", em contrapartida, para Bonelli (2001, p. 14-15), é uma "base metodológica" para o estudo da linguagem.

Berber Sardinha (2004, p. 35-38) explica que a LC claramente não é uma disciplina, pois se ocupa de várias áreas do conhecimento como: o léxico, a sintaxe e o texto e, em seu entendimento, é uma "metodologia" *instrumental*, que neste caso, pode ser colocada à disposição de outras disciplinas, e segue comentando que entre os linguistas há uma tendência a definir o *status* da LC não apenas como um instrumento, mas uma *abordagem filosófica.*

Entre as áreas que se beneficiam dos construtos da LC estão: a Lexicografia, a tradução, a estilística, a gramática, a Linguística Forense, a Linguística Computacional, a Linguística Aplicada, para citar alguns exemplos. A LC se ocupada de um corpus, que pode ser definido como um conjunto de

> [...] dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal,

[10] Foge ao escopo desta pesquisa a discussão acerca de ser LC uma 'metodologia' ou um 'método' de investigação, assim, considero os termos 'método' e 'metodologia' indistintamente.

> dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador (Berber Sardinha, 2004, p. 325).

Nesta pesquisa, todavia, exploramos um corpus especializado com o objetivo de extrair unidades formulaicas descontínuas[11]. Um corpus especializado tem por "propósito compreender tipos específicos de linguagem acadêmica e profissional" (Flowerdew, 2004, p. 13), permitindo investigações que mostram as suas múltiplas funcionalidades nos mais variados contextos linguísticos (Biber e Conrad, 1999; Biber e Barbieri, 2007; Gray e Biber, 2013; Staples *et. al*., 2013).

Para Biber (2010) a LC é mais que uma metodologia de pesquisa. O autor explica que LC amplia a capacidade sistêmica do pesquisador ao observar as variações linguísticas em um corpus de modo empírico e quantitativamente, além de "documentar a existência de construções linguísticas que não são reconhecidas pelas teorias linguísticas" (p. 194). Assim, propõe duas perspectivas metodológicas: baseada em corpus (*corpus-based*) e direcionada por corpus (*corpus-driven*). A abordagem baseada em corpus busca analisar padrões "sistemáticos de variação e uso das características pré-definidas" (Biber, 2012, p. 196-201), e na perspectiva direcionada por corpus as construções linguísticas emergem da análise do corpus, não fazendo "afirmações a *priori* em relação às características linguísticas empregadas na análise do corpus" (Biber, 2012, p. 201- 203)[12].

Bocorny (2021 *et. al.*, p. 5) explica que na perspectiva direcionada por corpus a categorização do objetivo de estudo não ocorre à *priori*. Em vez disso, ela ocorre a partir da "identificação do objeto de estudo no corpus com base na sua frequência de ocorrência, para posteriores análises e classificação".

1.2 Do objetivo geral aos objetivos específicos

Nas últimas décadas, houve um aumento significativo por pesquisas relacionadas às unidades formulaicas contínuas como recurso pedagogicamente útil ao ensino e aprendizagem

[11] A discussão sobre as unidades formulaicas serem contínuas e descontínuas e, simultaneamente convencionais e variáveis, será apresentada oportunamente no capítulo de revisão de literatura.
[12] Por ser de fundamental importância para esta pesquisa, as perspectivas metodológicas "baseada em corpus" e "direcionada por corpus" serão discutidas acuradamente no capítulo de revisão de literatura.

de língua adicional[13] (p. ex., *little attention has been paid to*). No entanto, esta pesquisa tem por objetivo geral identificar e analisar os espaços variáveis e classificar as unidades formulaicas descontínuas (p. ex., * [*little, scant, less, no*] *attention has been paid to*) que são simultaneamente convencionais e variáveis (Casal e Kessler, 2020, p. 3), especificamente, as estruturas lexicais[14] (doravante, ELexs) encontradas em um corpus especializado, com o propósito de compreender a linguagem em uso em artigos acadêmicos da área da Linguística Aplicada, considerando os princípios de variabilidade (*type-token ratio* - TTR) e previsibilidade.

Os termos comumente usados (*clusters*, n-gramas e pacotes lexicais, etc.) referem-se às sequências lexicais contínuas que partem da perspectiva metodológica baseada em corpus (*corpus-based*). Poucas pesquisas têm buscado identificar diretamente as unidades descontínuas em uma perspectiva direcionada por corpus (*corpus-driven*) a partir de dados estatísticos. Esta pesquisa busca diretamente por ELexs que descontinuamente podem preencher uma lacuna, usando uma metodologia direta de extração (*direct-approach methodology*) (Gray e Biber, 2013, p. 121).

Esse objetivo geral norteou minhas leituras e reflexões sob a perspectiva metodológica direcionada por corpus, levando-me a desenhar um estudo que identificasse as unidades formulaicas diretamente. A fim de atingir esse objetivo de identificar as ELexs *em um corpus especializado de artigo da linguística aplicada*, foram formuladas os seguintes objetivos específicos e as seguintes perguntas de pesquisa:

Os objetivos específicos são:

1. Investigar os ELexs mais frequentes no CorAAL;
2. Identificar as estruturas dos ELexs encontrados;
3. Identificar quais resultados sobre os ELexs do CorAAL podem nortear a criação de tarefas segundo a aprendizagem orientada por dados (*Data-Driven Learning*[15]).

[13] Neste trabalho é utilizado o conceito de língua adicional discutido por Leffa & Irala (2014, p. 33) como sendo uma língua "[...] construída a partir da língua ou das línguas que o aluno já conhece. O sistema, incorporando principalmente o léxico e a sintaxe, é construído sobre a língua já conhecida, às vezes estabelecendo contrastes,"[...] com a língua alvo.

[14] Nesta pesquisa utilizamos o termo Estruturas Lexicais (ELexs) como correspondente ao que em inglês é estudado como *Phrasal frame* (*p-frame*).

[15] Segundo Gilquin e Granger (2010, p. 359-370), a aprendizagem orientada por dados (em inglês, Data-Driven Learning - DDL) "consiste em usar as ferramentas e técnicas da Linguística de Corpus para fins pedagógicos". Hyland (2002, p. 120) argumenta que DDL é uma abordagem que "[...] na pedagogia da aprendizagem de línguas [...], fornece meios 'para que os alunos possam assumir papéis mais ativos, reflexivos e autônomos em sua aprendizagem".

As perguntas de pesquisa são:

1. Quais são as mais frequentes ELexs encontradas no corpus especializado – CorAAL?
2. Quais são as características estruturais das ELexs identificadas no corpus especializado - CorAAL?
3. Quais aspectos dos resultados das ELexs do CorAAL podem nortear a criação de tarefas que propiciem aprendizagem orientada por dados?

Além deste capítulo introdutório, no qual apresentei uma contextualização da pesquisa, bem como o objetivo geral e os específicos, este trabalho é composto de mais quatro capítulos. O segundo capítulo corresponde ao referencial teórico com o resultado de minhas leituras: sobre unidades fraseológicas contínuas e descontínuas; sobre o corpus e suas abordagens metodológicas; sobre o tratamento Estatístico aplicado aos dados linguísticos. O terceiro capítulo diz respeito à abordagem metodológica adotada nesta pesquisa, em que conceituo a pesquisa, descrevo todos os procedimentos adotados e, em seguida, os procedimentos metodológicos de coleta e análise de dados. O quarto capítulo aborda a análise e discussão dos dados coletados durante a pesquisa. Apresento evidências da possibilidade de se identificar as ELexs independentemente dos Pacotes Lexicais (PLs), destacando a análise estatística multivariada aplicada ao corpus especializado. No quinto capítulo são apresentadas as considerações finais, onde retomo os objetivos de pesquisa e discuto as implicações para futuras investigações na área em que esta pesquisa se insere.

CAPÍTULO II

2 REVISÃO DE LITERATURA

Estudos fraseológicos, enquanto disciplina com um escopo próprio, ganharam prestígio nas últimas décadas. Contudo, Sinclair (2008) explica que, no âmbito da Linguística, em virtude do tradicionalismo entre os pesquisadores, ou pela falta de clareza quanto à sua área de pesquisa (gramatical, lexical ou semântico), e a priorização de aspectos sintagmáticos em detrimento aos aspectos paradigmáticos, não havia espaço para a fraseologia "[...] no aparato tradicional de análise da linguagem, sendo muitas vezes ignorada" (p. xv)[16]. A análise sintagmática se concentra na estrutura gramatical, a abordagem paradigmática volta-se para padrões lexicais, ou seja, combinações formulaicas no texto.

O termo 'fraseologia' não é de fácil consenso. Em linhas gerais, o termo é definido como o estudo de combinação de palavras; no entanto, entre os especialistas a definição não é simples. Por exemplo, Howarth (1998, p. 25) critica a ausência de definição do termo fraseologia e argumenta que essa ausência faz surgir muitos termos correlatos, como: fórmulas, linguagem pré-fabricada ou pronta, *chunks* que, a depender do fenômeno estudado, diferem entre si. Ele propõe uma definição que a caracteriza como sendo uma "combinação de palavras com uma função sintática com constituintes de sentenças (como substantivos ou frases preposicionais ou construções de verbos e objetos)".

Gries (2009, p. 4) define fraseologia como a "coocorrência de uma forma de um item lexical ou mais elementos linguísticos de vários tipos que funcionam como uma unidade semântica em uma sentença e cuja frequência de coocorrência é maior do que o esperado". Para o autor, é necessário rigoroso parâmetro a ser observado quanto à frequência da coocorrência do item lexical para que haja um fenômeno fraseológico, confome Quadro 01.

[16] No original: "But there is a penalty for adopting a holistic strategy; there is no place for phraseology in the traditional apparatus of language analysis, so it is often just ignored".

Quadro 01 –Definição fraseológica segundo segundo Gries.

a) a ***natureza*** dos elementos envolvidos em uma unidade fraseológica;	Uma unidade fraseológica é coocorrência de uma forma ou item lexical ou qualquer outro tipo de elemento linguístico;
b) o ***número*** de elementos envolvidos em uma unidade fraseológica;	Podem consistir em apenas dois elementos (como pares de palavras) ou podem incluir um número maior de elementos.
c) o ***número de vezes*** na expressão deve ser observada antes de contar como uma unidade fraseológica;	Uma expressão é fraseologicamente aceita se a frequência de ocorrência observada for maior do que a esperada.
d) a ***distância*** entre os elementos envolvidos em uma unidade fraseológica;	Enquanto alguns trabalhos (especialmente estudos baseados em n-gramas no processamento de linguagem natural) preocupam-se apenas com elementos imediatamente adjacentes, o autor adota uma perspectiva mais ampla e mais difundida que também reconhece unidades fraseológicas descontínuas.
e) o ***grau*** de flexibilidade lexical e sintática dos elementos envolvidos;	Em unidades fraseológicas há padrões completamente inflexíveis e padrões relativamente flexíveis e, também, padrões parcialmente preenchidos lexicalmente.
f) o ***papel*** que a unidade semântica desempenha na unidade fraseológica.	Os elementos de uma unidade fraseológica - independentemente de como sejam distribuídos em uma sentença - são geralmente assumidos como funcionando como uma unidade semântica, ou seja, tendo um sentido exatamente como um único morfema ou palavra.

Fonte: Gries (2009, p. 4)

Römer (2010, p. 95) explica que em uma unidade fraseológica não há separação entre o léxico e a gramática, e a palavra isoladamente não constitui a principal unidade de significado, mas a sentença que, internamente, pode permitir variações (por exemplo: *would be interesting, it would be very interesting*). Assim, Römer (2010, p. 97) desenvolve um modelo do perfil fraseológico (do inglês, *phraseological profile model – PP model*) que permite ao pesquisador elaborar um inventário fraseológico de um corpus ou texto, conforme descrito abaixo no Quadro 02.

Quadro 02 – Inventário fraseológico segundo Römer

a) a identificação do item fraseológico	A extração das unidades fraseológicas deve ser feita automaticamente usado softwares especializados – p. ex.: Collocate (Barlow, 2004), ConcGram (Greaves, 2007), KfNGram (Fletcher, 2007), AntConc (Anthony, 2022)
b) a determinação interna do item lexical variável (p. ex.; A*CD, AB*D) – o * indica o item variável.	Nesta etapa, verifica-se a frequência (*type*) e a variação dos itens lexicais encontrados na fase anterior.
c) exame funcional do item lexical.	Identificar a função das unidades lexicais encontradas na fase anterior
d) análise da distribuição das unidades formulaicas no texto/corpus	A distribuição das unidades lexicais no texto através de software especializados - AntConc (Anthony, 2022)

Fonte: Römer (2010, p. 97)

Em termos gerais, a fraseologia investiga a combinação formulaica, ou seja, padronizada. No entanto, surge a questão: o que constitui e como se forma um padrão lexical? Hunston e Francis (1999, p. 37) dedicam-se a explorar esse tema e definem padrão como

> todas as palavras e estruturas que estão regulamente associadas a uma palavra que contribuem para seu significado. Um padrão pode ser identificado se a combinação de palavras ocorrer com relativa frequência e, se depender da escolha de uma palavra específica, e se houver um significado claro associado a ela[17].

Desse modo, o padrão em uma ocorrência se estabelece pela seleção de palavras que se encontram ao redor de uma palavra em uma linha de concordância. Por exemplo, normalmente o verbo exige complemento, então, é comum encontrá-los à sua direita, como em: *he <u>decided</u> to leave*, *he <u>hated</u> leaving*. Os autores Hunston e Francis (1999) explicam que os substantivos são análogos aos verbos, pois normalmente têm seus complementos à direita (*his <u>decision</u> to leave*, *the <u>theory</u> of evolution*), enquanto os adjetivos tendem a modificar substantivos quando posicionados à esquerda e, frequentemente, podem vir precedidos ou não por verbos de ligação.

[17] No original: “The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it”.

Quadro 03 - Quadro sinóptico dos principais padrões lexicais segundo Hunston e Francis.

The patterns of verbs	The verb is followed by a single noun group, adjective group, or clause, yielding the following patterns:
V n	I broke my left leg
V pl-n	The research compares two drugs
The patterns of nouns	Patterns with elements preceding the noun
***a* N; *the* N -**	The noun is preceded by an indefinite or definite article: *a cinch, a standstill; the blues, the bourgeoisie.*
poss N -	The noun is typically preceded by a possessive determiner like *my* or *your*, or a possessive formed from a noun group: *She had tidied away her possessions.* *I give you my word* *My husband's sister came to stay.*
The patterns of adjectives	
ADJ -ing	*I felt uncomfortable watching him.*
ADJ to-inf	*The print was easy to read.*

Fonte: Hunston e Francis (1999, p. 51-58)[18]

Segundo Wray (2002, p. 25), especialmente em análises de corpora linguísticos, a frequência tem sido um dos parâmetros mais comumente empregados para se estabelecer os padrões de distribuição das palavras no texto. O crescente interesse pela linguagem formulaica nas últimas décadas pode ser atribuído, em grande parte, à utilização de corpora como principal fonte de dados para análises linguísticas. Os corpora possibilitam a identificação de eventos linguísticos repetidos, permitindo o cálculo de frequência e o de uso de medidas estatísticas.

2.1 Linguística de Corpus - Do grande corpus ao pequeno corpus.

[18] O quadro sinóptico dos principais padrões lexicais segundo Hunston e Francis (1999) na íntegra, encontra-se em anexo.

Segundo Berber Sardinha (2004, p. 3), por corpus entende-se um "[...] conjunto de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística". E, em investigações de fenômenos linguísticos específicos, o corpus especializado tem por "propósito compreender tipos específicos de linguagem acadêmica e profissional" (Flowerdew, 2004, p. 13). O corpus especializado permite investigações formulaicas com múltiplas funcionalidades em diversos contextos linguísticos (Biber; Conrad, 1999; Biber; Barbieri, 2007; Gray; Biber, 2013; Staples *et al*., 2013).

Historicamente, o corpus grande, com o passar do tempo, torna-se pequeno. Inicialmente, faremos uma retrospectiva, mesmo que sucinta, quanto ao desenvolvimento das dimensões dos corpora; em seguida, comentaremos sobre a necessidade de grandes corpora. Desse modo, será possível explicar a aplicabilidade tanto de grandes quanto de pequenos corpora em estudos linguísticos, bem como destacar a diferença entre os dados encontrados; então, apontaremos os benefícios de estudos usando pequenos corpora; e por fim, a apresentação da Linguística de Corpus como uma metodologia investigativa.

Na década de 60, Nelson Francis e Henry Kucera, na Brown University, USA, compilaram usando tecnologia computacional uma amostra de um milhão de palavras do inglês americano formando o *Brown Corpus Manual*, publicado em 1961. E, na Escócia, na Universidade de Edimburgo, um corpus de transcrições de falas informais de falantes de inglês britânico, com cerca de 300.000 palavras, estava sendo formado. Segundo Ghadessy *et al*. (1996 *apud* Sinclair *et al.*, 1970), esses dados foram processados na Universidade de Manchester e o projeto foi transferido para Birmingham em 1965. Na década de 70, em Birmingham, outros *corpora* foram compilados contendo textos completos, em vez das 2000 palavras do estilo Brown. Para os padrões da época esses corpora eram considerados grandes, no entanto, para padrões atuais são pequenos.

A capacidade de armazenamento dos computadores atuais possibilita a compilação de dados infinitamente superiores aos daquela época e, provavelmente em um futuro não muito distante, os processadores atuais serão considerados obsoletos. Assim, *corpora* que atualmente são considerados grandes poderão ser considerados pequenos em virtude do desenvolvimento tecnológico, desse modo, o ciclo vai se repetir.

Berber Sardinha (2000, p. 345) problematiza qual a extensão necessária para que um corpus seja considerado representativo. O autor considera a representatividade dos dados como um dos critérios fundamentais para se caracterizar um corpus, no entanto, pouco se tem

pesquisado sobre os critérios mínimos para a validação de um corpus. Porém, sugere três abordagens possíveis: a impressionística, histórica e estatística.

A primeira, a impressionística, diz respeito à experiência dos especialistas quanto à criação e exploração de corpora e exemplifica apontando Aston (1997), que sugere de 20 a 200 mil palavras para corpus pequeno e 100 milhões ou mais para corpora maiores e, para Leech (1991), 1 milhão de palavras seria o mínimo para constituição de um corpus. Outros mais vagos, como Sinclair (2008, p. viii), que comenta que "o corpus deve ser o maior possível dentro do que pode ser atingido com a tecnologia da época", ou seja, depende do desenvolvimento tecnológico empregado à época de sua formação.

A segunda abordagem, a histórica, sugere o monitoramento dos corpora usados por especialistas. Berber Sardinha (2000, p. 109) explica que nesta abordagem faz-se um levantamento dos patamares dos *corpora* que estão sendo usados pelos pesquisadores. Este procedimento em muito se assemelha à prática de Linguística de Corpus - a observação dos dados em uso. Portanto, o critério na abordagem histórica é o da *aceitabilidade*, ou seja, o que é aceitável junto à comunidade de pesquisadores para delinear a extensão do corpus para um tipo específico de pesquisa.

Nas abordagens estatísticas, a fundamentação encontra-se nas teorias estatísticas. De acordo com Berber Sardinha (2000, p. 105) a abordagem estatística se divide em três perspectivas, a saber: interna, externa e relativa. A perspectiva interna analisa as várias maneiras pelas quais as características linguísticas estão distribuídas dentro e entre os textos; além de avaliar a distribuição e as implicações dessas distribuições no corpus, essa é a perspectiva seguida por Biber (1993), enquanto a perspectiva externa "depende de uma fonte de referência cuja dimensão é reconhecida" pela comunidade de pesquisadores; a perspectiva relativa se estabelece matematicamente quanto à "quantidade de vocabulário é necessário para incluir certas classes gramaticais"[19].

2.2 A aplicabilidade de corpus de "pequena" e de "grande" escala.

Até recentemente, a frequência com que determinadas combinações de palavras apareciam no corpus era o único aspecto considerado para determinar se o corpus era "pequeno" ou "grande", contudo, atualmente, a metodologia empregada na compilação do corpus é, também, um parâmetro contrastivo. Para Sinclair (2001, p. xi), um corpus pequeno

[19] Esta pesquisa considera uma perspectiva estatística de análise de agrupamento (*clusters*), e discutiremos suas propriedades no item 3.4.

“é visto como um conjunto de evidências relevantes e confiáveis, e é pequeno o suficiente para ser analisado manualmente, ou é processado pelo computador de forma preliminar”.

Metodologicamente, o autor segue classificando os pequenos corpora como projetados para rápida intervenção humana (do inglês, *early human intervention* - EHI), enquanto grandes corpora são projetados para intervenção humana tardia (do inglês, *delayed human intervention* - DHI). Na perspectiva DHI, o processo de compilação do corpus ocorre em sucessivas sessões do EHI, cabendo ao pesquisador intervir na interpretação dos resultados e, em contrapartida, nos processos de compilação EHI o corpus é compilado com um objetivo específico de pesquisa.

Nesse contexto, a LC se insere como proposta metodologica no campo da Linguística como ferramenta de investigação. Berber Sardinha (2000, p. 325) explica que Linguística de Corpus

> ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador.

A Linguística de Corpus, por décadas, tem demonstrado que há padronização na linguagem em unidades lexicais fixas ou semifixas (Sinclair, 1991; Biber, 2009; Hunston e Francis, 1999; Martinez e Schmitt, 2012) e, isso se dá em virtude do significado e o valor dos grandes corpora (*Large Corpora*), no entanto, pouco se tem falado sobre o valor experimental de pequenos corpora especializados (*Small Specialized Corpora*).

Sinclair (2001, p. vii) justificou que ao longo sua carreira acadêmica sempre trabalhou com grandes corpora e, por isso, muitas vezes foi mal interpretado como tendo uma opinião negativa quanto a corpus pequeno, no entanto, explica que o tamanho do corpus certamente traz limitações a certas pesquisas, mas não quanto à sua qualidade.

Assim, o autor segue comentando que o objetivo da pesquisa determina o tamanho do corpus. Por exemplo, se a pesquisa busca compreender os padrões contextuais de uma palavra, então quanto maior o corpus, maiores serão as possibilidades contextuais, no entanto, em uma pesquisa que busca compreender um fenômeno específico, um corpus pequeno poderá ser útil, pois neste caso, quando se conhece o padrão estudado, será mais fácil replicar os resultados em um corpus maior.

Biber (2012) explica que a identificação de padrões lexicais pode ser realizada por meio de uma abordagem baseada em corpus (*corpus-based*) ou orientada por corpus (*corpus-driven*). A primeira abordagem visa analisar os padrões "sistemáticos de variação e uso de características [semânticas/gramaticais] pré-definidas" pelo pesquisador (2012, p. 196-201), enquanto na abordagem orientada por corpus, as "construções linguísticas emergem da análise do corpus", isto é, não há "afirmações a priori em relação às características linguísticas empregadas na análise do corpus" (Biber, 2012, p. 201-203). Bocorny (2021 *et al.*, p. 5) acrescenta que na perspectiva baseada por corpus, a categorização do objetivo de estudo não ocorre a *priori*, mas sim a partir da "identificação do objeto de estudo no corpus com base na sua frequência de ocorrência, para posteriores análises e classificações".

2.3 Abordagem orientada por corpus e o padrão linguístico.

A abordagem direcionada por corpus é indutiva (Biber 2012, p. 196). Nesta perspectiva analítica, com o auxílio de ferramentas computacionais analíticas, os padrões das variações linguísticas[20] emergem do corpus devido à frequência da linguagem em uso. Tognini-Bonelli (2001, p. 84) argumenta que as evidências fornecidas pelo corpus são independentes de categorias linguísticas e a "ausência de padrões é considerada potencialmente significativa" na análise, além do mais, tais evidências podem conduzir a hipóteses e, estas, a afirmações teóricas. Desse modo, podemos definir a abordagem direcionada por corpus como sendo uma análise que parte de padrões recorrentes e das distribuições de frequência que emergem da linguagem em contexto.

As análises que se baseiam em corpora buscam explorar as capacidades de um corpus, não com o intuito de respaldar teorias preexistentes, sobretudo, para identificar categorias e unidades linguísticas que ainda não foram reconhecidas anteriormente. Sinclair (1991, p. 16), por exemplo, explica que não há equivalência nas formas no singular e plural dos substantivos "*eye*" e "*eyes*". Geralmente, o plural ocorre com adjetivos (*blue eyes, brown eyes, manic eyes*) ou com pronomes possessivos (*your eyes, his eyes, her eyes*), enquanto a forma singular

[20] Embora a Linguística de Corpus utilize o termo "variações linguísticas" para realizar uma análise mais sistemática, proporcionando uma compreensão mais aprofundada de como a língua é empregada em diferentes situações e por diversos grupos de falantes. Isso inclui influências geográficas, sociais, históricas ou contextuais e manifesta-se em vários níveis linguísticos, como fonético, fonológico, morfológico, sintático, lexical e pragmático, nesta pesquisa, a utilização do termo "variações linguísticas" é geral e está alinhada com a perspectiva de Gray (2015, p. 10).

raramente faz referência a partes do corpo, frequentemente é utilizada em expressões fixas (*make eye contact, catch your eye, in my mind's eye*). Assim, o autor afirma que a variedade de significados das palavras, na verdade, se destaca por meio de suas aplicações específicas.

Por outro lado, conforme observado por Francis (1993, p. 139-150), certos padrões "selecionam palavras com significados específicos". Por exemplo, os substantivos presentes em *that-clauses* (*argument that, decision that, fact that, problem that, sorrow that*) podem ser categorizados em seis grupos distintos: processos ilocucionários (*observation that, recommendation that*); atividade linguística de algum tipo (*message that, point that*); estados mentais em relação a questões específicas (*conviction that, hypothesis that*); processos de pensamento ou seus resultados (*analysis that, conclusion that*); sentimentos e atitudes (*anger that, fear that*); e substantivos gerais (*advantages that, fact that*).

Considerando o exposto, evidencia-se um padrão na configuração determinada pela escolha das palavras próximas a um termo específico em uma linha de concordância. Desse modo, cada palavra em um corpus pode manifestar uma diversidade de padrões, conforme enfatizado por Hunston e Francis (1999).

É importante destacar que, em uma abordagem orientada por corpus, a pesquisa não se desenvolve de maneira mecânica. Em outras palavras, os dados não surgem de forma espontânea; eles são constantemente moldados pela intervenção do pesquisador-linguista, que aplica seu conhecimento e experiência em todas as fases da pesquisa (Tognini-Bonelli, 2001, p. 85).

2.4 Pacotes Lexicais vs. Estruturas Lexicais.

Como anteriormente destacado, a abordagem direcionada por corpus baseia-se em padrões recorrentes e distribuições de frequência que surgem da linguagem em contexto. Assim, embora "classes gramaticais e estruturas sintáticas não tenham *status* a priori na análise", o ponto de partida é a palavra (Biber, 2010, p. 201). Neste contexto, apresentamos duas contribuições provenientes dessa abordagem metodológica: os pacotes lexicais e as estruturas lexicais.

Os Pacotes Lexicais (Biber, 2007; Cortes, 2002) e as Estruturas Lexicais (Römer, 2009; Yoon e Casal 2020, p. 463) são termos utilizados para descrever distintos tipos de composições linguísticas. Em outras palavras, são fenômenos linguísticos que requerem uma análise por parte do pesquisador, uma vez que a diferenciação nem sempre é evidente.

Resumidamente, os pacotes lexicais (doravante, PLs) referem-se a expressões contínuas que ocorrem juntas (p. ex.: *on the other hand*), enquanto as estruturas lexicais (doravante, ELexs) consistem em estruturas descontínuas, nas quais as lacunas podem ser preenchidas com palavras diferentes mantendo a mesma estrutura básica (p. ex.: [*little, scant, less, no*] * *attention has been paid to*). Respectivamente, procedemos à descrição das diferenças entre ambas.

Os PLs constituem as sequências recorrentes de múltiplas palavras que se repetem com frequência e são distribuídas em diferentes textos. Contudo, é evidente que essas sequências apresentam características singulares, tanto estruturais quanto funcionais, embora raramente se alinhem às estruturas linguísticas completas reconhecidas pelas teorias linguísticas contemporâneas. Conforme Biber (2010, p. 203), os PLs "não são estruturalmente completos e não têm significado idiomático".

Em uma abordagem direcionada por corpus, os PLs são estabelecidos exclusivamente como base nos critérios de distribuição e sequência das palavras nos textos. Entretanto, após a identificação, torna-se possível realizar análises estruturais e funcionais de maneira sistemática. Além disso, diferem das expressões idiomáticas em, pelo menos, três aspectos: a) são extremamente comuns; b) carecem de significado literal e não se destacam no texto e c) geralmente, não apresenta uma unidade lexical completa (Biber *et al.*, 1999, p. 990-991). Ao contrário, as expressões idiomáticas possuem estruturas e significados definidos, muitas vezes adquiridos pelo uso ou contexto (p. ex.: *kick the bucket, a slap in the face*).

Embora não constituam unidades estruturalmente completas, Biber (2010, p. 205) destaca que esses agrupamentos apresentam fortes correlações gramaticais. Por exemplo, o PL *you want me to* é construído a partir de estruturas gramaticais (V-pron), enquanto *in the case of* (N-prep), e diversos outros PLs incorporam fragmentos verbais, como: *it's going to be* e *what do you think*. Adicionalmente, alguns PLs são compostos por orações dependentes, exemplificadas por *when we get to* e *that I want to*.

Adicionalmente, Biber e*t al*. (1999, p. 377) observam que a maioria dos PLs atua como ponte entre duas estruturas linguísticas. Em conversações, estas estruturas conectam duas partes distintas do discurso (p. ex.: *I want to know, well that's what I*), enquanto nos discursos acadêmicos, geralmente, conectam sentenças (p. ex.: *in the case of, the base of the*).

Os agrupamentos lexicais atuam como um suporte ou estrutura discursiva, inserindo informações adicionais no texto. Em outras palavras, comumente expressam novos significados interpretativos para o discurso em progresso por meio de informações

proposicionais. Segundo Biber (2010, p. 206), os PLs fornecem estruturas interpretativas para o discurso, por exemplo:

> *I want you to write a very brief summary of his lecture. Hermeneutic efforts are provoked by the fact that the interweaving of system integration and social integration [. . .] keeps societal processes transparent . . .*

Quanto às funções discursivas, Biber *et al*. (2004, p. 384) classificam os PLs como sendo: a) expressões referenciais, b) expressões de jugamento e c) organizadores de discurso. As expressões referenciais apontam para coisas tangíveis ou conceituais, ou para o contexto. Elas servem para identificar a entidade mencionada ou enfatizar algum atributo específico como sendo mais importante. Essencialmente, são construções linguísticas que direcionam a atenção para algo específico no discurso, seja real ou abstrata, ou ainda elementos presentes no texto (p. ex.: *that's one of the, and this is a*).

As expressões de jugamento referem-se às atitudes ou avaliações de certeza, refletindo, em outras palavras, opiniões, pontos de vista ou níveis de convicção (p. ex.: *I don't know if, it is possible to*). Enquanto os organizadores de discurso desempenham o papel de conectar o que foi dito anteriormente com o que será abordado em seguida. Esses organizadores são estruturas linguísticas que ajudam a guiar a compreensão estabelecendo conexões lógicas entre as partes do discurso (p. ex.: *what do you think, take a look at*). Isso é crucial para a compreensão e continuidade do discurso mantendo a clareza, coesão e fluidez na escrita acadêmica.

Para complementar a abordagem apresentada por Biber *et. al* (2004), Tan e Römer (2022, p. 4) argumentam que, dada a diversidade nas unidades lexicais, é possível que surjam outras funções nos textos, destacando-se as expressões de conversação e as associadas a atividades (p. ex.: *to * for the* [*apply, pay, compete, look,wait*]). Importante ressaltar que, ao longo desta pesquisa, direcionamos nossa atenção para as expressões associadas a atividades, as quais estão intrinsecamente ligadas às ações do sujeito oracional.

Ao contrário dos PLs, as ELexs são caracterizadas por sua flexibilidade (Römer, 2009; Yoon e Casal 2020, p. 463). Essas estruturas permitem que suas lacunas sejam preenchidas com palavras diferentes, enquanto a estrutura básica permanece a mesma (p. ex.: [*little, scant, less, no*] * *attention has been paid to*). Isso as classifica como estruturas fraseológicas descontínuas, destacando sua capacidade de adaptar-se a diversas expressões sem alterar sua estrutura fundamental.

Em comparação com as extensivas pesquisas sobre PLs, nota-se uma relativa escassez de pesquisas dedicadas às unidades fraseológicas descontínuas, também conhecidas como ELexs, nas últimas décadas (Renouf e Sinclair, 1991; Biber 2009; Römer 2010; Gray & Biber, 2013; Garner 2016; Tan e Römer, 2022).

Em um estudo pioneiro sobre unidades fraseológicas descontínuas, Renouf e Sinclair (1991, p. 128), investigaram ELexs, as quais denominaram *collocational frameworks*, ao examinar as palavras funcionais que preenchiam as lacunas em *a + * + of*. Os autores observaram que os espaços não eram preenchidos aleatoriamente, mas sim por agrupamentos semânticos específicos. Além disso, notaram que os padrões estavam relacionados aos elementos circundantes.

Biber (2009) introduz a abordagem direcionada por corpus em sua pesquisa, visando investigar variações em PLs em conversações e em textos acadêmicos. O autor emprega uma metodologia por ele denominada *bundles-to-frame*, a qual consiste a partir dos PLs, identificar as variações possíveis nos espaços das unidades lexicais marcadas por * (p. ex.: 1*34, 12*4, *234, 123*). Em 2010, Römer (p. 309-325) utilizando metodologia *bundle-to-frame* proposta por Biber (2009), destaca a importância do perfil fraseológico de um tipo de texto. Em outras palavras, compreender como certos termos ou expressões são utilizados de maneira recorrente em um determinado tipo de texto proporciona *insights* sobre a natureza fraseológica da língua em questão. Além disso, Römer (2010) sugere que esse entendimento contribui para a compreensão da criação de significado no discurso, destacando a relevância das expressões utilizadas na construção do sentido em textos acadêmicos.

Gray e Biber (2013, p. 128) examinaram PLs e ELexs em textos acadêmicos e em conversações utilizando a metodologia de abordagem direta (*Direct Approach*) buscando identificar e classificar padrões de previsibilidade nas unidades lexicais. Os pesquisadores constataram que as unidades lexicais com baixo índice de previsibilidade, de modo geral, não estão associadas a PLs, e concluíram que a variação fraseológica das ELexs na escrita acadêmica está intrinsecamente ligada às construções gramaticais.

Dessa forma, é possível observar algumas características importantes que diferenciam as ELexs dos PLs: o grau de variabilidade e previsibilidade, bem como a estrutura e os preenchimentos variáveis das lacunas (Biber, 2009, Gray e Biber, 2013). No intuito de estudar o grau de variabilidade dessas estruturas lexicais, foi utilizada nesta pesquisa a razão *Type/Token ratio* (TTR), conforme proposta por Tan e Römer (2022, p. 4). O TTR capta uma relação entre diferentes palavras que preenchem o espaço em branco (*) em relação ao número total de ELexs. Os valores de TTR variam de 0 a 1, onde o TTR próximo a 1 indica

alta variabilidade, isto é, uma alta proporção de tipos de variantes na lacuna (*) por estrutura lexical. O cálculo do TTR é derivado da porcentagem de um tipo específico de variante em relação à frequência total das ELexs. A fórmula pode ser expressa da seguinte forma:

$$\frac{\text{Frequência de tipos variantes (preenchimento)}}{\text{Frequência (tokens)das estruturas lexicais}} \text{ X } 100$$

Por meio desta fórmula, calcula-se a porcentagem ao obter a razão da frequência de um tipo específico de variante em relação à frequência total das ELexs e, em seguida, multiplica-se o resultado por 100 para expressá-la como uma porcentagem.

Além da variabilidade, as ELexs são caracterizadas pela sua previsibilidade. Conforme Gray e Biber (2013), o grau de previsibilidade determina se uma ELex possui preenchimento fixo em seus espaços, indicando o nível de incerteza de uma distribuição de probabilidade (Kumar *et al*., 1986). Os tipos de variantes nas lacunas das ELexs variam de 0 a 1, onde valores mais próximos de 1 indica uma distribuição mais uniforme com maior probabilidade de ocorrência. Assim, a média de previsibilidade é calculada considerando o número de *tokens* do preenchimento mais frequente nos espaços. Gray e Biber (2013) destacam que ELexs com níveis elevados de previsibilidade e alta frequência estão associadas a PLs, ao passo que aquelas com baixa previsibilidade não possuem preenchimento fixo de seus espaços e, assim, não estão vinculadas a PLs. A fórmula pode ser expressa da seguinte forma:

$$\frac{\text{Frequência de preenchimento}}{\text{Frequência das estruturas lexicais}} \text{ X } 100$$

Desse modo, o número de vezes que um determinado preenchimento (variante) ocorre em uma ELex é proporcional à razão do número total de ocorrências de todas as ELexs resultando em uma medida em porcentagem. Isso fornece uma indicação da frequência do preenchimento. Em outras palavras, quanto maior o resultado, maior a representatividade desse preenchimento em relação ao total de ELexs.

Além dos níveis de variabilidade e previsibilidade, Gray e Biber (2013), propuseram uma classificação quanto à estrutura dos elementos das estruturas lexicais: com palavra de conteúdo (p. ex.: *results * that*), com palavras funcionais (*a * of*) e baseadas em verbo (*is * to*). As palavras de conteúdo incluem substantivos, verbos, adjetivos e advérbios, enquanto as palavras de função abrangem determinantes, preposições e pronomes. Cada ELex está

associada a uma dessas categorias de preenchimento, facilitando a classificação e a compreensão de como as palavras se inserem nas lacunas das ELexs.

Embora haja poucas pesquisas dedicadas à análise de sequências fraseológicas descontínuas, os estudos existentes desempenham um papel significativo na compreensão da escrita acadêmica, revelando em que medida as variações nas ELexs são de importância fundamental para entender as tendências de padrões textuais (Römer, 2010). Ademais, observa-se uma lacuna na literatura quanto à análise de sequências fraseológicas descontínuas na área da Linguística Aplicada, especialmente no que diz respeito à aplicação de recursos estatísticos.

Quadro 04 – Resumo das características dos PLs e das ELexs.

Pacotes Lexicais	Estruturas Lexicais
São expressões contínuas (p. ex.: *on the other hand*).	São expressões descontínuas. As lacunas podem ser preenchidas com palavras diferentes mantendo a mesma estrutura, (p. ex.: [*little, scant, less*] * *attention has been paid to*)
São sequências recorrentes de múltiplas palavras que se repetem com frequência.	São flexíveis. As lacunas são preenchidas com palavras diferentes, enquanto a estrutura básica mantem-se preservada. Assim, adaptam-se a diversas expressões sem alterar sua estrutura fundamental.
É possível fazer análise estrutural e funcional de modo sistemático.	Quanto aos elementos estruturantes, podem ser palavras de conteúdo (substantivos, verbos, adjetivos, etc) e/ou funcional (determinantes, preposições e pronomes).
Geralmente conectam sentenças (p. ex.: *in the case of, the base of the*).	
São interpretativas, isto é, atuam como suporte ou estruturas discursivas.	
Não são expressões idiomáticas.	
Não possuem unidades lexicais completas, porém possuem fortes ligações gramaticais (p. ex.: you want me to (V pron) e in the case of (N prep.).	

Fonte: Própria

2.5 Estatística Multivariada e a Linguística

Neste contexto, esta pesquisa se insere ao investigar a aplicação da Estatística como uma ferramenta científica nos estudos linguísticos. Em outras palavras, nesta seção, dedicamo-nos a explorar a análise quantitativa de dados linguísticos utilizando a linguagem R para análises estatísticas. Diversos dados linguísticos (morfológicos, sintáticos, semânticos) apresentam variações, e isso inclui as unidades fraseológicas. Essa variabilidade destaca a diversidade e a adaptabilidade da linguagem em diversos contextos. Assim, o tratamento estatístico faz-se necessário, não somente para a quantificação dos dados, mas também para o propósito interpretativo dos fenômenos linguísticos.

Segundo Oushiro (2022), a análise estatística dos dados possibilita descrever um fenômeno por meio de gráficos e tabelas. Devido à quantidade extensa de dados, muitas vezes, é impraticável resumi-los em dimensões compreensíveis. Assim, as tabelas e os gráficos tornam-se valiosos para representar os dados, embora sua eficácia em explicá-los seja limitada quando consideradas isoladamente. A interpretação dos padrões deve estar sempre em sintonia com as teorias e modelos linguísticos adotados no processo de análise.

Esta pesquisa se baseia em uma das vertentes da Estatística, mais precisamente na Análise Multivariada aplicada a dados linguísticos (AMD). O termo "multivariado", em princípio, é abrangente, englobando todas as técnicas estatísticas que analisam simultaneamente múltiplas medidas relacionadas a indivíduos ou objetos sob investigação. Hair *et al*. (2009) afirmam que, inicialmente, qualquer análise que envolva mais de duas variáveis pode ser classificada como multivariada.

Para o escopo desta pesquisa, alinhamos nossa perspectiva com o conceito de Hair *et al*. (2009, p. 23), os quais observam que, em uma análise multivariada,

> todas as variáveis devem ser aleatórias e inter-relacionadas de tal maneira que seus diferentes efeitos não podem ser significativamente interpretados em separado. Alguns autores estabelecem que o objetivo da análise multivariada é medir, explicar e prever o grau de relação entre variáveis estatísticas (combinações ponderadas de variáveis). Assim, o caráter multivariado reside nas múltiplas variáveis estatísticas (combinações múltiplas de variáveis), e não somente no número de variáveis ou observações.

Assim, o objetivo central dessa abordagem é quantificar, explicar e antecipar a intensidade das relações entre as variáveis estatísticas[21], as quais são essencialmente combinações ponderadas de diferentes variáveis. Ressaltamos que a ênfase na análise multivariada não se limita à quantidade de elementos, mas sim à complexidade das interações entre os elementos, destacando a necessidade de uma análise abrangente para compreender as relações existentes.

Em uma análise multivariada lidamos simultaneamente com diversas variáveis. Devido à própria essência, essa técnica permite identificar relações complexas entre as variáveis tornando-as acessíveis para interpretação e compreensão. Em outras palavras, as técnicas multivariadas são valiosas para explorar e compreender relações que não seriam facilmente percebidas ao analisar cada variável isoladamente.

A análise multivariada, segundo Hair *et al.* (2009), compreende um amplo conjunto de técnicas aplicáveis a um vasto domínio de situações de pesquisa. Algumas dessas técnicas incluem a análise de componentes principais, análise de fatores comuns, regressão múltipla, correlação múltipla, análise de relação canônica, análise multivariada de variância e covariância, análise de correspondência e análise de agrupamento.

Esta pesquisa concentra-se na análise de agrupamento, também conhecida como análise de *clusters*. Essa abordagem analítica visa desenvolver subgrupos ao categorizar elementos em grupos menores com base em suas similaridades. Ao contrário de abordagens com grupos predefinidos, na análise de agrupamento, os grupos são identificados por meio da técnica de agrupamento. O processo de agrupamento geralmente envolve pelo menos três etapas: a) avaliação das similaridades ou associações entre as entidades para determinar a quantidade de grupos existentes na amostra, b) com base nessas similaridades, os elementos são particionados em grupos menores e c) estabelecimento do perfil das variáveis na composição dos subgrupos (Hair *et al.*, 2009; Mingoti, 2023). Essa abordagem possibilita otimizar as similaridades entre as variáveis, resultando em grupos mais homogêneos, ao mesmo tempo em que reduz as semelhanças entre grupos distintos (Mingoti, 2023, p. 143).

A análise de agrupamento é uma técnica amplamente empregada em diversas áreas, incluindo biologia, psicologia, economia, sociologia e administração. Recebendo diferentes

[21] Por variável estatística, entende-se: "[...] as medidas tomadas nas unidades experimentais, após terem sido submetidas aos tratamentos, constituem os valores das variáveis dependentes ou variáveis respostas. Estas são, no geral, predeterminadas pelo pesquisador, ou seja, ele sabe o que vai medir, contar, observar dentre outras. Grosso modo, podemos considerar as variáveis como sendo quantitativas contínuas e descontínuas e qualitativas atributivas. Muito frequentemente, são usadas as variáveis derivadas a partir de duas ou mais variáveis, como é o caso das percentagens, dos índices ou das relações" (Assis *et al.* 2019, p. 731).

denominações como análise Q, análise de classificação e taxonomia numérica, essa abordagem possui o objetivo comum de agregar variáveis com base nas relações entre os elementos do grupo e suas distâncias (ou proximidade).

Na área da Linguística poucos estudos têm empregado a análise de agrupamento como uma técnica multivariada. Gómez (2002, p. 234) aponta que a resistência dos linguistas quanto ao uso de dados estatísticos (ou quantitativos) tem sido frequentemente por medo, falta de treinamento ou aversão. Para o autor, a resistência é motivada pela crença de que as técnicas quantitativas não devem ser empregadas às humanidades e, sim, a área das exatas, além, "da sensação que esses métodos podem destruir a "magia" da literatura".

Contudo, mesmo diante dessa resistência, o crescente aumento de pesquisas na área da LC tem levado as análises quantitativas a ganharem notoriedade entre os linguistas. A LC, que coleta e explora o corpus, ou conjunto de dados linguísticos textuais[22], busca empiricamente padrões linguísticos por meio de ferramentas computacionais. Assim, a análise de agrupamento classifica as variáveis definidas pelo pesquisador de modo que cada elemento tenha características semelhantes.

Os agrupamentos formados devem apresentar homogeneidade interna (dentro do grupo) e, ao mesmo tempo, elevada heterogeneidade externa (entre os grupos). Em outras palavras, se a categorização ocorrer corretamente, os elementos pertencentes aos grupos estarão visualmente próximos quando representados graficamente, enquanto conjuntos distintos estarão visualmente distantes.

Nesse contexto, a definição de similaridade desempenha papel fundamental na análise estatística de agrupamento. Conforme observado por Hair *et al.* (2009, p. 440), a similaridade é definida como "uma medida empírica de correspondência, ou semelhança, entre objetos a serem agrupados". Na presente pesquisa, medida de similaridade foi calculada para todos as ELexs encontradas considerando as características especificadas pelo pesquisador[23] utilizando a ferramenta N-gram[24] no software AntConc[25].

---

22 Usamos a expressão "dado linguístico textual" de modo genérico, significando todos os tipos de textos: escrito, oral, visual.

23 As especificações a que fazemos referência constam na seção 3.3, incluem uma frequência mínima de 20 vezes por milhão de palavra e um alcance (dispersão) mínimo de 10 vezes em que a sequência lexical deva aparecer no corpus.

24 N-gram é uma ferramenta digital do software AntCon que examina todo o corpus em busca de agrupamentos (*clusters*) de tamanho 'N' (por exemplo, clusters de 2, 3, 4, 5, n palavras). Isso permite encontrar expressões comuns no corpus.

25 AntConc é um software dedicado a pesquisas em linguística de corpus, aplicando métodos específicos e contribuindo para o aprimoramento da aprendizagem de línguas por meio de dados. A versão utilizada neste estudo é 4.2.0/2022, Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 27 dez. 2022.

A similaridade entre os elementos pode ser medida de diversas formas, e nas análises de agrupamento, três abordagens são predominantes: medidas correlacionais, medidas de associação e medidas de distância, dependendo do objetivo da pesquisa e do tipo de dados estatísticos. Na medida correlacional é avaliada a relação linear entre duas variáveis. Ou seja, essa medida analisa o grau de associação entre as variáveis, auxiliando na compreensão de se a mudança em uma variável está associada à mudança em outra variável. A medida de distância é usada para calcular a proximidade ou distância entre as variáveis, contribuindo para a criação de grupos de *clusters*.

A medida de distância é uma medida de dissimilaridade, em que valores maiores denotam menor similaridade. Assim, a distância é convertida em uma medida de similaridade pelo uso de uma relação inversa. Em outras palavras, quanto maior a distância, menos parecidos são os elementos analisados. No entanto, para facilitar a interpretação e a consistência com o conceito de similaridade, muitas vezes é preferível utilizar uma relação inversa. Assim, a distância é convertida em uma medida de similaridade, onde valores maiores agora indicam uma maior similaridade entre os elementos (Hair, *et al*., 2009, p. 442).

Existem diversos tipos de medidas de distância, e entre os inúmeros exemplos, destacam-se a distância euclidiana, distância euclidiana quadrada, distância de Manhatttan, distância de Chebychev, distância de Mahalanobis. Nesta pesquisa optamos por utilizar a Distância Euclidiana (doravante, DE), também conhecida como distância em linha reta. A DE é uma medida que compreende o comprimento da hipotenusa de um triângulo retângulo formado pelas diferenças nas coordenadas dos pontos no espaço. Essa técnica calcula a menor distância entre dois pontos, proporcionando uma representação da dissimilaridade entre eles (Hair *et al.*, 2009).

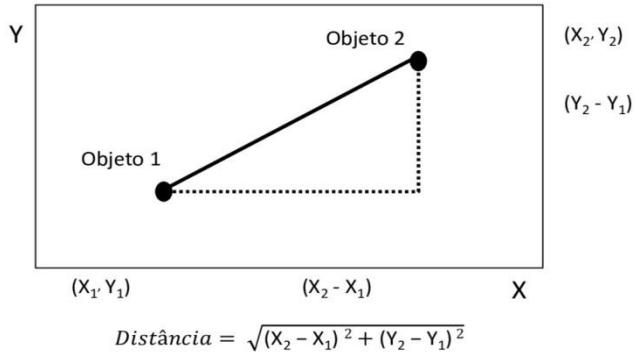


Figura 01 - Distância euclidiana entre dois elementos considerando duas variáveis, *X* e *Y*.

Segundo Hair *et al*. (2009, p. 448), a análise de agrupamento se encarrega de mensurar as características estruturais inerentes a um conjunto de dados. Em virtude disso, a metodologia apresenta propriedades matemáticas, enquanto premissas de normalidade, linearidade e homocedasticidade[26], tão essenciais em outras abordagens estatísticas, demonstram ter pouca influência ou significado neste contexto de análise. A normalidade diz respeito à distribuição normal dos dados, ou seja, assume-se que as variáveis se distribuem de maneira aproximadamente normal. A linearidade refere-se à relação linear entre variáveis e a homocedatiscidade significa que a variabilidade das respostas é constante em todos os níveis das variáveis independentes, assim, não há padrões sistêmicos de variação nas dispersões.

Após o cálculo da distância entre os elementos, torna-se essencial determinar o número de agrupamentos. Existem diversos algoritmos computacionais disponíveis para auxiliar os pesquisadores na minimização das diferenças entre os agrupamentos, considerando as variações internas. Utilizando o software de Linguagem R (R CORE TEAM, 2023)[27], em conjunto com editor RStudio, elaboramos *scripts* na formação dos *clusters* a fim de analisar as estruturas lexicais identificadas no corpus. Essa ferramenta visa proporcionar uma abordagem eficiente para encontrar um número apropriado de agrupamentos, contribuindo para uma análise mais precisa e significativa da estrutura dos dados.

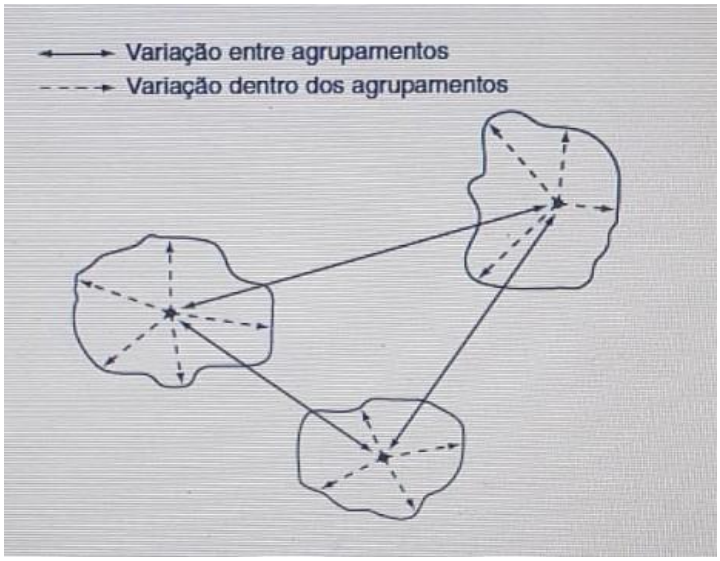


Figura 02 - Diagrama de agrupamento de variáveis internas e externas (Hair *et al*., 2009, p. 444)

[26] Segundo Assis *et al*. (2019, p. 400), homocedático é a "propriedade de um conjunto de dados cuja variância residual de todos os tratamentos é igual".

[27] The R Founndation for Statistical Computing. Vesão 4.3.1/2023. Disponível em: https://www.r-project.org/. Acessado em 20 de abril de 2023.

Os algoritmos mais utilizados na análise de agrupamento podem ser categorizados como sendo hierárquicos e não-hierárquicos (Lattin *et al.,* 2011; Mingoti, 2023). O método hierárquico envolve uma série de n-1 decisões de agrupamentos, e podendo ser aglomerativos ou divisivos. Hair *et al.* (2009) explica que no método hierárquico aglomerativo, cada elemento é inicialmente considerado como um agrupamento individual. Em contrapartida, nos divisivos, todas as observações começam como parte de um único agrupamento e são subsequentemente subdivididas até que se torne um agrupamento isolado (Figura 3).

Nos processos de agrupamentos não-hierárquicos, os números de grupos a serem estudados são pré-estabelecidos pelo pesquisador. Conforme Hair *et al.* (2009) esclarecem, esses agrupamentos também podem ser gerados aleatoriamente por meio de um software de computador, como o FASTCLUS ou SPSS, levando em consideração, por exemplo, a distância mínima entre os elementos. É comum, neste contexto, utilizar a distância nula como padrão.

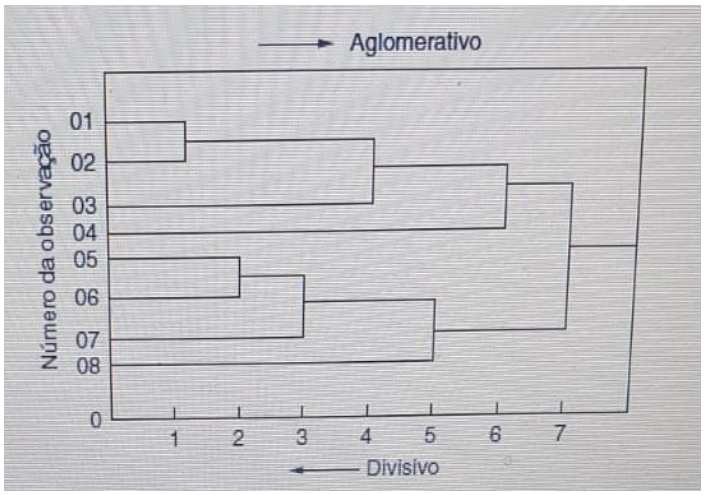


Figura 03 – Dendrograma de grupamento hierárquico (Hair *et al*., 2009, p. 435).

Conforme mencionado anteriormente, os resultados de um estágio anterior são integrados aos resultados de um estágio subsequente, dando origem a um gráfico em formato de árvore ou dendrograma (Figura 13). No entanto, ao organizar o agrupamento, é necessário definir como os elementos serão agrupados. Mingoti (2023) esclarece que os cinco algoritmos aglomerativos mais populares são: ligação individual, ligação completa, ligação média, método centróide e método de Ward. Podemos definir o método de ligação simples como a distância mínima de qualquer elemento de um agrupamento em relação a qualquer elemento do outro agrupamento. Já a ligação completa (método do diâmetro), baseia-se na distância máxima entre os elementos do agrupamento. Hair *et al*. (2019, p. 450) explicam que neste

método “todos os objetos em um agrupamento são conectados uns com os outros a alguma distância máxima. Assim a similaridade interna se iguala ao diâmetro do grupo”.

Ao contrário das ligações simples e completa, a ligação média não depende de valores extremos, como pares mais próximos ou mais afastados. Em vez disso, a similaridade é calculada em consideração a todos os elementos. Assim, ela se situa como um ponto intermediário entre as ligações simples e completas, inclinando-se para pequenas variações internas. No método centróide, realiza-se o cálculo entre dois agrupamentos por meio do centróide[28].

Dentro dessas abordagens, optamos pelo método de ligação de Ward na presente pesquisa (Mingoti, 2013, p. 162). O método de Ward se destaca das técnicas anteriores de agrupamento baseando-se na soma dos quadrados dentro dos agrupamentos, abrangendo todas as variáveis consideradas. Em outras palavras, o método de Ward busca minimizar a variância interna dos agrupamentos, promovendo a formação de *clusters* mais coesos e homogêneos. Assim, combina os agrupamentos com um menor número de elementos, uma vez que a soma de quadrados está diretamente relacionada ao número de elementos envolvidos. Dessa forma, o método de Ward não apenas promove a coesão intragrupo, mas também favorece a formação de *clusters* mais compactos e representativos, resultando em uma análise mais refinada e informativa dos dados.

Assim, em uma perspectiva holística, torna-se evidente a interconexão entre os diferentes elementos explorados nesta pesquisa, cujo propósito fundamental é a análise de unidades fraseológicas descontínuas a partir de um corpus especializado, fundamentada na perspectiva da Teoria Estatística. A metodologia multivariada adotada não apenas possibilita a identificação, mas também a compreensão das relações complexas entre as variáveis do corpus, tornando-as acessíveis por meio da organização em agrupamentos, conhecidos como clusters. Destaco, ainda, que essa abordagem amplia o escopo da pesquisa, preenchendo uma lacuna de pesquisa pouco explorada na área da Linguística Aplicada. A seguir, apresentamos os procedimentos metodológicos adotados na identificação e classificação das estruturas lexicais descontínuas identificadas no corpus especializado.

[28] Em contextos matemáticos e estatísticos, o termo centróide refere-se ao ponto médio em um conjunto de dados. No contexto de agrupamento de dados, o centróide representa o "centro" de um grupo ou *cluster* de pontos. Isso é útil em algoritmos de agrupamento (Assis *et al.*, 2019, p. 99).

CAPÍTULO III

3 METODOLOGIA

No presente capítulo delineamos a metodologia utilizada para a coleta e análise dos dados com o objetivo de responder às perguntas de pesquisa. Inicialmente, (i) apresentamos o corpus, seguido pela (ii) contextualização da pesquisa e descrição das ferramentas usadas na extração das ELexs. Na sequência, a (iii) apresentação dos procedimentos metodológicos utilizadas na identificação das ELexs.

3.1 A descrição do corpus especializado

Para esta pesquisa usamos o Corpus de Artigos de Linguística Aplicada – CorAAL (do inglês, *Corpus of Articles of Applied Linguistics*)) (Prina Dutra e Berber Sardinha, 2021), constituído por artigos acadêmicos organizados na seguinte ordem: resumo, introdução, metodologia, resultado, discussão e conclusão. São 900 textos acadêmicos publicados em seis revistas especializadas de alto impacto na área da Linguística Aplicada (LA), entre 2014 e 2018, com 973.844 palavras (Tabela 02). O CorAAL foi balanceado quanto à abordagem metodológica utilizada na pesquisa de cada artigo: quantitativa, qualitativa ou mista. A Tabela 01 reporta o número de artigos coletados e as respectivas fontes.

Tabela 01 - Fontes e número de textos compilados

Revistas	**Qualitativo**	**Quantitativo**	**Qualitativo e Quantitativo**
Journal of English for Academic Purposes – (EAP)	12	01	13
Language Learning and Technology- (LLT)	11	01	13
TESOL Quarterly – (TQ)	13	05	08

Applied Linguistics – (AL)	02	19	04
International Review of Applied Linguistics in Language Teaching - (IRAL)	-	19	05
English Language Teaching Journal - (ELT)	12	05	07
TOTAL	**50**	**50**	**50**

Fonte: GECEA[29]

Tabela 02 – Extração de dados no AntConc versão 4.2.0 (Anthony, 2022)

Composição	Textos	Nr. de Palavras
CorAAL - QL	300	315.422
CorAAL - QT	300	340.094
CorAAL - QQ	300	318.328
Total	**900**	**973.844**

Fonte: Própria

3.2 Da contextualização da pesquisa e descrição das ferramentas usadas na extração das ELexs.

Ao longo de décadas, a Fraseologia tem estudado sequências lexicais contínuas e, a partir das sequências contínuas (Pacotes Lexicais), têm-se verificado as descontinuidades nas sequências lexicais.

Em 2009, Biber analisou 234 sequências lexicais contínuas em textos acadêmicos, oral e escrito, com a finalidade determinar até que ponto as "palavras" que preenchiam as lacunas eram fixos ou variáveis. O autor percebeu que em escritos acadêmicos os intervalos são variáveis (1*34, 12*4, 1*3*4) e, quando as lacunas são fixas, são preenchidas com palavras funcionais e, quando variáveis, são preenchidas com palavras de conteúdo. Enquanto nos discursos orais as sequências lexicais são contínuas e as lacunas tanto fixas quanto variáveis são preenchidas com palavras funcionais (Gray e Biber, 2013, p. 111).

Römer (2010, p. 98), com o objetivo de "demonstrar o perfil fraseológico de um texto e criar um inventário de itens fraseológicos em um idioma" analisou 280 sequências lexicais

[29] Descrição detalhada da compilação do corpus feita pelos membros do GECEA (Grupo de Estudos de Corpora Especializados e de Aprendizes).

descontínuas usando uma metodologia similar à utilizada por Biber (2009), em que inicialmente identificou sequências lexicais contínuas, para então, determinar as sequências lexicais descontínuas com lacunas variáveis.

Gray e Biber (2013, p. 111) comentam que, neste caso, há uma "pressuposição de que todas as altas frequências descontínuas estão associadas ao menos a uma sequência contínua moderadamente frequente", por exemplo: "*on the * of is*" está associado ao recorrente PL "on the ***basis*** of"[30], identificada através da metodologia direta (do inglês, *direct approach*) na identificação das ELexs.

Diferentemente do exposto, nesta pesquisa, partimos da perspectiva direcionada por corpus (*corpus-driven*) buscando as descontinuidades de sequências formulaicas a partir de observações da variabilidade (*type-token ration* – TTR) e dos diferentes tipos de entropias internas (previsibilidade) nas sequências formulaicas, através da metodologia *frame-to-frame methodology* (Gray e Biber (2013).

Biber (2010, p. 196) explica que, em oposição à perspectiva baseada em corpus, a abordagem direcionada por corpus "difere da prática padrão porque faz suposições mínimas à *priori* sobre as características linguísticas empregadas na análise do corpus"[31].

Desse modo, a partir do corpus especializado, a presente pesquisa caracteriza-se como um estudo indutivo, isto é, parte de um raciocínio particular para o geral. Stubbs (2007, p. 89), explica que em um corpus de grande escala, identificar e interpretar a forma e a função das unidades formulaicas constitui um desafio aos pesquisadores, porém, esse desafio pode ser contornado usando "amostras pequenas para gerar hipóteses que podem ser testadas em larga escala"[32].

Após a constituição do corpus, os dados foram lidos por um software desenvolvido em Python[33] e Qt[34] usando um compilador PyIntaller[35], conhecido como AntConc[36]. A versão

[30] No original: "This approach is based on the assumption that all high-frequency discontinuous frames will be associated with at least one moderately frequent continuous sequence. For example, the extremely high-frequency frame *on the * of is* associated with the frequent lexical bundle *on the basis of."*

[31] No original: "The corpus-driven approach differs from the standard practice of linguistics in that it makes minimal a priori assumptions regarding the linguistic features that should be employed for the corpus analysis. In its most basic form, corpus-driven analysis assumes only the existence of words, while concepts like "phrase" and "clause" have no a priori status".

[32] No original: "The only realistic strategy is to start small: to use a restricted sample to generate hypotheses which can be tested on larger samples"

[33] Python é uma linguagem de programação de alto nível e interpretada, o que significa que o código é executado linha por linha pelo interpretador. Isso facilita a depuração e a execução interativa do código. Além disso, o Python é multiplataforma, funcionando em sistemas operacionais como Windows, macOS e Linux. Criada por Guido van Rossum e lançada pela primeira vez em 1991. É uma linguagem de programação popular devido à sua sintaxe clara e suporte extensivo para bibliotecas. Sua versatilidade amplia seu propósito geral, sendo usada em diversos campos, incluindo desenvolvimento web, ciência de dados, automação de tarefas, desenvolvimento de jogos (Haslwanter, 2016).

4.2.0 (2022) do AntConc é gratuita e capaz de gerar listas automáticas de n-gramas com lacunas múltiplas e variáveis. Os textos do corpus foram codificados em UTF-8[37], padrão usado pelo AntConc para leitura dos dados.

Além disso, a Linguagem R, através de seu editor RStudio, foi usada no desenvolvimento de *scripts* (ANEXO F) na produção de gráficos, pois *software* como Collocate (Barlow, 2004), WordSmith Tools (Scott, 1998) e o AntConc, não possuem, modo geral, a funcionalidade gráfica estatística. O R é uma linguagem de programação voltada para a computação gráfica e estatística de código aberto disponível para Linux, MacOS e Windows totalmente gratuita, possibilitando compilar corpora, processar e fazer anotações de textos, bem como, realizar análises estatísticas e gráficas.

Em pesquisas quantitativas é de fundamental importância uma análise estatística dos dados para a compreensão geral dos dados, desta forma, a Linguagem R corrobora permitindo ao pesquisador fazer afirmações seguras. Para Oushiro (2022, p. 11)

> tabelas e gráficos permitem descrever o que ocorre nos dados, mas isoladamente não permitem explicá-los: a explicação sobre certos padrões deve sempre estar alinhada às teorias e aos modelos linguísticos com que se trabalha. Análises estatísticas, no entanto, podem auxiliar na interpretação de padrões ao revelar correlações entre aspectos concomitantes.

3.3 Apresentação dos procedimentos metodológicos.

Esta pesquisa parte dos pressupostos que (i) a língua é formulaica, ou seja, é padronizada (Wray e Perkins, 2000) e, (ii) em uma lista de frequências gerada automaticamente sem que haja parâmetros previamente determinados e usando um software

[34] O termo Qt refere-se a uma estrutura de desenvolvimento de software multiplataforma utilizada para criar aplicativos e interfaces gráficas. Criada pela empresa The Qt Company, fornece um conjunto de ferramentas e bibliotecas que facilitam o desenvolvimento de aplicativos em várias plataformas, como, Windows, MacOS, Linux, Android e IOS.

[35] PyInstaller é uma ferramenta de código aberto para Python que permite empacotar aplicativos Python em executáveis independentes, que podem ser distribuídos e executados em máquinas que não têm o interpretador Python instalado. Ele suporta os principais sistemas operacionais, incluindo Windows, macOS e Linux (Haslwanter, 2016).

[36] O AntCon versão 4.2.0 de 2022, Disponível em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 27 dez. 2022.

[37] UTF-8 (Unicode Transformation Format - 8bit) é um sistema de codificação de caracteres usado para representar texto em computadores e outros dispositivos digitais. Ele foi projetado para ser compatível com ASCII (American Standard Code for Information Interchange) e é amplamente utilizado na web e em diversas aplicações de software.

especializado, há sequências formulaicas contínuas e descontínuas em um mesmo corpus (Gray e Biber, 2013).

Assim, inicialmente, geramos uma sequência lexical geral contendo *5-palavras* usando uma das ferramentas do AntConc - o N-gram (Anthony, 2022), e encontramos o total de 924.521 unidades formulaicas, conforme Figura 04.

Em seguida, uma segunda lista de frequências formulaicas com *5-palavras* foi gerada automaticamente pelo AntConc considerando uma lacuna variável. A busca por apenas uma lacuna variável se justifica, em consonância com Gray e Biber (2013, p. 112), pela complexidade em se identificar, em uma abordagem direta, múltiplas frequências de diferentes frames preenchendo as lacunas. A tabela 03 identifica o padrão considerado na presente pesquisa (* indica o local de variabilidade).

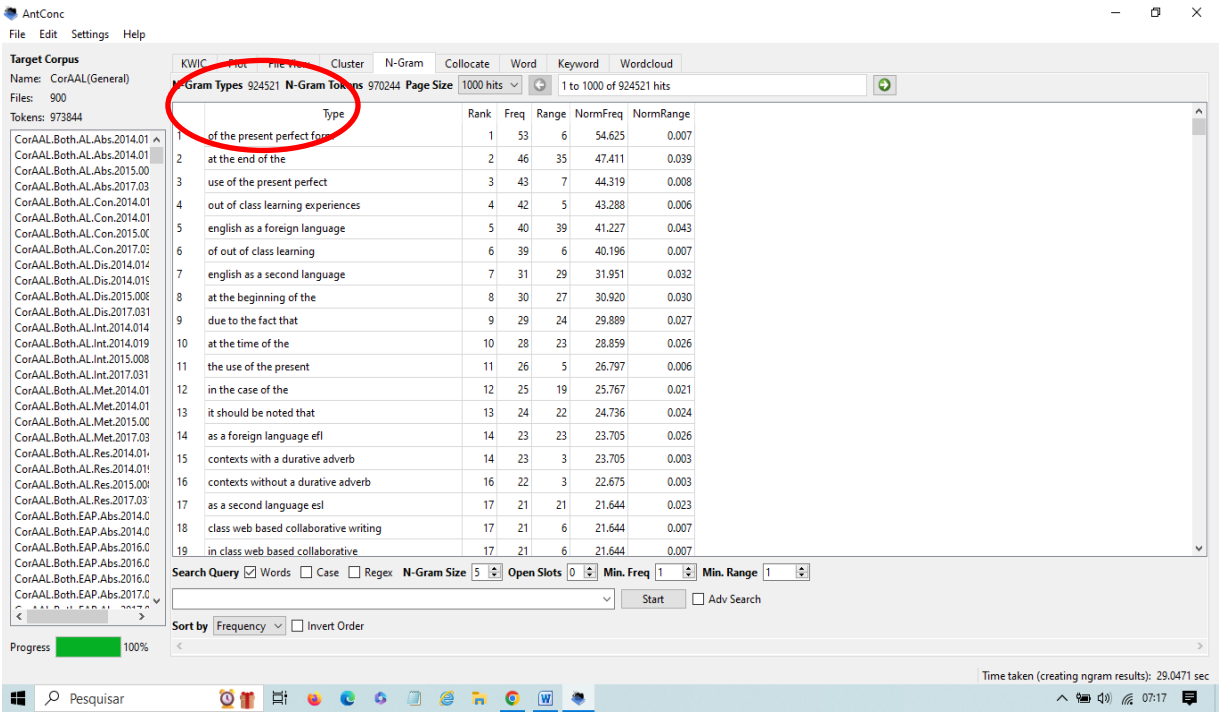


Figura 04 – Inserção do corpus na ferramenta N-gram do software AntConc.

Tabela 03 – O padrão das ELexs nesta pesquisa.

5-palavras	**Exemplos**
1*345	the + of this study
12*45	it is + to note
123*5	due to the + that

Fonte: Própria

Além dos parâmetros, sequência de 5-palavras e uma lacuna variável, outras configurações foram aplicadas à ferramenta N-gram (Anthony, 2022) na extração das sequências lexicais. Por exemplo, após carregar os arquivos do corpus no software, a) a frequência mínima considerada é de 20 vezes por milhão de palavras, isso significa o número de vezes que o n-grama aparece no corpus, b) o mínimo de *range* de 10 vezes, isto é, o alcance mínimo refere-se ao número de arquivos diferentes nos quais o n-grama aparece. Isso garante que os n-gramas selecionados sejam representativos e não apareçam apenas em um único texto, c) sendo ordenada pela frequência, e d) o AntConc deverá mostrar os valores normalizados.

Então, após os referidos parâmetros serem aplicados à ferramenta N-gram (Anthony, 2022), o software identificou 158 sequências formulaicas, como mostramos abaixo na Figura 05. No entanto, como a literatura já previa (Gray e Biber, 2013), houve impossibilidade de distinguir diretamente dentre as sequências formulaicas encontradas, os PLs das ELexs, assim, as sequências lexicais foram analisadas de modo semimanual, isto quer dizer que cada unidade foi verificada separadamente em seu contexto.

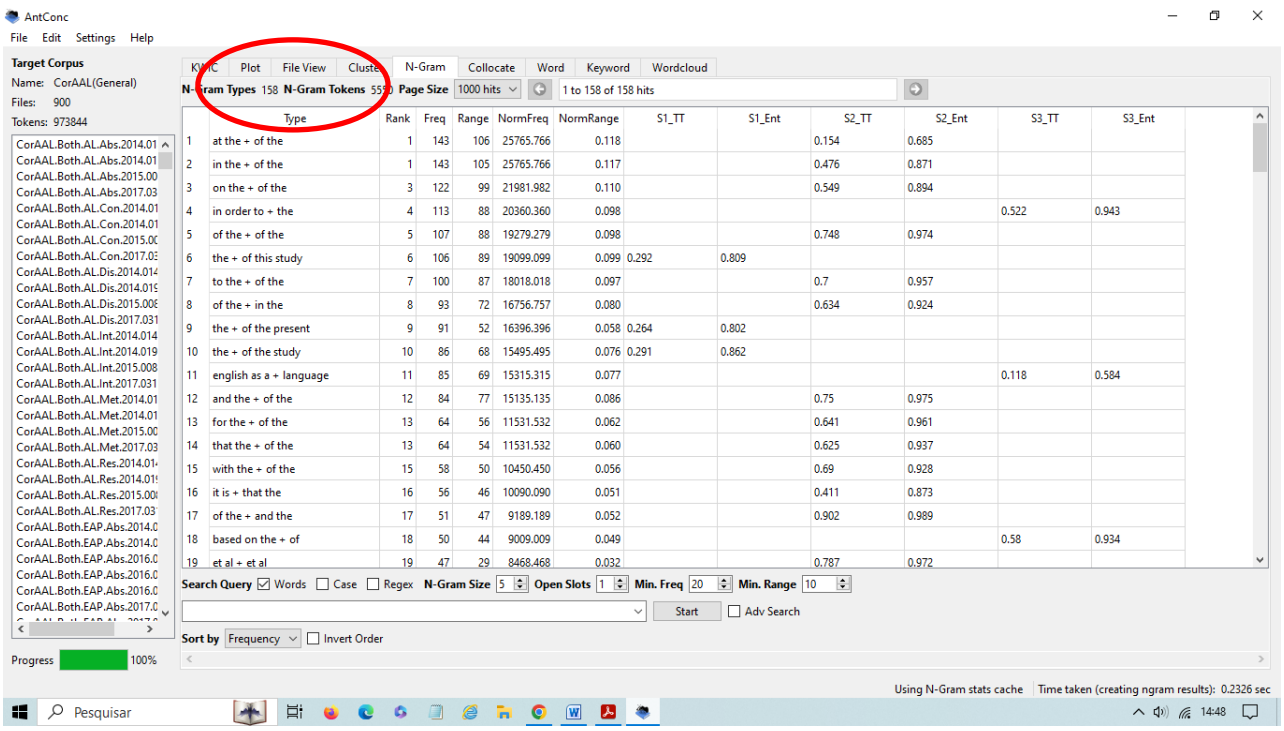


Figura 05 – As 158 sequências formulaicas identificadas pelo N-gram.

Além disso, as unidades não significativas (por exemplo, *et al. + et al.;* e.g. + *et al.*) foram excluídas e, de igual modo, não foram aceitas as variantes com números (p. ex., *is * o'clock in*), nome de pessoas ou de cidade (p. ex., *name is *, live in* *). Tan e Römer (2022, p. 4) asseveram que essas ELexs falseiam as "pontuações gerais de variabilidade e, portanto, não seriam indicadores confiáveis do conhecimento linguístico" em termos de variedade linguística.

Ademais, unidades formulaicas semelhantes com lacunas variáveis foram mantidas. Desse modo, nesta análise, todas as ULs foram consideradas válidas, no entanto, somente as que ocorreram pelo menos em cinco textos do corpus foram aceitos nos testes de previsibilidade e variabilidade.

Tabela 04 – Unidades formulaicas semelhantes

P-frame	**Frame**	**Freq.**
english as a + language	foreign	85
english + a foreign language	as	41
english as + foreign language	a	40
as a + language efl	foreign	23
as a foreign + efl	language	23

Fonte: Própria

Assim, após a análise individual e as exclusões acima citadas, foi possível identificar 66 ULs neste corpus. Porém, ainda não foi possível distinguir os PLs das ELexs, conforme o Tabela 10 em anexo.

De posse desse resultado, buscamos por ELexs não oriundos de PLs. Neste intento, comparamos os PLs identificados por Biber *et al.* (1999), com as ULs identificadas nesta pesquisa. Os autores explicam que os PLs contendo cinco e seis sequências lexicais são significativamente menos comuns quando comparados com quatro sequências lexicais (Biber 1999 *et al*., p. 990) e estes, quando identificados em trabalhos acadêmicos, podem ser agrupados em doze categorias (p. 1014) conforme o Quadro 05.

Quadro 05 – As categorias lexicais identificadas em Biber *et al*. (1999)

As doze categorias mais encontradas	**Exemplos**
Noun phrase with of-phrase fragment;	+ the point of view
Noun phrase with other post–modifier fragments;	the way in which the
Preposition phrase with embedded of-phrase fragment;	* as a result of the as in the case of
Other prepositional phrase (fragment);	and at the same time^
Antecipatory it + verb phrase/adjective phrase;	it is not possible to
Passive verb + prepositional phrase fragment;	is to be found in ^
Copula be + noun phrase/adjective phrase;	is one of the most
(verb phrase +) that-clause fragment;	+ should be noted that the ^
(verb/adjective +) to-clause fragment;	+ is not to say that
Adverbial clause fragment;	-
Pronoun/noun phrase + be (+ ...)	this does not mean that ^
Other expressions	-

Fonte: Biber *et al.* (1999, p. 1014)

Durante a extração das unidades fraseológicas no corpus, identificamos um total de 66 unidades com 5-palavras (Tabela 06). Dessas unidades, 09 foram identificadas tanto na pesquisa conduzida por Biber *et al.* (1999) quanto nesta pesquisa. Além disso, 46 unidades fraseológicas foram exclusivamente identificadas na pesquisa conduzida por Biber *et al.* (1999), caracterizando-se como PLs. Em contrapartida, 11 ULs não estão associadas aos PLs identificados no estudo dirigido por Biber *et al*. (1999). Entretanto, a ausência dessas 11 ULs na pesquisa conduzida por Biber *et al.* (1999) não implica automaticamente que sejam ELexs. Assim, restando-nos a tarefa de examinar em que medida as 11 ULs se distinguem das 66 ULs identificadas inicialmente.

3.4 Dos dados estatísticos – análise de *clusters*

Na análise de dados através de métodos estatísticos multivariados as variáveis são medidas simultaneamente, isso quer dizer que os elementos de uma amostra são agrupados a partir de suas similaridades. Nesta pesquisa usamos a técnica exploratória de sintetização (ou simplificação) da estrutura das variáveis, isto é, análise de agrupamentos (*clusters*). Essa técnica permite maximizar as semelhanças entre as variáveis tornando o grupo homogêneo e, ao mesmo tempo, minimiza as semelhanças entre grupos diferentes (Mingoti, 2023, p. 143).

Com o auxílio do software da Linguagem R (R CORE TEAM, 2023) e usando o editor RStudio[38], desenvolvemos *scripts* para criação *clusters* ou conglomerados (agrupamentos) para analisar as ELexs encontradas no corpus, como pode ser verificado na Figura 03. Segundo Gries (2009, p. 105), a linguagem de programação R é apropriada para personalizar *scripts* com fins estatísticos e gráficos.

Desse modo, as sequências lexicais (Tabela 10, em anexo) foram divididas em grupos de modo que os elementos pertencentes a um mesmo grupo sejam similares entre si considerando suas características internas de variabilidade e de previsibilidade, e os elementos em grupos diferentes sejam heterogêneos em relação às mesmas características.

Tan e Römer (2022, p. 4) explicam que o software AntConc acessa as informações de variabilidade e previsibilidade das ULs. A variabilidade é acessada através da razão *type-token ratio* (TTR), em que os valores de TTR variam de 0 a 1, onde o TTR próximo a 1 indica alta variabilidade, isto é, uma alta proporção de tipos de variantes na lacuna * por UL (seção 2.4).

Enquanto a previsibilidade das ULs é acessada pelo valor de previsibilidade normalizada, ou seja, o valor de previsibilidade indica o nível de incerteza de uma distribuição de probabilidade (Kumar *et al*., 1986). Os tipos de variantes nas lacunas de uma UL variam de 0 a 1, e isso significa que quanto mais próximo de 1 indica uma distribuição mais uniforme com maior probabilidade de ocorrência.

Uma questão central na análise de *clusters* é determinar a técnica de agrupamento das variáveis observadas. Os processos de agrupamentos são: não hierárquicos e hierárquicos, sendo que nos agrupamentos não hierárquicos os números de grupos a serem estudados são pré-estabelecidos pelo pesquisador, enquanto nos agrupamentos hierárquicos não há grupos pré-definidos pelo pesquisador, podendo ser classificados como aglomerativos ou divisivos (Mingoti, 2023, p. 151).

A técnica aglomerativa parte do princípio inicial da existência “n” grupos, sendo “n” o número de elementos do conjunto de dados considerados isoladamente, isso significa que *clusters* específicos são agrupados pelo algoritmo segundo os valores de semelhança. Diferentemente, a técnica divisiva assume inicialmente que todos os elementos fazem parte de um grupo e inicia a divisão dos elementos mais distantes em grupos diferentes.

38 www.rstudio.com

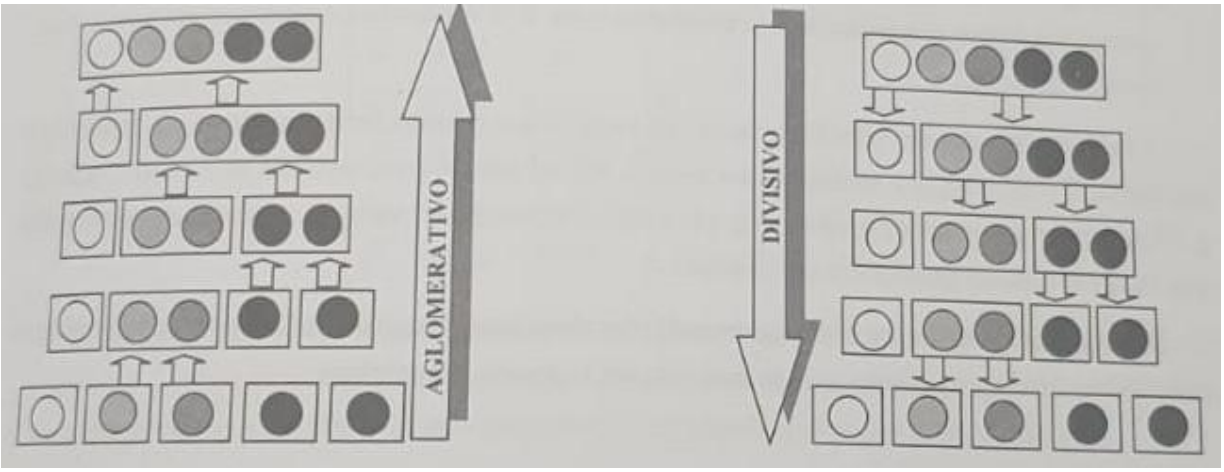


Figura 06 - Procedimentos hierárquicos aglomerativos e divisivos (Mingoti, 2023, p. 151)

Nesta pesquisa usamos a técnica hierárquica aglomerativa, isto significa que para cada elemento observado na Tabela 10 (ANEXO A), as colunas *type-token ratio* (TTR) e a de previsibilidade respectivamente, são considerados como um *cluster* isolado. Assim, no estágio inicial cada elemento amostral é considerado um *cluster* de tamanho 1 e no último estágio do agrupamento será formado um único cluster contendo todos os elementos amostrais (Figura 3).

Segundo Mingoti (2023, p. 152), em

> termos de variabilidade, no estágio inicial, tem-se a partição com a menor dispersão interna possível, já que todos os conglomerados têm um único elemento e, logo, a variância de cada um deles é igual a zero. No estágio final, tem-se a maior dispersão interna possível, já que todos os elementos amostrais estão num único cluster.

Na concepção de Mingoti (2023, p. 144-145), as etapas apresentadas na Figura 03 podem ser sintetizadas da seguinte forma: a) na fase inicial, cada elemento constitui um *cluster,* portanto, "n" *clusters*; b) o algoritmo combina *clusters* similares formando novo conglomerado, isto é, n-1, pois em cada etapa apenas um conglomerado é formado; c) na terceira etapa, ocorre a consolidação dos agrupamentos, ou seja, um novo agrupamento de *clusters* é formado pela incorporação do agrupamento anterior. Isso significa que, quando dois elementos amostrais são combinados em um novo *cluster*, eles permanecerão juntos em agrupamentos subsequentes e não poderão ser separados. Esse processo é conhecido como *propriedade de hierarquia*; d) a partir da propriedade de hierarquia, é possível construir a "árvore" hierárquica do agrupamento, isso significa construir a história do agrupamento.

Assim, o objetivo é agrupar os elementos considerando suas medidas de similaridades ou dissimilaridades, ou seja, para cada elemento amostral j, tem-se o vetor de medidas X_j definido por:

$$X_j = [\ X_{1j}, X_{2j}, ...X_{pj}]' ,\ j = 1,2,...,n$$

onde, X_{1j} representa o valor observado da variável i medida no elemento j. Nesta pesquisa usamos a Distância Euclidiana (DE) por se tratar de uma medida apropriada para variáveis quantitativas. Essa é uma distância de dissimilaridade, isso significa que "quanto menor os valores, mais similares serão os elementos que estão sendo comparados" (Mingoti, 2023, p.145).

Assim, o cálculo da DE, nesta pesquisa, dá-se em função dos dados da Tabela 10, colunas TTR e de previsibilidade respectivamente, isto é, a distância entre dois conjuntos de dados, sendo definida por:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

onde, x refere-se à coluna type/token ratio (TTR), y à entropia interna (Previsibilidade) e o i ao elemento de cada coluna.

A DE permite calcular a distância entre os elementos dos conjuntos de dados (x, y), de forma a identificar as características de dissimilaridade entre eles, permitindo desse modo, o agrupamento hierárquico. Considerando a matriz de distância entre os elementos (Figura 07), as análises foram conduzidas usando técnicas hierárquicas aglomerativas usando a função *dist*, do pacote *stats* do software R. (Figura 06) (Kaufman e Rousseeuw, 1990, p. 44-48).

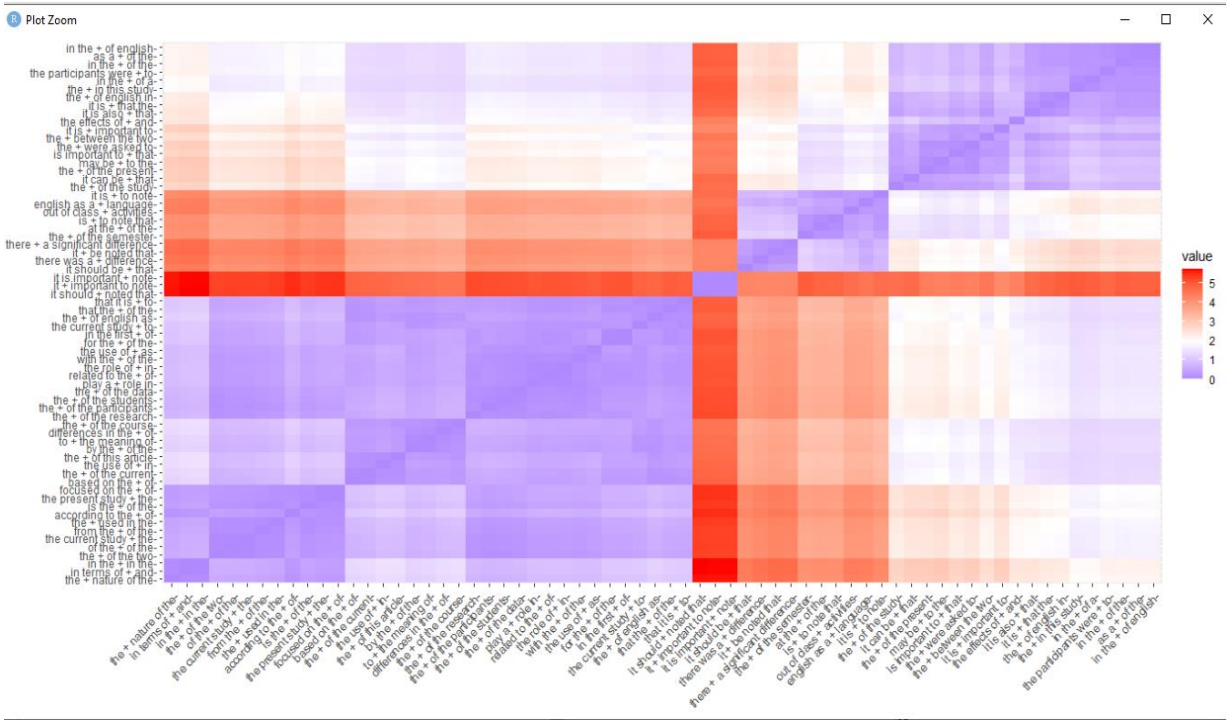


Figura 07 – Matriz de Distância [39]

A escala de valores de 0 a 5, do azul ao vermelho (ordem ascendente), indica o índice de semelhanças entre os pares de elementos (x, y). Desse modo, após a normalização dos dados, a partir de 0 (em azul) indicamos as sequências lexicais que possuem características mais semelhantes, e 5 (em vermelho) assinalamos as sequências lexicais que possuem características mais discrepantes.

No eixo x, o índice de variabilidade (TTR), os valores variam de 0 a 1, onde o TTR próximo a 1 indica alta proporção de tipos de variantes na lacuna; e no eixo y, o índice de previsibilidade, isto é, o nível de incerteza de distribuição de probabilidade, que varia de 0 a 1, em que quanto mais próximo de 1 mais uniforme com maior probabilidade de ocorrência.

Por exemplo, na Figura 08, o ponto de interseção (em vermelho) entre o eixo x e y indica que, quando comparado os índices de variabilidades (TTR) e de previsibilidades entre os elementos: "*it + important to note*" (e suas variantes: *it is + to note*, *it is important + note*) e "*in terms of + and*", não há características semelhantes entre os elementos, ou seja, a distância entre as características dos elementos é elevada.

[39] Em virtude da quantidade de dados e, por consequência a dificuldade de leitura nesta página, adicionamos uma Matriz de Distância Euclidiana em anexo para melhor visualização.

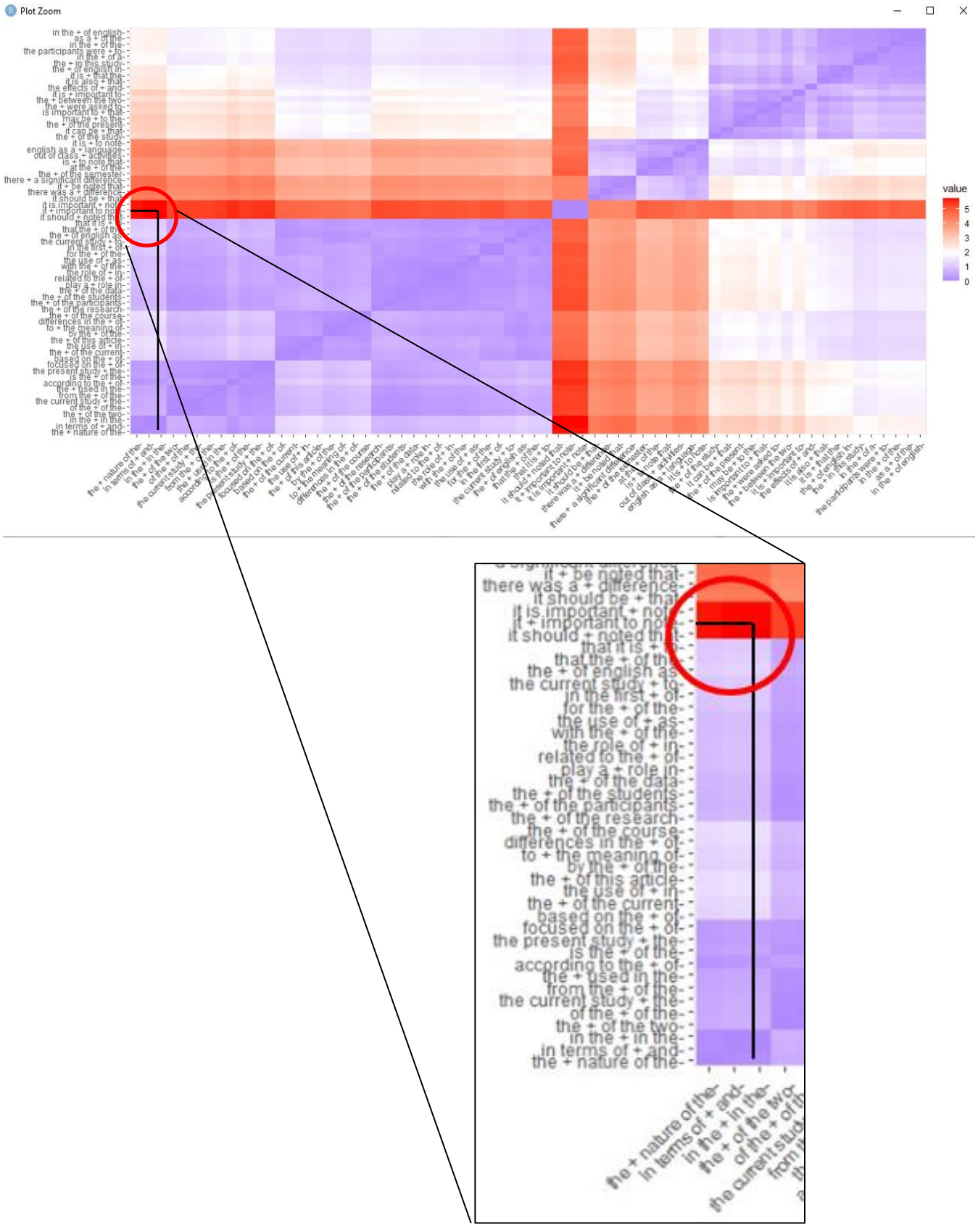


Figura 08 – Distância Euclidiana entre dois elementos discrepantes

O ponto de interseção destacado na Figura 08 evidencia a inexistência de conexão entre as ULs *it + important to note* e *in terms of + and*, apontando para a ausência de proximidade na relação dos índices de variabilidades e previsibilidades. Na UL *it + important to note*, a lacuna é preenchida com verbo, enquanto na UL *in terms of + and*, a lacuna é preenchida predominantemente com substantivos. O Quadro 06 exemplifica a distância entre o índice de TTR e a entropia interna das ULs apresentadas na Figura 08. A análise das disparidades entre esses elementos será abordada na seção 4.

Quadro 06 – ULs com estruturas discrepantes

Unidade Lexical	(x) TTR	(y) Prev	Variante (*)
it + important to note	0.048	0.0	is (21)
in terms of + and	0.897	0.99	number (2); experience (2); equality (2); frequency (1); scores (1); process (1); opening (1); function (1); preparation(1); argument(1); sociocultural(1); content(1); discrete (1); time (1); instruction(1); quantity (1); cognitive (1); déficit (1); cooperation(1); support (1); accurate (1); unity (1); potentiality (1); preparedness (1); speaking (1); phonetics (1); participants(1); conceptualization (1); accuracy (1); lemas (1); comprehending(1); syntactic (1); quality (1); comprehensibility (1); satisfaction (1)

Diferentemente, o ponto de interseção (em azul) entre o eixo x e y indica que, quando comparado os índices de variabilidades (TTR) e de previsibilidade entre os elementos "*that the + of the*" e "*for the + of the*", há características semelhantes entre os elementos, ou seja, a distância entre os elementos não é elevada (Figura 09). Assim, em ambas as ULs, o preenchimento das lacunas é realizado com palavras de conteúdo.

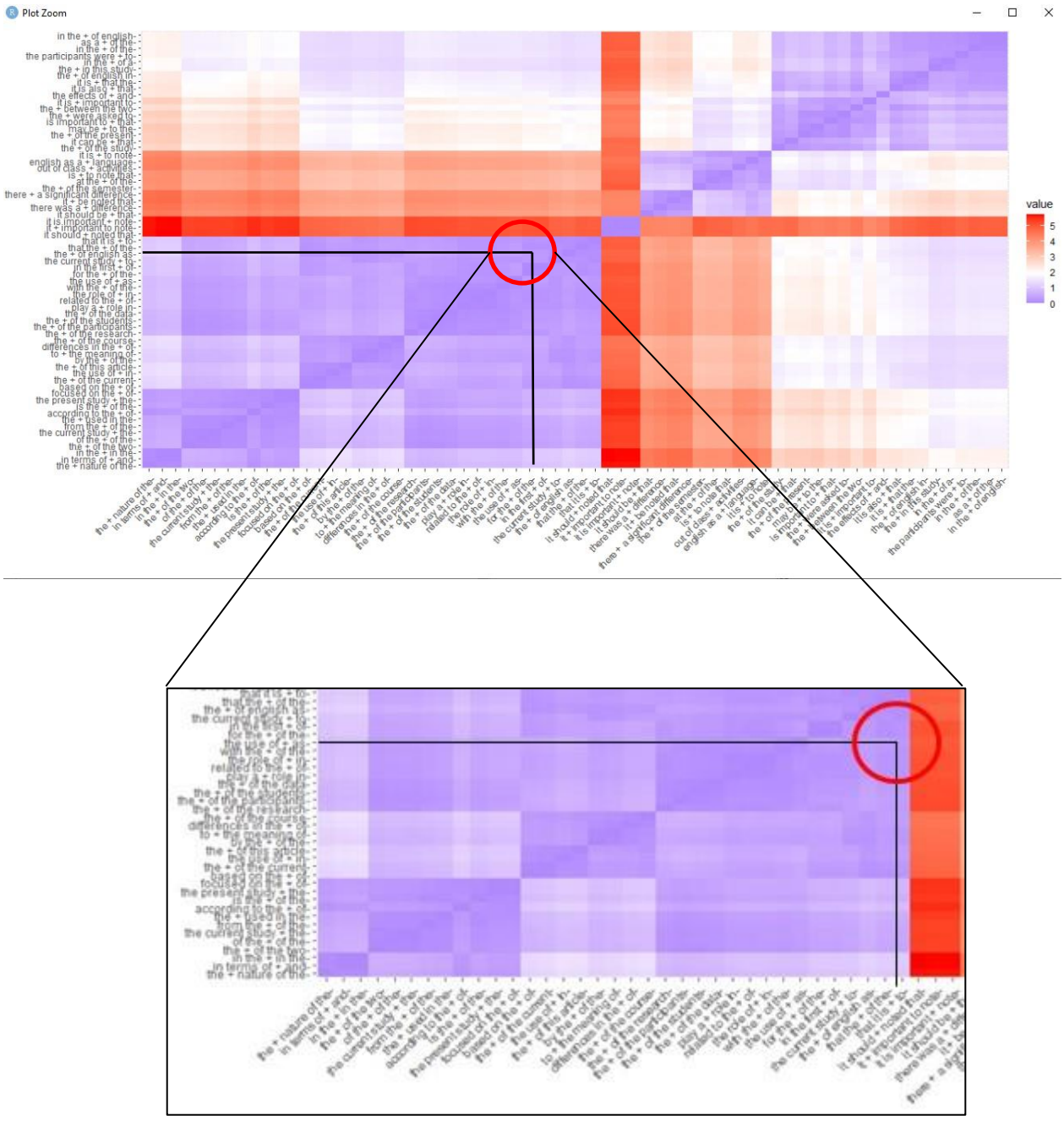


Figura 09 – Distância Euclidiana entre dois elementos semelhantes

O ponto de interseção na Figura 09, em destaque, oferece informações sobre as relações entre os elementos *that the + of the e for the + of the*, evidenciando características semelhantes sugerindo uma proximidade nos índics de variabilidades e previsibilidades. O Quadro 07 exemplifica essa proximidade ao analisar as ULs apresentadas na Figura 09. É relevante notar que uma discussão aprofundada sobre as características entre esses elementos será abordada na seção 4, proporcionando uma compreensão mais abrangente das complexidades adjacentes a essa análise gráfica.

Quadro 07 – ULs com estruturas semelhantes

Unidade Lexical	(x) (TTR)	(y) Previsibilidade	Variante (*)
that the + of the	0.625	0.937	content (6); use (6); majority (5); combination (3); strength (3); performance (3); nature (2); presence (2); authors (2); direction (2); considerations (1); frequency (1); facts (1); quality (1); stability (1); correlations (1); reviewing (1); framing (1); design (1); findings (1); voices (1); judgments (1); weight (1); value (1); effectiveness (1); scope (1); alternatives (1); analysis (1); effects (1); first (1); structure (1); acceptability (1); participants (1); robustness (1); administration (1); acquisition (1); impact (1); purpose (1); teacher(1); results (1).
for the + of the	0.641	0.961	purposes (5); purpose (4); teaching (3); validity (3); use (3); majority(3); rest (2) evaluation (2); development (2); interpretation (2); completion (2); analysis (2); frequency (2); comparison (2); coding (1); infrequency (1); scope (1); investigation (1); success (1); triangulation (1); readers (1); project (1); strengths (1); shape (1); findings (1); features (1); drafting (1); duration (1); coherence (1); end (1); other (1); índices (1); comprehension (1); ability (1); definition (1); each (1); effectiveness (1); absence (1); meanings (1); convenience (1); strength (1).

Assim, após o cálculo da distância entre os elementos, a partir da DE, o agrupamento dos elementos foi realizado usando a função *hclust* do software R e o método de Ward (Ward, 1963)[40]. A função *hclust* permite usar diferentes métodos de ligação entre os elementos para construir um dendrograma e, nesta pesquisa, usamos o método Ward.D2 do software R.

40 Método Ward foi proposto em 1963 por JoE H. Ward Jr., e consiste em classificar n número de parcelas reunindo-as progressivamente em grupos por meio de minimização de uma função objetiva para cada passo de fusão (n-2).

O método Ward, também conhecido como método de “Mínima Variância”, propõe um agrupamento baseado na variação entre os grupos e dentro dos grupos que estão sendo formados, sendo que cada passo do agrupamento fundamenta-se nos seguintes princípios (Mingoti, 2023, p. 162): a) inicialmente, cada elemento é um *cluster*; b) no agrupamento calcula-se a soma dos erros quadráticos (em inglês, *error sum of squares* - ESS) entre os *clusters*, isto é, a soma das distâncias euclidianas ao quadrado de cada elemento amostral entre os clusters, ou seja,

$$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

onde k é o agrupamento em questão, n é o número total de elementos do agrupamento k e x_i é o i-ésimo objeto do agrupamento k. Portanto, em cada passo do agrupamento as distâncias são minimizadas e combinadas.

Desse modo, o método Ward combina os dois agrupamentos considerando a diferença dos tamanhos dos agrupamentos comparados resultando no menor valor de ESS na produção do dendrograma. Além disso, segundo Mingoti (2024, p. 163)

> é importante ressaltar que a aplicação do método Ward não depende do fato de os dados serem provenientes ou não de uma população com distribuição normal multivariada. Para usá-lo, basta que as p-variáveis sejam quantitativas e passíveis, portanto, de cálculo de médias.

O método Ward deve ser aplicado somente para variáveis quantitativas, já que tem por base a comparação de vetores de médias, produzindo agrupamentos com números de elementos semelhantes considerando o princípio de análise de variância entre os elementos.

De igual modo, usamos a função *fviz_dend* do pacote *factoextr*a do software R, para definir, a partir da árvore hierárquica, o dendrograma (Figura 13). O número de *clusters* do dendrograma pode ser definido de modo aleatório, ou segundo a conveniência do pesquisador[41] e, ainda, através do método hierárquico. O número final n de *clusters* define a

[41] Lattin *et al.* (2011, p. 215-252) explica que a escolha aleatória, dá-se, talvez por motivos lógicos ou teóricos ou, ainda, por razões práticas e pessoais do pesquisador em determinar o número de clusters que deseja explorar; quanto à conveniência, o pesquisador opta pelo critério que melhor lhe convém, como: a familiaridade com certo tipo de análise ou por ser de fácil interpretação, como por exemplo, a média de similaridade ou dissimilaridade interna do grupo ou em virtude das distâncias estabelecidas nas etapas de agrupamento.

partição do conjunto dos dados analisados. Nesta pesquisa, optamos pelo método hierárquico, ou seja, a razão entre a variância total interna dos grupos e a variância entre os grupos em relação ao número de agrupamentos e, a função *fviz_cluster* do pacote *factoextra* do software R nos possibilitou a visualização dos agrupamentos.

Segundo Lattin *et al.* (2011), o método hierárquico apresenta a vantagem de ser aplicado tanto a grande, quanto a pequenos volumes de dados, não sendo necessário calcular e armazenar uma nova matriz de dissemelhança em cada passo do algoritmo, além de possuir a capacidade de "reagrupar os objetos em clusters diferentes daqueles que foram calculados inicialmente".

O método de Ward permitiu a visualização dos *clusters* em forma aproximadamente elíptica, como mostrado na Figura 14, que será discutida no próximo capítulo. Esse método promove uma minimização homogênea dentro dos grupos e, por consequência, um distanciamento entre os *clusters*. A configuração final de *clusters* deve-se ao ponto de fusão (distância) agrupando elementos que levam à menor soma de quadrados dentro de cada etapa, produzindo, assim, agrupamentos conexos de tamanhos semelhantes.

Em síntese, os aspectos estatísticos metodológicos acima descritos não constituem a única possibilidade de análise multivariada dos dados, mas creio ter feito opções apropriadas aos objetivos desta pesquisa.

CAPÍTULO IV

4 ANÁLISE E DISCUSSÃO DOS DADOS

Para uma compreensão mais aprofundada da análise proposta, reiteramos o objetivo geral desta pesquisa: identificar, analisar e classificar as unidades formulaicas descontínuas por meio de uma perspectiva direcionada por corpus, com base em dados estatísticos. Dessa forma, esta pesquisa visa identificar os espaços variáveis das ELexs e como essas lacunas são preenchidas, usando uma metodologia direta de extração (*direct-approach methodology*) (Gray e Biber, 2013, p. 121). Temos o propósito de compreender a linguagem em uso em artigos acadêmicos da área da LA, levando em consideração os princípios de variabilidade (*type-token ratio* - TTR) e previsibilidade. Especificamente, neste caso, adotamos uma metodologia estatística multivariada para os agrupamentos (*clusters*) das ELexs.

Este capítulo é subdividido em quatro seções, nas quais apresentamos os resultados e analisamos os dados obtidos à luz de pesquisas e teorias que fundamentam o presente trabalho. Na primeira seção apresentamos os resultados vinculados à primeira questão de pesquisa: quais são as ELexs mais frequentes encontradas no corpus especializado? Na segunda seção, partindo das ELexs identificadas, analisamos os dados referentes à variabilidade (TTR) e previsibilidade das variantes internas das estruturas lexicais, buscando por agrupamentos. Isso responde à segunda questão da pesquisa: quais são as características das ELexs encontradas no corpus especializado? Na terceira seção, exploramos os agrupamentos das ELexs e apresentamos os resultados relacionados à terceira questão de pesquisa: quais aspectos das estruturas lexicais identificadas no corpus podem nortear a criação de tarefas que propiciem aprendizagem orientada por dados (*Data-Driven Learning*)? Por fim, na quarta seção, apresentamos a classificação Funcioanl e a Análise Estrutural das ELexs encontradas no corpus.

4.1 As ELexs mais frequentes encontradas no corpus especializado.

A primeira questão desta pesquisa visa identificar as ELexs existentes no corpus. Durante a extração das unidades fraseológicas no corpus, identificamos um total de 66 unidades com 5-palavras (Tabela 06).

Com base na taxonomia de Biber *et al.* (1999), conforme mencionado no capítulo III, seção 3.3, nove ULs se destacam tanto na pesquisa conduzida por Biber *et al.* (1999) quanto nesta pesquisa. Além disso, quarenta e seis unidades fraseológicas foram identificadas exclusivamente na pesquisa conduzida por Biber *et al.* (1999), sendo classificadas como PLs. No entanto, onze unidades fraseológicas revelaram-se independentes, ou seja, a princípio, não se encontram vinculadas aos PLs. Dessa forma, nesta pesquisa, nos concentamos na análise e classificação dessas ULs para verificar se elas se configuram como PLs ou ELexs.

Assim, ao examinar as construções linguísticas presente no corpus especializado (CorAAL - *Corpus of Articles of Applied Linguistics*), observamos uma predominância de ULs contínuas (PLs) (Biber, 2007; Cortes, 2002) em detrimento às descontínuas (ELexs) (Römer, 2009; Yoon e Casal 2020).

À primeira vista, este resultado corrobora, de certa maneira, com a pesquisa conduzida por Gray e Biber (2013, p. 115), os quais afirmam que os PLs são mais frequentes quando comparados às ELexs em um corpus quando há critérios mínimos de frequência especificada. Com isso, segundo os autores, os PLs conteriam as ELexs em suas unidades, sendo possível a distinção apenas após a verificação das variáveis nas unidades lexicais.

Porém, ao adicionar o parâmetro de variabilidade ao limite de frequência, Gray e Biber (2013) perceberam que houve uma inversão quanto ao padrão de recorrência, isto é, houve aumento das ELexs e, por consequência, diminuição dos PLs. Essa descoberta levou os autores à hipótese de que, nem todas as ELexs estejam vinculadas a sequências contínuas de alta frequência e, também, ao questionamento da eficácia da abordagem que identificava ELexs a partir de PLs. Dessa maneira, perceberam que, mesmo variando os critérios de frequência, o número de ELexs continuava alto quando comparado aos PLs. Isso os levou à revisão da pesquisa conduzida por Biber em 2009, a qual sugeria uma dependência entre ELexs dos PLs.

Resumidamente, a pesquisa conduzida por Gray e Biber (2013) utilizou os critérios de frequência e variabilidade dos preenchimentos das lacunas em ULs como base de análise e, contrariando estudos anteriores, chegaram à conclusão de que há a possibilidade de unidades descontínuas não estarem necessariamente associadas a sequências contínuas.

A perspectiva adotada em estudos fraseológicos sugere que a identificação de ULs está associada a padrões, os quais são possíveis somente a partir de critérios mínimos de

recorrências, como afirmam Römer (2010) e Hunston e Francis (1999). Nesse contexto, a maioria das pesquisas sobre fraseologia, baseando-se em corpora, define critérios mínimos de ocorrência para reconhecer padrões fraseológicos, revelando que cada palavra em um corpus pode demonstrar uma diversidade de padrões, conforme enfatizado por Hunston e Francis (1999).

De igual modo, nesta pesquisa, adotamos o critério de frequência, conforme especificado na seção 3.3, estabelecendo um limite de 20 ocorrências por milhão de palavras, com uma lacuna variável em uma sequência de 5-palavras, com um alcance mínimo de 10 vezes em que a sequência lexical deve aparecer no corpus, resultando na identificação de 66 ULs (Tabela 06).

Na Tabela 10, observa-se que as ULs identificadas com o símbolo (√) são exclusivas da pesquisa conduzida por Biber *et al.* (1999). No entanto, é importante destacar que o objetivo do referido estudo é identificar e classificar PLs. Ainda assim, esse estudo foi fundamental para fornecer parâmetros na identificação das ELexs nesta pesquisa. Os símbolos (*/√) referem-se as nove ULs que aparecem tanto no estudo conduzido por Biber *et al.* (1999) quanto nesta pesquisa (Tabela 05), apresentando características de PLs e de ELxs.

Ao examinarmos as informações da seção 2.4 sobre os parâmentos distintivos entre os pacotes e as estruturas lexicais e, contradizendo as observação feitas por Biber *et al*. (2009) quanto à razão type-token baixa (menos variáveis), que tendem a ser identificadas pelo método dos PLs, e a razão type-token alta (mais variáveis) que não é verificada pela mesma metodologia, observamos que as unidades lexicais da Tabela 05 apresentam baixa variabilidade (0.02 - 0.05) e baixa previsibilidade (0.0 - 0.0). Assim, por exemplo, as unidades lexicais ((*at, in) the + of the* [*end, beginning, time*]), (*english as a + language* [*foreign, second*]) e (*it is + to note* [*important*]), exibem características de ELexs por apresentarem descontinuidade em sua estrutura lexical e flexibilidade quanto ao preenchimento das lacunas com palavras funcionais e de conteúdo. Dessa forma, ao identificarmos as ELexs apenas a partir de unidades lexicais contínuas, acabamos por excluir aquelas que possuem baixa variabilidade como evidenciado pela análise mencionada.

Tabela 05 – ULs identificadas nesta pesquisa e em Biber *et al*. (2009).

Unidades Lexicais	**Variante mais comum**	**Freq**	**TTR**	**Previsibilidade**
at the + of the	end	46	0.02	0.0
at the + of the	beginning	30	0.03	0.0
at the + of the	time	28	0.04	0.0

in the + of the	case	25	0.04	0.0
the + of this study	findings	20	0.05	0.0
the + of this study	results	20	0.05	0.0
english as a + language	foreign	40	0.02	0.0
english as a + language	second	31	0.03	0.0
it is + to note	important	21	0.05	0.0
it should be + that	noted	24	0.04	0.0

Fonte: Própria

Nesta pesquisa, identificamos onze ELexs que não foram identificadas no estudo dirigido por Biber *et al.* (1999). As ULs da Tabela 06, marcadas com (*) na Tabela 10 (em anexo), apresentam em sua estrutura interna, preenchimento dos espaços (1*345, 12*45, 123*5) com palavras de múltiplas funções. Por exemplo, a UL *for the + of the* [*purpose(s), validity, teaching, use, majority*, etc.], é uma ELex que aparece 64 vezes no corpus (Tabela 06), com preenchimento do espaço com palavra de conteúdo de base nominal, não contendo, neste caso, verbos em sua composição interna, sendo precedido de palavra funcional (preposição) (Figura 10).

Observamos que todas as ULs da Tabela 06 são preenchidas com palavras de conteúdo de base verbal, nominal ou adjetival (Tabela 09). Essas observações alinham-se com Gray e Biber (2013, p. 122) quanto às variáveis que preenchem as lacunas como sendo, por exemplo:

- De estrutura verbal – contém um ou mais verbo modais, auxiliares ou principais.

  *...the participants were **asked** to ...(33),*

  *...to **determine** the meaning of ...(*23)
- De estrutura com palavras de conteúdo – contém substantivos, adjetivos, mas nenhum verbo.

  *... the **differences** between the two ...(*30)

  *... the focused on the **use** of ...* (26)

  *... play a **crucial** role in ...* (25)

Além do mais, essa distinção se tornou evidente durante o cálculo da matriz de distância entre as estruturas lexicais (seção 3.4), que serão discutidas mais adiante na análise estatística de agrupamento. Portanto, destacamos que a predominância das ELexs, neste

corpus, é de base nominal, como exemplificado pela Figura 10, em que o AntConc mostra linhas de concordância do frame *for the + of the* com preenchimento variável da lacuna com substantivo.

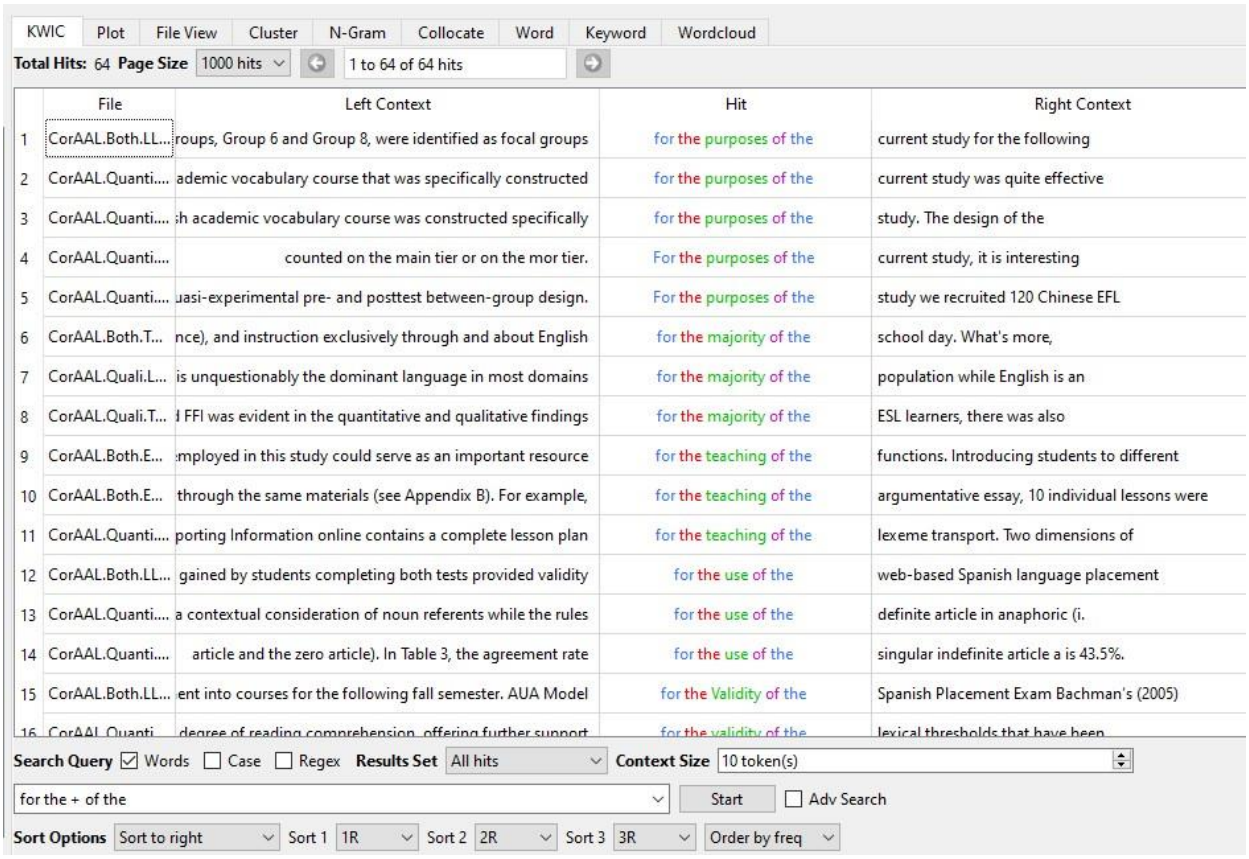


Figura 10 – Linhas de concordância do frame *for the + of the* : preenchimento da lacuna variável com substantivos.

Tabela 06 – ULs identificadas nesta pesquisa

Estrutura Lexical	Variante mais comum	Freq	TTR	Previsibilidade	
for the+of the	purposes	64	0.64	0.96	(2,6%)
based on the+of	results	50	0.58	0.93	(2,0%)
related to the+of	use	37	0.67	0.94	(1,5%)
the participants were+to	asked	33	0.48	0.82	(1,3%)
the+between the two	differences	30	0.37	0.79	(1,2%)

focused on the+of	use	26	0.81	0.97	(1,0%)
play a+role in	crucial	25	0.68	0.94	(1,0%)
according to the+of	level	25	0.84	0.97	(1,0%)
out of class+activities	learning	25	0.08	0.63	(1,0%)
the+of english in	use	24	0.42	0.89	(1,0%)
to+the meaning of	determine	23	0.61	0.86	(0,9%)

Fonte: Própria

4.2 Quais são as características das ELexs encontradas no corpus especializado.

Nesta segunda seção, a partir da abordagem direta empregada no presente estudo, a qual permitiu a identificação das sequências descontínuas, com destaque para as mais recorrentes, passamos à análise das variantes internas das estruturas lexicais, considerando os dados referentes à variabilidade e aos diferentes tipos de entropias internas (previsibilidade), utilizando a técnica estatística multivariada para sintetizar as estruturas lexicais em agrupamento (*clusters*).

Na seção 2.4, a medida *type-token* foi introduzida com o objetivo de capturar o número de diferentes palavras que preenchem o espaço variável em relação ao número total de ELex. Assim, uma proporção próxima a 1 (razão type-token alta) indica uma maior variabilidade, enquanto uma proporção mais próxima de 0 sugere uma variabilidade fixa, ou seja, um conjunto mais limitado de palavra de conteúdo que ocorrem no espaço variável (Gray e Biber, 2013, p. 124). Em 2009 (p. 209), Biber propôs a razão type-token como uma medida para cálculo de variáveis contínuas, dividindo-as em três categorias: ≤ 0,30 (variabilidade relativamente fixa), de 0,30 a 0,70 (variável) e > 0,70 (alto índice de variabilidade).

Tabela 07 – Variação Interna das ELexs

	Unidades - Variantes		
Fixas (≤.30)	02	25	1,0%
Variável (.30 - .70)	08	286	11,5%
Alta variabilidade (>.70)	01	51	2,0%

Fonte: Própria

A Tabela 07 apresenta uma análise da distribuição da variação interna das ELexs que são consideradas relativamente fixas, variáveis e altamente variáveis no corpus especializado

em estudo. De acordo com os resultados, observamos que oito ELexs encontradas no corpus têm preenchimento de ocorrência variável (286 variantes - c. 11,5%), enquanto duas ELexs são de alta variabilidade (51 variantes – c. 2%) e uma ELex é classificada com relativamente fixa (25 variante – c. 1%). O maior índice de variabilidade (≥.30 - ≤.70) das ELexs, são predominantemente baseadas em palavras de conteúdo (b-N, b-A, b-V), conforme Tabela 09 Esses dados sugerem que, neste corpus, há uma distribuição de predominância variável nas ELexs, o que está em concordância com as pesquisas anteriores conduzidas por Biber (2009) e Gray e Biber (2013), que destacaram a considerável variabilidade de palavras de conteúdo na escrita acadêmica.

Adicionamos à nossa pesquisa o índice de previsibilidade. Essa métrica quantifica a incerteza associada à predição do próximo elemento em uma sequência lexical. Assim, 0 e 1 podem ser interpretados como diferentes graus de previsibilidade: quando se aproximam de 1, indica uma distribuição mais uniforme com mais probabilidades de ocorrências e, quando se aproximam do 0, significa que o padrão é aleatório, ou não há padrão discernível na ocorrência (Gray e Biber, 2013).

Portanto, ao integrarmos os dados de variabilidade com os de previsibilidade (conforme discutido na seção 3.4 e apresentado na Tabela 10), as variáveis foram comparadas e agrupadas com base em suas dissimilaridades, resultando no dendrograma da Figura 13. Desse modo, todos os elementos (*clusters*) são agrupados, de forma que os elementos de um grupo sejam semelhantes entre si e, ao mesmo tempo, diferentes dos elementos de outro grupo.

No dendrograma, identificamos inicialmente dois agrupamentos de ELexs identificados pelas cores vermelha e preta (Figura 13). As linhas verticais indicam a distância entre os elementos agrupados, enquanto as horizontais indicam o ponto de fusão. Assim, o agrupamento destacados no dendrograma na cor vermelha, exibem características distintas dos demais. Esses *clusters* demonstram níveis elevados de variabilidade (de .54 a .90) e previsibilidade (de .86 a .99) interna, sendo divididos em subgrupos de acordo com a redução da similaridade dos *clusters* que estão sendo fundidos. Por exemplo, a altura dos nós dos *clusters*: *the + nature of the*, *the term of + and* e *in the + in the*, representa maior similaridade entre si, em comparação com os *clusters*: *the + of the study*, *it can be + that* e *the + of the present*. Portanto, este agrupamento (Figura 11) apresenta características internas de variabilidade e previsibilidade diferentes do agrupamento identificado em preto (Figura 12). Ressaltamos que neste agrupamento estão localizadas a maior parte das ELexs identificadas

através da metodologia direta de identificação (Tabela 06), com exceção das ELexs *out of class + activities*, *the + between the two* e *the + of english in*".

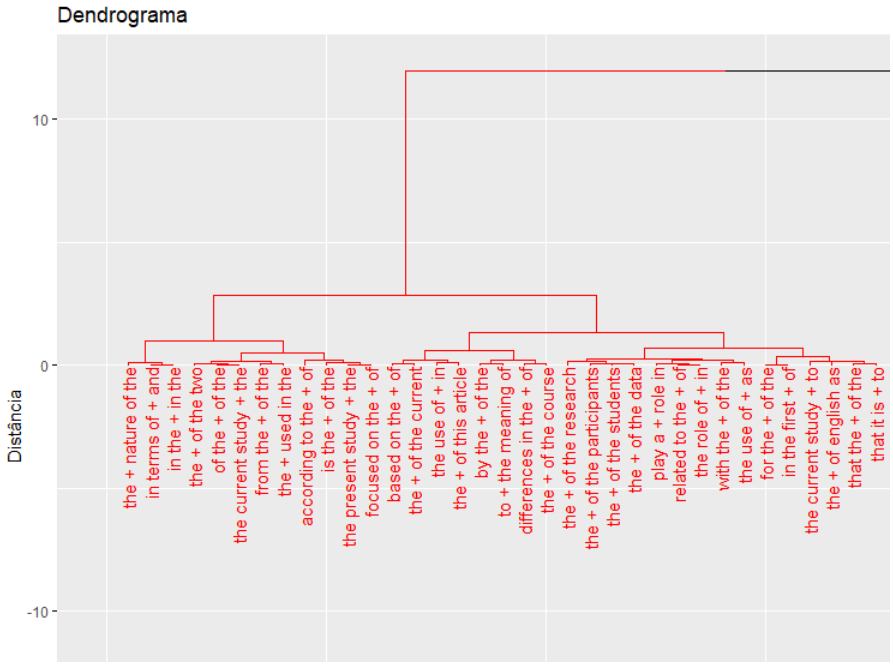


Figura 11 – Primeiro agrupamento: similaridades internas entre as ELexs

O segundo agrupamento, destacado em preto, apresenta subdivisões no dendrograma nas cores em azul, preto e laranja (Figura 12). Isso se alinha aos resultados da Tabela 09, sugerindo que à medida que a variabilidade interna aumenta, decorrente do preenchimento das lacunas das ELexs com palavras de conteúdo e suas dissimilaridades (ou seja, tipos de variantes nas lacunas), as ELexs tendem a se agrupar, destacando-se de outros grupos.

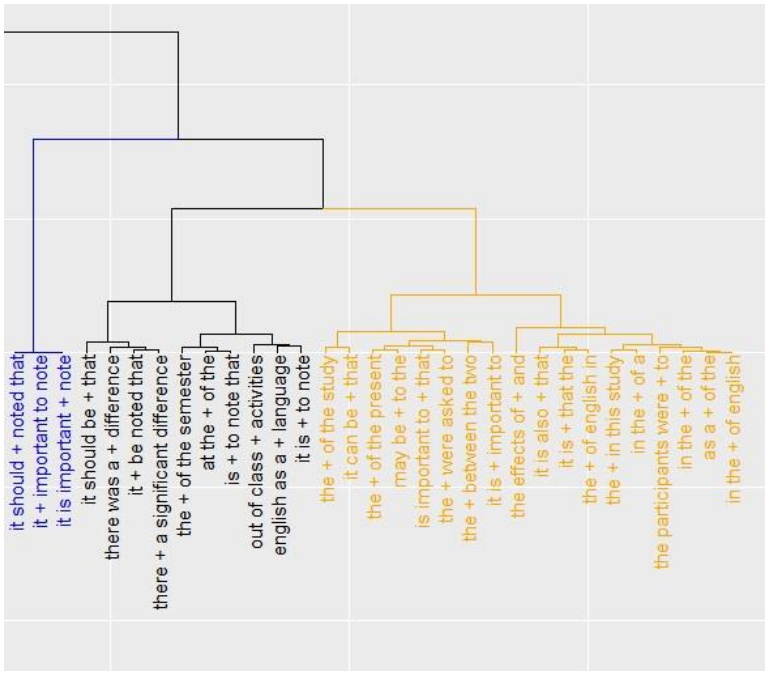


Figura 12 – Segundo agrupamento: baixa similaridade em comparação ao primeiro agrupamento

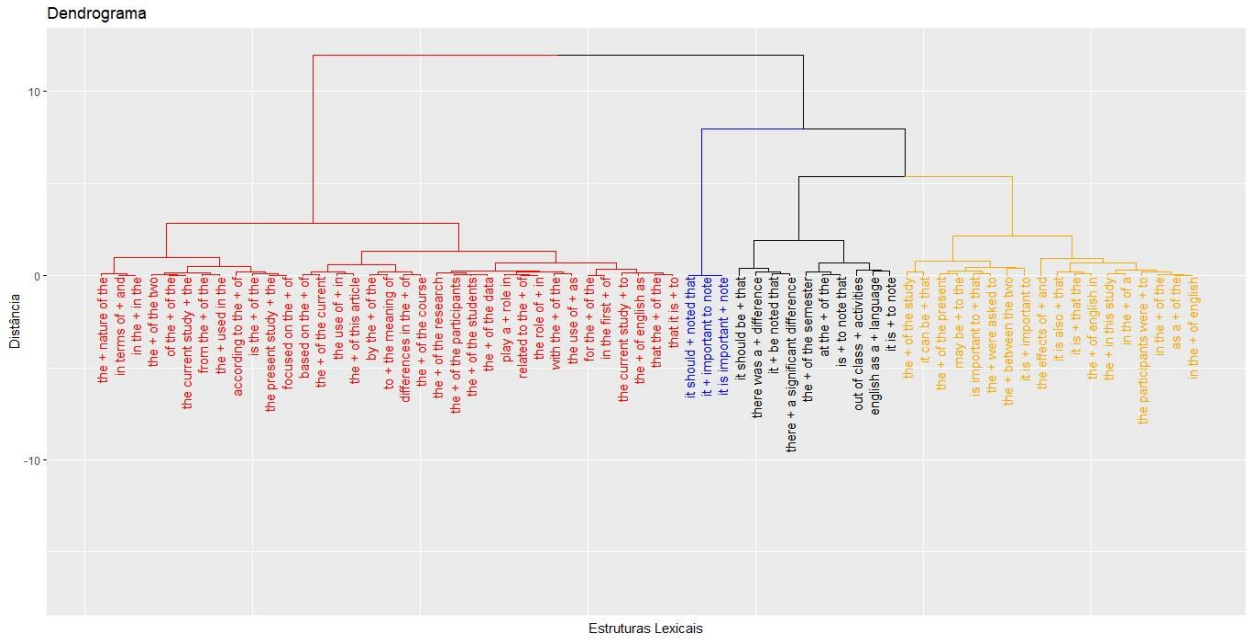


Figura 13 – Dendrograma a partir da Matriz de Distância[42]

[42] Em virtude da quantidade de dados e, por consequência a dificuldade de leitura nesta página, adicionamos um dendrograma em anexo para melhor visualização.

Desse modo, a hipótese discutida no estudo conduzido por Gray e Biber (2013) se confirma, assim como na seção 4.1 desta pesquisa, sugerem que a identificação de ULs está correlacionada com padrões de frequência. Esses padrões, como afirmam Römer (2010), Hunston e Francis (1999), selecionam palavras específicas ao estabelecer os critérios de limites e alcance. A análise estatística dos critérios de variabilidade e a entropia interna (previsibilidade), possibilitou a identificação de ELexs (unidades descontínuas) que não se originam de PLs (unidades contínuas). Além disso, segundo Biber (1993), Berber Sardinha (2000) e na seção 2.1 desta pesquisa, a abordagem estatística possibilita avaliar a distribuição interna e os diferentes contextos das ELexs, bem como avaliar as implicações destas distribuições no corpus.

O método de Ward possibilitou a visualização dos *clusters* em forma aproximadamente elíptica (Figura 14), indicando uma minimização homogênea dentro dos grupos e, por consequência, um distanciamento entre os *clusters*. A configuração final de *clusters* deve-se ao ponto de fusão (distância) agrupando elementos que levam à menor soma de quadrados dentro de cada etapa, produzindo, assim, agrupamentos conexos.

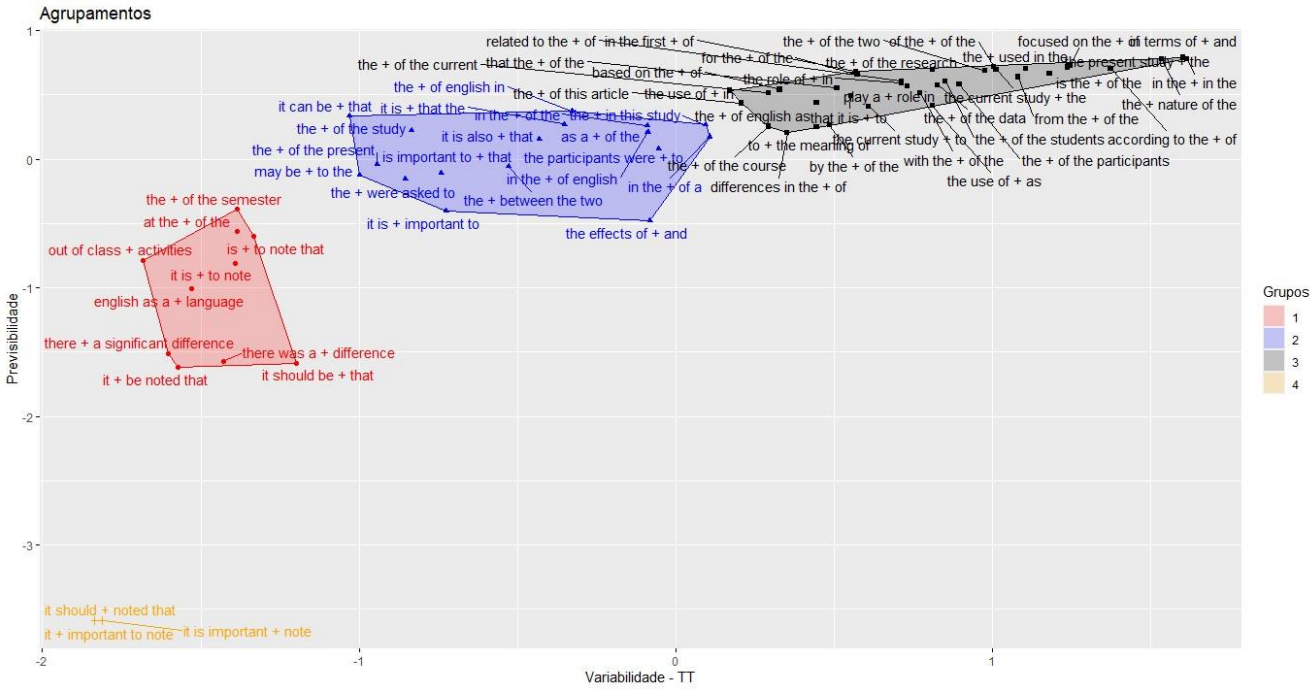


Figura 14 – Agrupamento a partir do Método Hierárquico

Esta abordagem busca agrupar elementos de modo que a soma dos quadrados das distâncias dentro de cada cluster seja minimizada, promovendo a homogeneidade interna dos

grupos. Nesta formação, os agrupamentos são organizados de maneira a mostrar a distância de dissimilaridade entre as ELexs,. Dessa forma, as ELexs com similaridades internas tendem a se agrupar, enquanto as ELexs com dissimilaridades significativas formam agrupamentos distintos. Essa disposição facilita a visualização das relações de similaridade e dissimilaridade entre as Elexs.

4.3 – Da Classificação Funcional e Análise Estrutural das ULs.

A seguir, passaremos a uma análise funcional baseado nos trabalhos de Biber *et al.* (2004), que apontam três funções discursivas dos PLs, a saber: expressões referenciais, expressões de postura e organizadores de discurso. Além disso, Tan e Römer (2022, p. 4) sustentam que, em virtude da variabilidade das ULs, outras funções podem surgir nos textos, resultando na adição de outras categorias ao sistema funcional proposto por Biber *et al.* (2004): as expressões de conversações especiais e expressões de atividades (p. ex., *to * for the* [*apply, pay, compete, look,wait*]).

Em virtude do objetivo desta pesquisa, qual seja o de identificar unidades fraseológicas descontínuas, serão exploradas as seguintes categorias taxonômicas: expressões referenciais (eR), expressões de postura (eP), expressões de discurso (eD) e expressões de atividade (eA), que originalmente foram aplicadas aos PLs. As expressões referenciais são as que identificam ou especificam partes de objetos ou entidades (por exemplo, *that's one of the, there's a lot of*), as expressões de postura transmitem avaliações ou atitudes (por exemplo, *I don't want to, it is import to*), as expressões de discurso são as que fazem conexões entre discursos anteriores e futuros (por exemplo, *if you look at, on the other hand*) e por expressões de atividades são as que referem-se às ações do sujeito oracional (por exemplo, *to apply for the, to pay for the*) (Tan & Römer (2022, p. 4).

Tabela 08 – Classificação Funcional das ELexs.

Estruturas Lexicais	Variantes mais comuns	Freq	Classificação Funcional
for the+of the	purposes	64	eD
based on the+of	results	50	eD
related to the+of	use	37	eP
the participants were+to	asked	33	eP
the+between the two	differences	30	eD
focused on the+of	use	26	eD
play a+role in	crucial	25	eA

according to the+of	level	25	eD
out of class+activities	learning	25	eA
the+of english in	use	24	eA
to+the meaning of	determine	23	eD

Fonte: Prórpia

Além da análise funcional, realizamos uma análise estrutural com o objetivo de determinar a existência de diferenças ou similaridades quanto à composição das Elexs nos diferentes textos. Segundo Hunston e Francis (1999, p. 51-58)[43] a estrutura das ELexs podem ser classificadas em três grupos: de base verbal (bV - *I broke my left leg*), de base nominal (bN – (*a* N; *the* N), *a cinch, a standstill, the blues, the bourgeoisie*) e de base adjetival (bA – (adj+ing) *I felt uncomfortable watching him*).

Tabela 09 – Análise Estrutural das ELexs

Estruturas Lexicais	**Variante mais comum**	**Freq**	**Range**	**Structure**
for the+of the	purposes	64	56	b-N
based on the+of	results	50	44	b-N
related to the+of	use	37	32	b-N
the participants were+to	asked	33	25	b-V
the+between the two	differences	30	24	b-N
focused on the+of	use	26	23	b-N
play a+role in	crucial	25	23	b-A
according to the+of	level	25	23	b-N
out of class+activities	learning	25	11	b-V
the+of english in	use	24	20	b-N
to+the meaning of	determine	23	15	b-V

Fonte: Própria

Destacamos que as análises funcionais foram aplicadas à Tabela 06, englobando as unidades fraseológicas (ELexs) identificadas como não provenientes de pacotes lexicais.

4.4 As ELexs que podem nortear a criação de tarefas que propiciem aprendizagem orientada por dados.

[43] Quadro geral da classificação estrutural baseado em Hunston e Francis *et al*. (1999) encontra-se em anexo.

Esta terceira seção visa explorar os agrupamentos das ELexs extraídas do corpus especializado CorAAL (*Corpus of Articles of Applied Linguistics*) com o propósito de orientar a criação de tarefas que promovam a aprendizagem orientada por dados (DDL - *Data-Driven Learning*). Para essa finalidade, o método estatístico multivariado foi empregado, que permitiu agrupar todos os elementos amostrais, sintetizando as estruturas das variáveis formando *clusters*.

Entendendo que os resultados apontados pela análise do corpus especializado fornecem uma abordagem alternativa para o ensino da linguagem acadêmica, que proporciona ao aluno o desenvolvimento de competências como escritores ao criarem textos que atendam aos padrões estabelecidos por especialistas das áreas nas quais se inserem. Nesta seção, exploramos as ELexs identificadas nesta pesquisa, conforme apresentadas na Tabela 06, e ilustramos sua aplicação na elaboração de tarefas pedagógicas.

Assim, os princípios que norteiam a elaboração desta tarefa pedagógica estão fundamentados no trabalho de Welp, Didio e Finkler (2019), que tem por finalidade desenvolver as habilidades de análise linguística, especificamente da escrita acadêmica em língua inglesa, dos alunos por meio da interação com o corpus especializado. Neste sentido, a sequência didática proposta abordará a linguagem utilizada em artigos acadêmicos na área da LA e está dividida em duas etapas (Etapa 1 e Etapa 2). Na Etapa 1, abordaremos o reconhecimento do software AntConc, incluindo como carregar o corpus no dispositivo, como realizar as buscas, como utilizar as funções da ferramenta N-Gram e, por fim, a produção de um parágrafo com as ELxs identificadas. Na etapa 2, abordaremos a leitura e interpretação dos agrupamentos (*clusters*) obtidos através da análise estatística multivariada.

A tarefa pedagógica apresentada a seguir é um exemplo de como podemos utilizar resultados de pesquisas motivadas por corpus em aplicações pedagógicas. Como a tarefa foi desenhada para ser utilizada por professores e seus alunos do nível intermediário avançado, elas estão totalmente descritas em inglês.

Level (s): Upper intermediate and above

Aims:

To develop students' awareness of formulaic language and of linguistic research and analytical skills through use of and corpus tools. Interacting with specialized corpus, students will develop their research and analytical skills, as well as their ability to think critically and independently.

- Familiarize students with searching for formulaic sequences;
- Introduce specialized corpus: The concept of specialized corpora and its relevance to academic writing. Explain how the corpus can provide insights into language use that may be difficult to obtain through intuition.
- Specialized corpus: To explore the specialized CorAAL (*Corpus of Articles of Applied Linguistics*). Familiarize students with the corpus by exploring its contents and features.
- Explore the tool *N-gram* by AntCon: Show to students how to search specific p-frames, how to filter results by frequency and context, and how to analze the corpus (Table 06).
- Conduct a corpus-driven study: Encourage the students to use the corpus to identify patterns and to interpret the results (Table 06).
- To promot a brief discussion about the limitations of specialized corpus: There are limitations in specilized corpus, such as the need for careful interpretation, and the importance of triangulating data with other sources.
- Encourage critical thinking: To encourage students to think critically about the data and the conclusions. The purpose is that they will gain valuable skills that are applicable to a wide range of academic writing.

Number of class: 2 classes

Class time: 50 min (each)

Resources: Computer, Internet access, Data projector.

Introduction

Formulaic language has been the focus of various studies (Sinclair, 1991; Hunston e Francis, 1999; Wray, 2002; Biber, 2009), especially, due to interest in investigating how language is naturally used in different registers. This lesson plan aims to learn discontinuous lexical units, specifically phrasal frames (P-frames), which are recurring and discontinuous lexical sequences with variable gaps (Fletcher, 2003) (e.g., *it is * to* [*easy, hard, important*]), considering the variability of internal slots through a corpus-driven methodology (Biber, 2012). The specialized corpus of articles in Applied Linguistics (Corpus of Articles of Applied Linguistics - CorAAL - 150 articles, totaling 973,844 from 6 high-impact journals)

was investigated with the use of multivariate statistical methods and we got to the frame structure and their variables through cluster analysis. In academic texts, the gaps in p-frames can be variable, for example, 1*345 (*the*of this study* [*findings, result*]), 12*45 (*it is*to note* [*important, interesting*]), and 123*5 (*due to the*that* [*fact, perspective*]). These gaps are filled with content words, where * indicates the location of variability (Gray e Biber, 2013). This type of analysis considers observations of variability and different types of internal entropies (predictability) in formulaic sequences (frame-to-frame methodology) (Biber, 2012). This proposed lesson explores P-frames from specialized corpus to help students develop their abilities in academic writing.

Steps 1

Before the class

1. The teacher should help students get familiar with the software AntConc interface.

2. Explain how to upload the Corpus of Articles of Applied Linguistics – CorAAL into the software (Figure 15).

3. To explain the parameters that will be used for the task. The parameters are: The tool: N-Gram; Page size: 1000 hits; N-Gram-size: 5; Open Slots: 1; Min. Freq. 20; Min. Range: 10 (Figure 16).

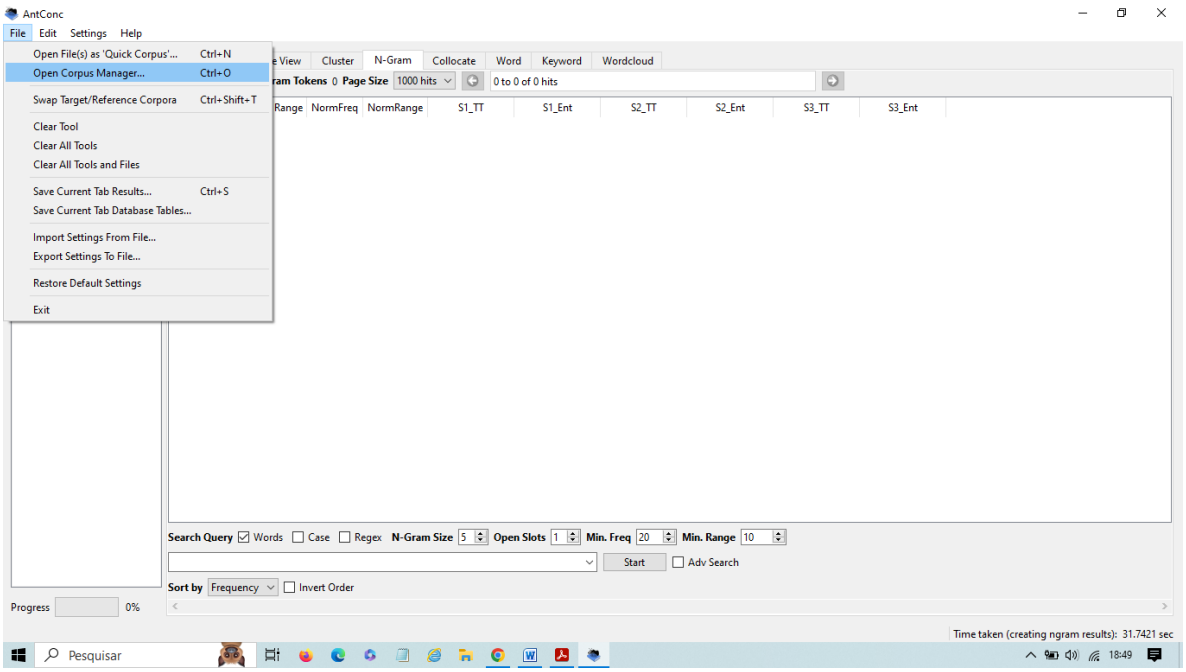


Figure 15 – Upload the file AntConc

In the class

1. Divide the class into two groups (A and B) or pair them up. Write on the board two lists of P-frames. Initially, give the students opportunity to think about which word can complet in each frame.

Group A	***Group B***
for the_____of the	play a_____role in
based on the_____of	according to the_____of
related to the_____of	out of class_____activities
the participants were_____to	the_____of english in
the_____between the two	to_____the meaning of
focused on the_____of	

2. After the exercise, explain the concept of P-frame and their origin. The teacher then checks whether the words suggested by students are among those identified by the research as common P-frames. The teacher can use the Table 06 as a reference to guide students' learning.

Tabela 06 – ULs identificadas nesta pesquisa

Estrutura Lexical	**Variante mais comum**	**Freq**	**TTR**	**Previsibilidade**	
for the+of the	purposes	64	0.64	0.96	(2,6%)
based on the+of	results	50	0.58	0.93	(2,0%)
related to the+of	use	37	0.67	0.94	(1,5%)
the participants were+to	asked	33	0.48	0.82	(1,3%)
the+between the two	differences	30	0.37	0.79	(1,2%)
focused on the+of	use	26	0.81	0.97	(1,0%)
play a+role in	crucial	25	0.68	0.94	(1,0%)
according to the+of	level	25	0.84	0.97	(1,0%)
out of class+activities	learning	25	0.08	0.63	(1,0%)
the+of english in	use	24	0.42	0.89	(1,0%)
to+the meaning of	determine	23	0.61	0.86	(0,9%)

Fonte: Própria

3. Invite students to access the software AntConc (Anthony, 2022) at https://www.laurenceanthony.net/software/antconc/. Instruct students that among all the tools available in the software, the N-gram tool will be used in this task. Explain to them the importance of the parameters.

4. Instruct students to select the N-Gram tool and add all required parameters. Then, the teacher should demonstrate the task by searching for a P-frame (e.g., *for the + of the*) as shown in Figure 17, and explain the results that were found. "The Open Slot Viewer" screen shows all words can be used to complete the gaps in a formulaic sequence (Figure 18).

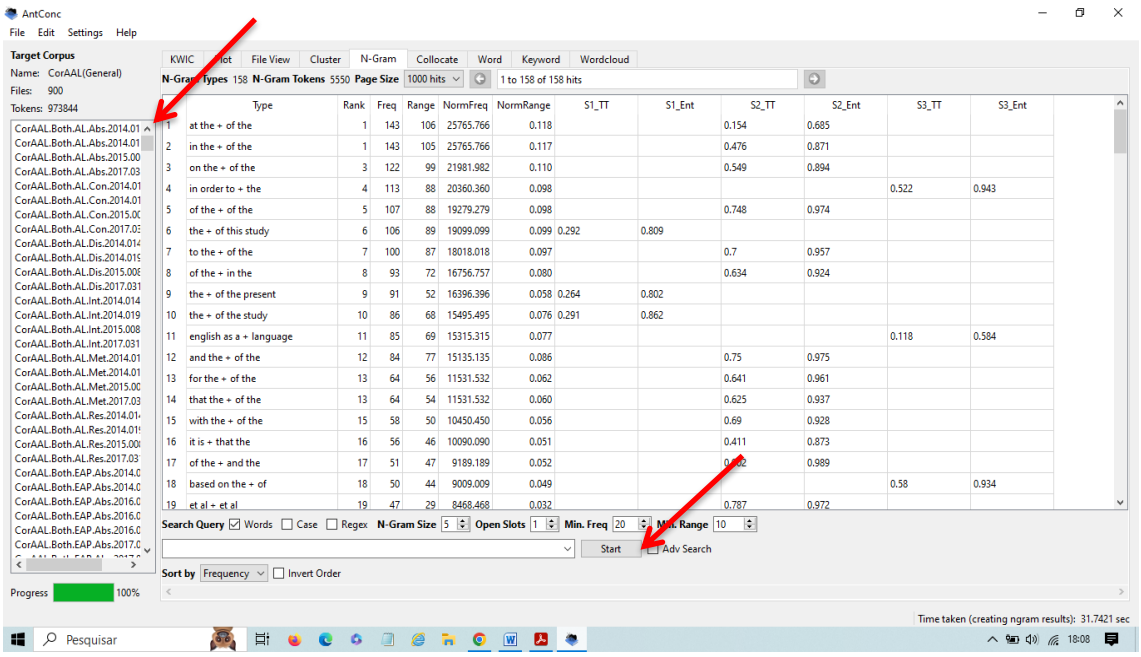


Figure 16 - Insert the search parameters in the N-gram tool.

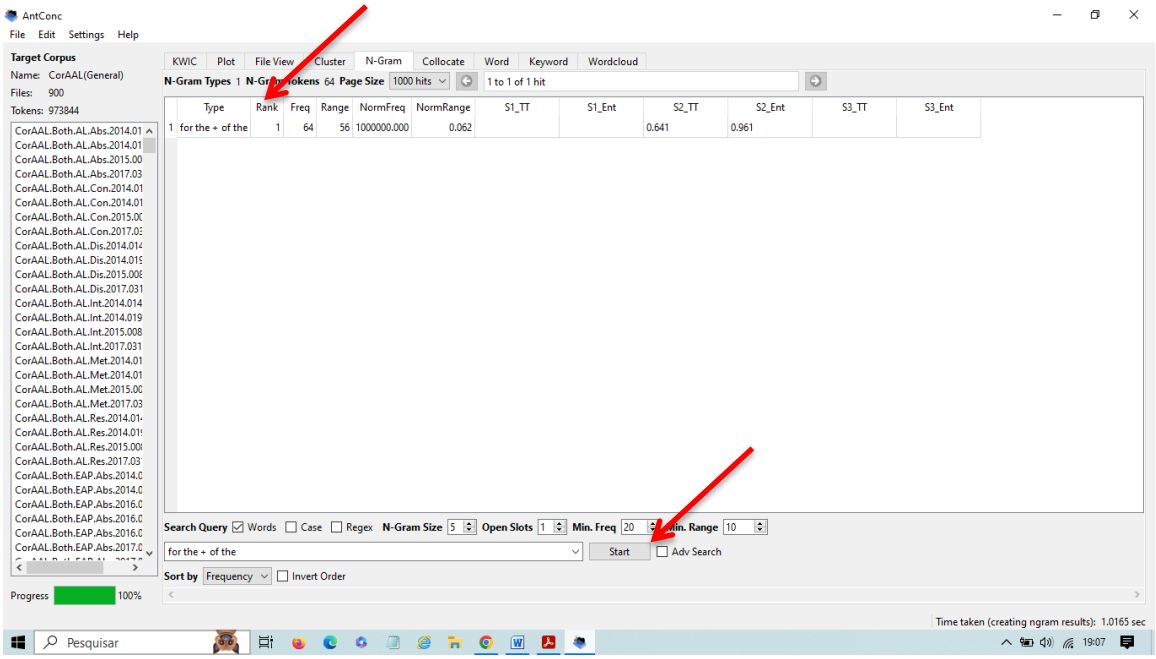


Figure 17 – Showing p-frame section in AntCon

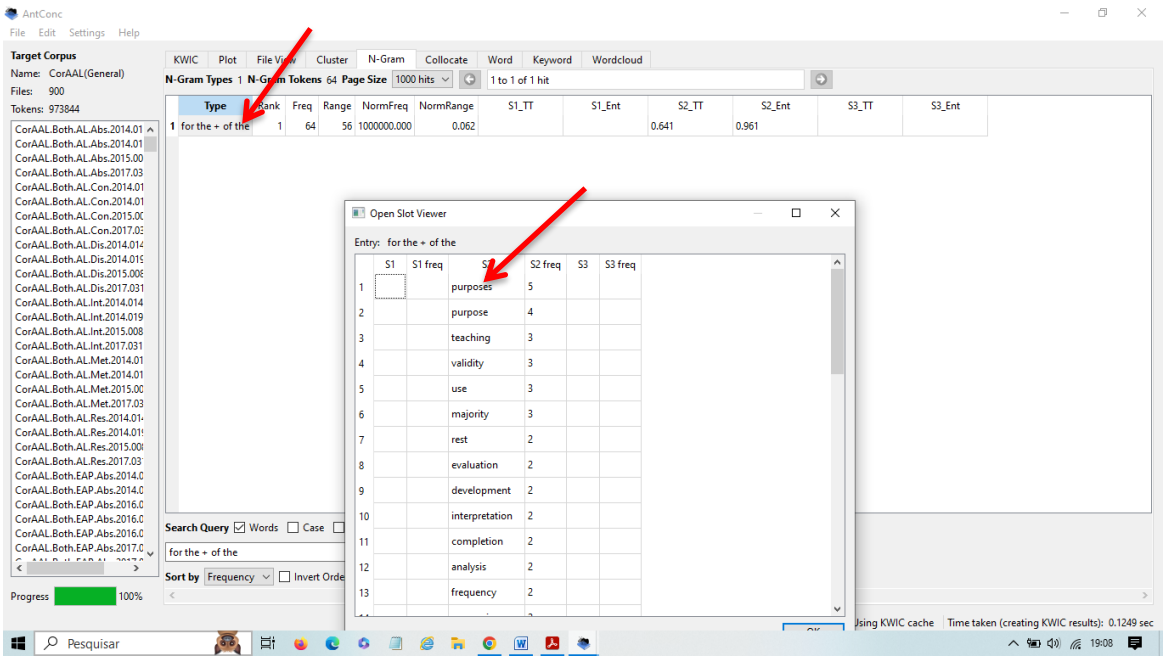


Figure 18 – The open slot viewer, - how to select a P-frame

5. Have students examine the results of their search (Figure.19). The output displays examples of the formulaic unit search where the different parts of speech have been colored (e.g., the main word (verb, noun, adjective, etc. are in green)

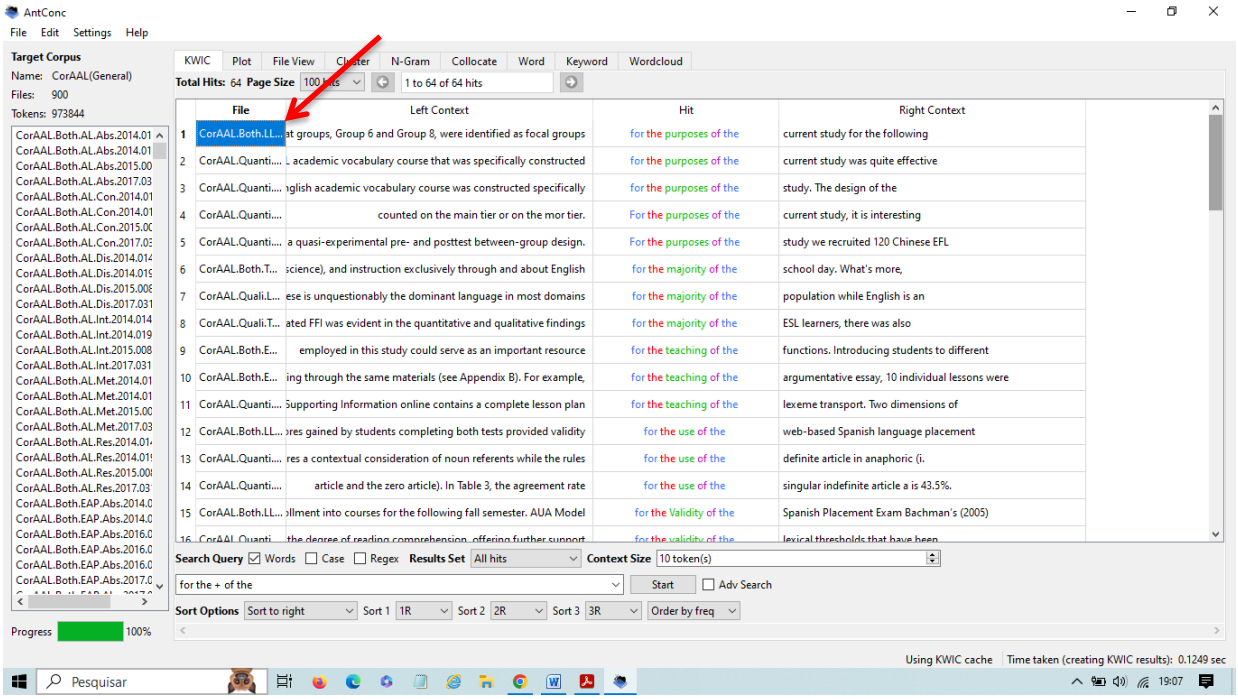


Figure 19 - Selecting a specific file to acess the texts in which the p-frame appears

6. Scrolling down the webpage allows to view all concordances lines. At first, the entire sentence may not be visible in the concordance line, but the students can access more context by clicking on file column (Figure 19, e.g., *for the + of the* = 64 hits). Encourage students to analyse the concordance line and create their own examples using the same context. The teacher explains, for example, that the P-frame "*for the + of the*" can be filled with various words content (Figure 18), Open Slot Viewer : *purposes* (5 hits), *purpose* (4 hits), *teaching* (3 *hits*), etc.), and so on. In this specialized corpus, it can also see in which section of the academic article the P-frame exemplified was used (Figure 20):

CorAAL.Both.LLT.**Met**.2018.005.txt (Methodology section)
CorAAL.Quanti.TQ.**Dis**.2015.005.txt (Discussion section)
CorAAL.Quanti.TQ.**Met**.2015.005.txt (Methodology section)

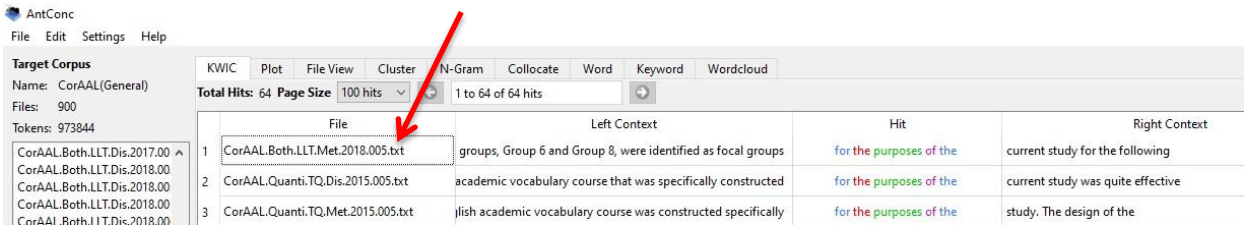


Figure 20 – Selecting a specfic file

7 . The teacher should encourage students to analyse the P-frame in the context in which it was used. When the File was selected (Figure 21) the "File View" tool is opened and they can see original text and the location of the P-frame.

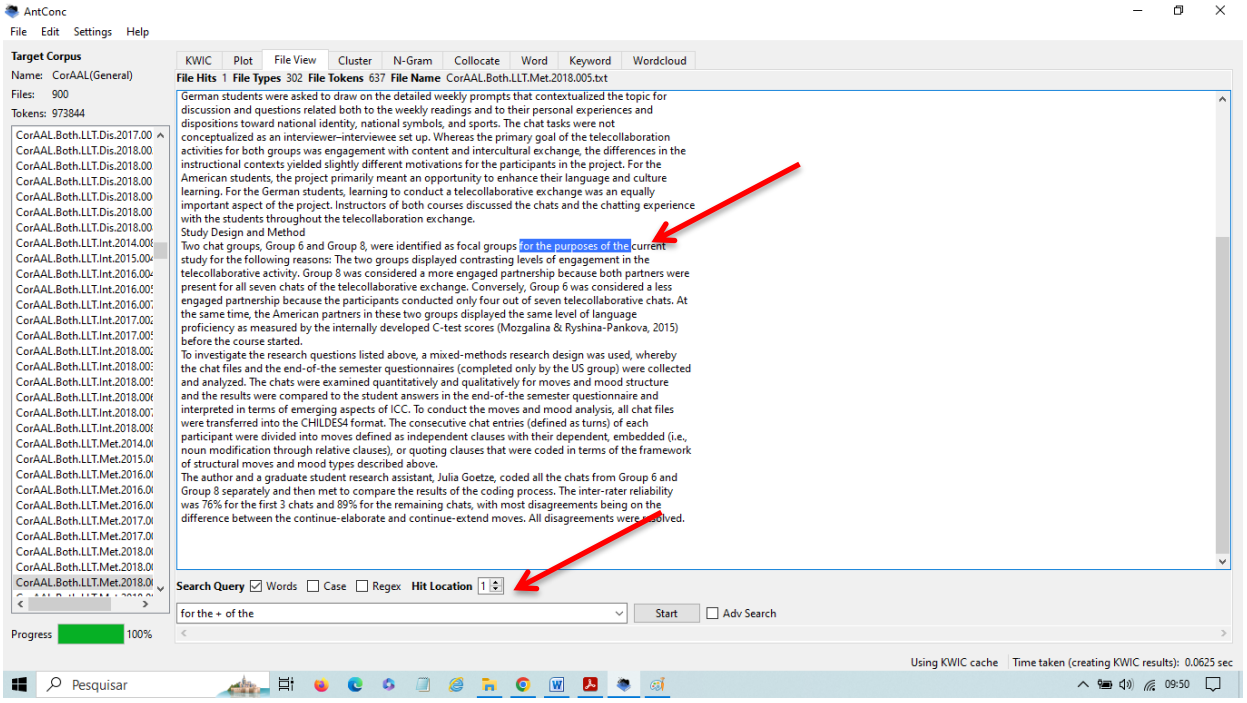


Figure 21 – Reading the file

8. Supervise students and provide feedback as/if needed.

9. Encourage students to make a note of the sentences they create and/or select for the gap-fill activity. Encourage them to revisit Activity 1 and complete the gaps.

10. Virify the grammatical accuracy of these sentences.

11. Pair up students or put them in group to review the P-frames and to present them to classroom. The students should give the definition considering the most common academic article section: Abstract, Introduction, Methodology, Result, Discussion and Conclusion.

12. Ensure the class uses the words in the same context within which they were learned.

13. Each student is required to compose a short paragraph utilizing at least one or two P-frame, focusing on one of sections of an academic article. Subsequently, the teacher should collect all the work and share it all with the students.

Step 2

Aims:

- To analyse the statistical results: it is important to approach statistical analysis with a critical and informed perspective, and to consider the broader context in which the data was collected and analyzed from Figure 8.

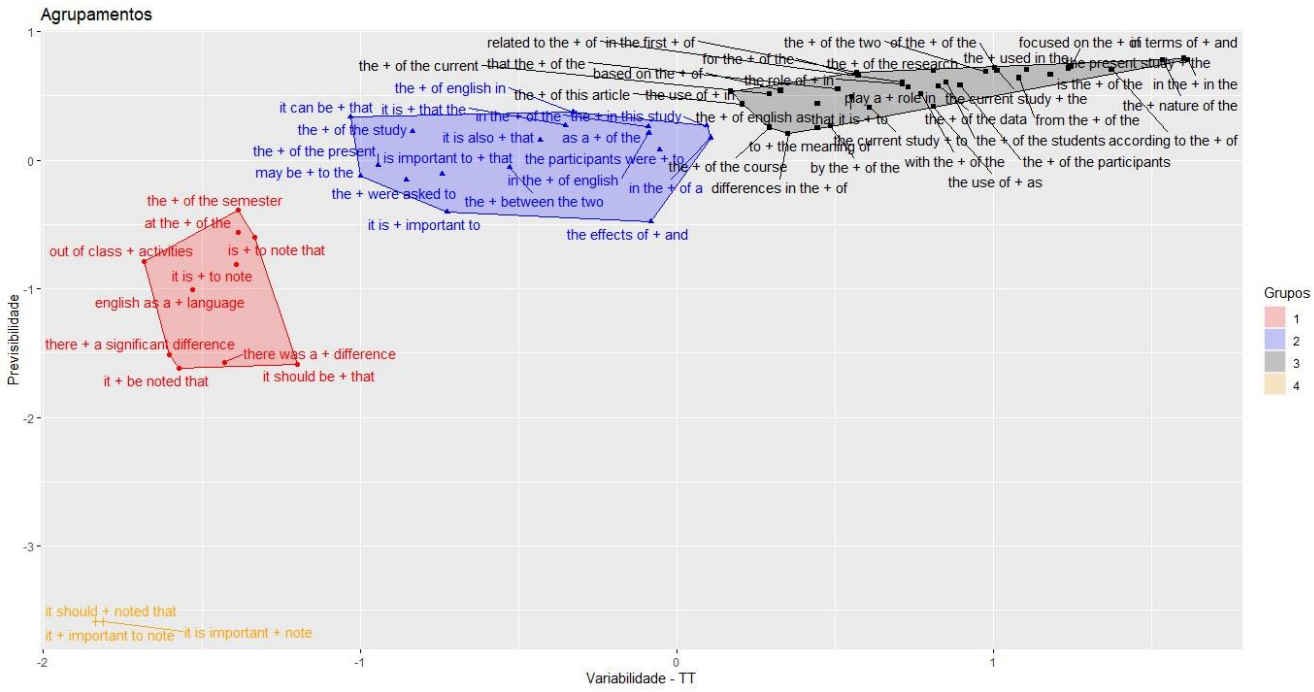


Figure 14 – Hierarchical cluster method.

1. Initially, encourage students to analyze the image of the cluster grouping (Table 8). The teacher writes the students' observations on the board (e.g., amount of grouping, colors, distance between p-frames, distance between groupings, and so on.).

2. Give students' the statistical Reading. The teacher explains to the students that this type of elliptic grouping allows for visualization of homogeneous minimization within the groups and, consequently, the establishment of distance between the clusters. The final configuration

of clusters by melting point (distance) grouping elements that lead to the smallest sum of squares within each step.

3. Encourage students to think critically about the implications of establishing distances between clusters and how it affects data analysis (Figure 8).

4. Divide the class into two groups (A and B) or pair them up. Give them a sheet of paper with P-frames from specialized corpus (CorAAL - Corpus of Articles of Applied Linguistics). The P-frames were identified using the N-gram tool compiled from AntConc. In groups or pairs, complete the chart below according to the example given. Then, each group (or pair) should share their results with all members in the classroom. The function is listed in the table below:

Table 08 – Functional Classification.

Estruturas Lexicais	**Variantes mais comuns**	**Freq**	**Classificação Funcional**
for the+of the	purposes	64	eD
based on the+of	results	50	eD
related to the+of	use	37	eP
the participants were+to	asked	33	eP
the+between the two	differences	30	eD
focused on the+of	use	26	eD
play a+role in	crucial	25	eA
according to the+of	level	25	eD
out of class+activities	learning	25	eA
the+of english in	use	24	eA
to+the meaning of	determine	23	eD

GROUP A (or in pairs)

P-frames	**Functional Classification (Table 8)**	**Section location (Abs, Int, Meth, and so on)**	**Other possible variants**	**Gouping according to Figure 8 (1, 2, 3, 4)**
...for the **teaching** of the...				

...based on the **scores** of...				
related to the **nature** of				
...the participants were **encouraged** to...				
...the **relationship** between the two...				
...focused on the **transformation** of...				

NOW, IT IS YOUR TURN

Utilize P-frames and explore other potential fillers to compose a text. Ensure to take into account the appropriate context or section where the P-frame is typically employed.

_____

_____

_____

_____

_____

_____

_____

_____

_____

5. After this activity, the teacher should help students understand that the statistical multivariate analysis reveals the proximities between P-frames, indicating their internal similarities. This helps underscore the importance of learning and applying in academic writing. The teacher should exemplify it by explaining the dendrograma below:

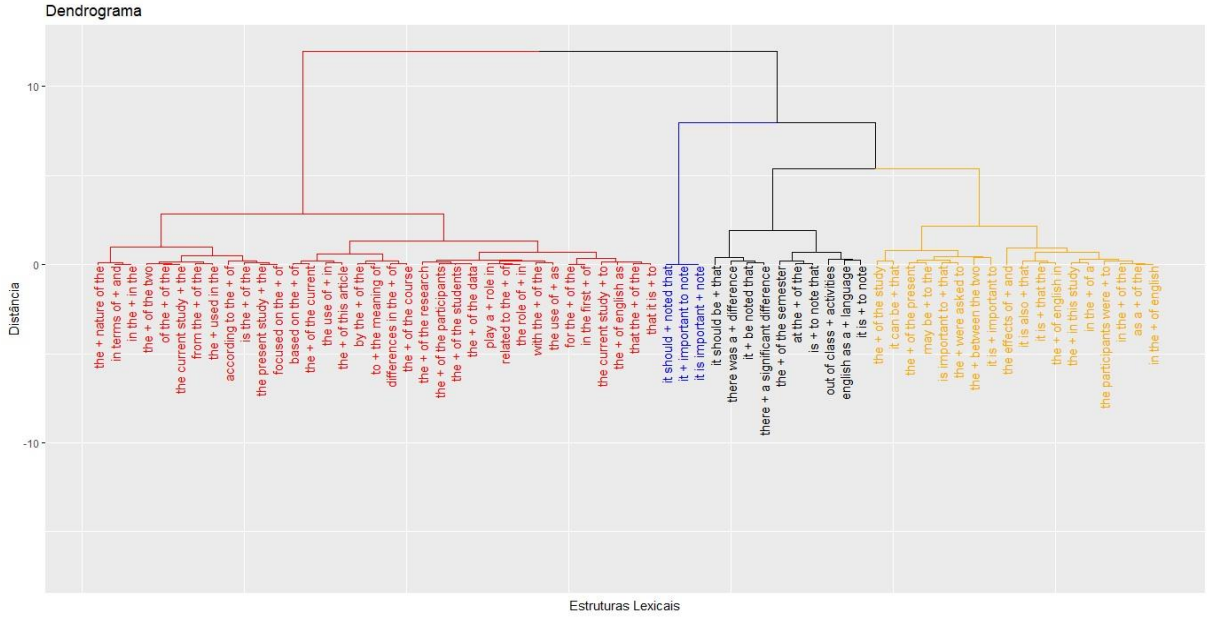


Figura 13 – Dendrograma a partir da Matriz de Distância

GROUP B (or in pairs)

P-frames	**Functional Classification (Table 8)**	**Section location (Abs, Int, Meth, and so on)**	**Other possible filling**	**Grouping according to Figure 8 (1, 2, 3, 4)**
...play a **major** role in...				
...according to the **results** of...				
...out of class **communication** activities...				
...the participants were **encouraged** to...				
...the **role** of english in...				
...to determine the **effectiveness** of...				

NOW, IT IS YOUR TURN

Utilize P-frames and explore other potential fillers to compose a text. Ensure to take into account the appropriate context or section where the P-frame is typically employed.

_____

_____

_____

_____

_____

_____

_____

_____

_____

5. After this activity, the teacher should help students understand that the statistical multivariate analysis reveals the proximities between P-frames, indicating their internal similarities. This helps underscore the importance of learning and applying tin academic writing. The teacher should exemplify it by explaining the dendrograma below:

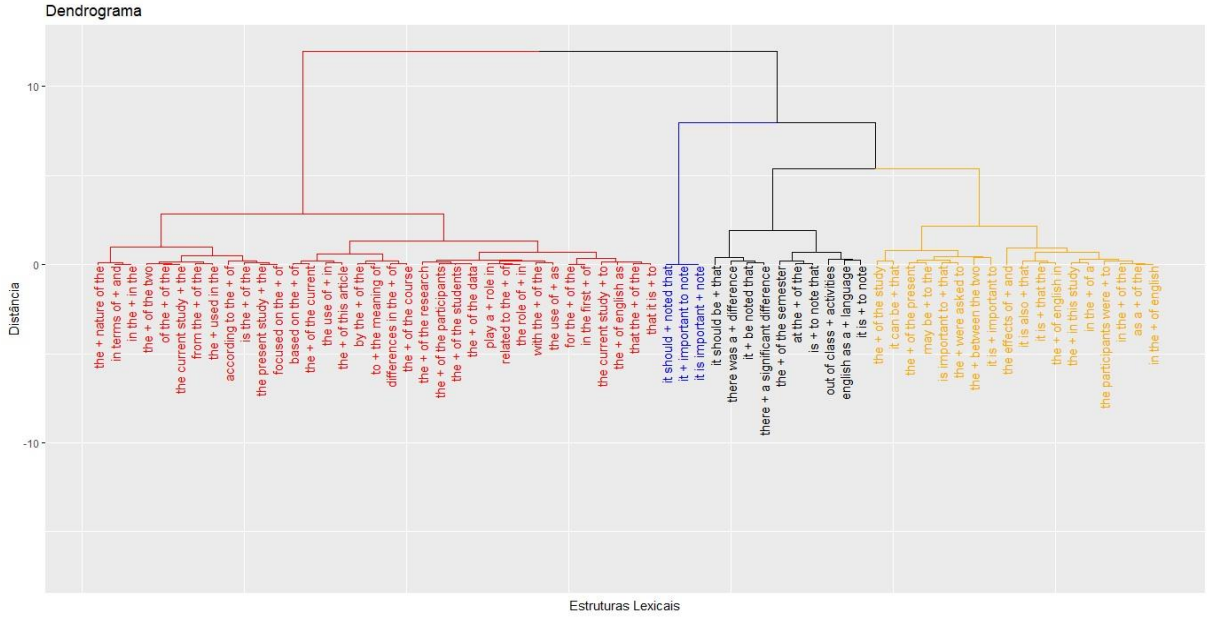


Figura 13 – Dendrograma a partir da Matriz de Distância

CAPÍTULO V

5. CONSIDERAÇÕES FINAIS

Este capítulo está estruturado em quatro seções. Na primeira seção, revisito os objetivos de pesquisa delineados no Capítulo 1, fundamentando-me nos resultados obtidos no capítulo anterior. Em seguida, abordo as implicações desses resultados para o campo da Linguística Aplicada. Na sequência, discuto as limitações do estudo e proponho algumas direções para pesquisas futuras. Finalizo com algumas considerações finais.

5.1 Retomando os objetivos da pesquisa

Ao propor a pesquisa, pretendia-se investigar a viabilidade de identificar Estruturas Lexicais (ELexs) utilizando o método direto (*Direct-Approach methodology*) descrito por Gray e Biber (2013), em um corpus especializado (CorAAL – *Corpus of Articles of Applied Linguistics*) de artigos da área da Linguística Aplicada, sob a perspectiva metodológica direcionada por corpus. No estudo conduzido por Gray e Biber (2013), a frequência é o único parâmetro para se identificar os PLs e, a partir dessa identificação, extrair as Estruturas Lexicais (*bundle-to-frame methodology*). Sintetizando, as ELexs eram identificadas a partir da identificação de PLs.

No entanto, nesta pesquisa, além de investigar a possibilidade de identificação das ELexs independente dos PLs, isto é, a frequência deixa de ser o único parâmetro investigativo, e consideramos analisar os princípios de variabilidade (*type-token* - TTR) e previsibilidade (*normalizad entropy*) nas estruturas internas das ELexs encontradas, por suspeitar que suas estruturas internas apresentavam similaridades. Para isso, aplicamos uma técnica estatística pouco usada em estudos linguísticos, a Análise Multivariada de Dados (AMD) aplicada a dados linguísticos.

Dentre as inúmeras possibilidades de análise multivariada (seções 2.5 e 3.4), optamos pela análise de agrupamento, também conhecida como análise de *clusters*. Essa abordagem possibilita otimizar as similaridades entre as variáveis, resultando em grupos mais homogêneos, ao mesmo tempo em que reduz as semelhanças entre grupos distintos (Hair *et al.*, 2009; Mingoti, 2023).

Para isso, três objetivos específicos foram definidos. O primeiro objetivo específico era identificar as ELexs mais frequentes no corpus especializado. Assim, nós partimos da premissa que a linguagem é formulaica (Wray, 2002) e, consequentemente, encontraríamos ELexs por todo o corpus. Desse modo, a abordagem delineada para esta pesquisa veio preencher uma lacuna investigativa em alguns aspectos.

Primeiramente, a Linguística de Corpus, enquanto metodologia de pesquisa tem sido amplamente adotada ao longo das últimas décadas em corpora de grande escala. No entanto, nesta pesquisa, optamos por utilizar um corpus pequeno (seções 2.1 e 2.2). Em concordância com Sinclair (1996, p. vii), reconhecemos que, embora um corpus de menor escala possa apresentar limitações em investigações de fenômenos linguísticos específicos, isso não compromete a qualidade da pesquisa. Além do mais, o pequeno corpus oferece um conjunto de evidências relevantes e confiáveis, podendo ser manipulado de forma ágil para cumprir um objetivo específico.

Desse modo, nesta pesquisa, o corpus especializado (CorAAL), de acordo com suas especificações (seção 3.1), possibilitou a investigação das múltiplas funcionalidades das ELexs em diferentes contextos linguísticos (Biber e Conrad, 1999; Biber e Barbieri, 2007; Gray e Biber, 2013; Staples *et al*., 2013). Isso resultou inicialmente na elaboração do Quadro Geral das unidades formulaicas (Tabela 10) em anexo. Após as apropriadas análises (seção 3.3), nos foi possível identificar as ELexs não provenientes de PLs, conforme detalhado na Tabela 06.

Além disso, considerando que o segundo objetivo específico, que consiste em compreender as características internas das ELexs e, sendo auxiliado pela linguagem R, conseguimos analisar as ELexs utilizando a técnica exploratória de sintetização da estrutura das variáveis, conhecida como análise de *clusters*, conforme a Figura 03.

O tratamento estatístico possibilitou a quantificação dos dados em análise, ampliando o propósito interpretativo dos fenômenos linguísticos por meio de gráficos e tabelas (Oushiro, 2022). A análise multivariada não se restringiu apenas à quantificação das variáveis que compõem as estruturas das ELexs, mas também explorou a complexidade das interações entre os elementos do corpus, tornando as relações entre as unidades lexicais, tanto interna quanto externamente, acessíveis para interpretação e compreensão (Hair *et al*., 2009).

A pesquisa conduzida por Gray e Biber (2013), pressupôs que as ELexs estão associadas a pelo menos uma sequência contínua. Porém, por suspeitar que as unidades lexicais apresentassem características internas de similaridades, adicionamos ao parâmetro frequência, os parâmetros de variabilidade e previsibilidade das ELexs (Tan e Römer, 2022)

(seção 2.4). Desse modo, com o auxílio da Linguagem R (formação de *scripts* – ANEXO F), as sequências lexicais (Tabela 10) foram divididas em grupos (*clusters*) de modo que os elementos semelhantes se agrupavam (homogeneidade) e elementos com características diferentes se distanciavam (heterogeneidade) entre grupos. A técnica aglomerativa parte do princípio que *clusters* específicos são agrupados pelo algoritmo segundo os valores de semelhança.

Assim, o método hierárquico, isto é, a razão entre variância total interna dos grupos e a variância entre grupos possibilitou a visualização dos *clusters*, conforme o dendrograma da Figura 13 e o agrupamento das ELexs, conforme a Figura 14. Com isso, concluímos que as ELexs possuem características próprias e podem ser identificadas e analisadas independentemente dos PLs.

O terceiro objetivo específico aborda como os resultados obtidos nesta pesquisa pode nortear a criação de tarefas que propicie aprendizagem orientada por dados. Entendemos que esta pesquisa, em última instância, contribui para *letramentos* (no plural) *acadêmicos*. Isso porque, segundo Green (2020, p. 41-42), a prática de letramento acadêmico é plural, na medida em que entrelaça "conhecimentos de uma gama muito ampla de áreas, mas é possível agrupar a maioria desses conhecimentos sob os títulos de conhecimento contextual, conhecimento declarativo e conhecimento procedimental". Segundo o autor, o desenvolvimento dos *letramentos acadêmicos* é gradual e, para aqueles que estão envolvidos no processo, resulta na sensação de pertencimento a uma comunidade específica.

Desse modo, como mencionado anteriormente, nosso ponto de partida pretendia investigar a possibilidade de se identificar ELexs utilizando uma abordagem direta, sob a perspectiva metodológica direcionada por corpus. Contudo, indo além, da análise baseada apenas na frequência, decidimos examinar os princípios de variabilidade (*type-token*) e previsibilidade nas estruturas internas da ELexs, devido à nossa hipótese de que as estruturas internas apresentassem similaridades. Para tal análise, empregamos a análise multivariada (técnica de agrupamento de *clusters*) ao conjunto de dados linguísticos. Essa abordagem estatística é incomum em estudos linguísticos e em análise de ELexs, porém, nos auxiliou na confirmação de que as ELexs são unidades com identidades próprias.

Assim, os resultados da análise do corpus (Tabela 06) revelam uma alternativa para o ensino da linguagem acadêmica, ajudando os alunos a desenvolver competências como escritores ao criarem textos que atendam aos padrões estabelecidos por especialistas em suas respectivas áreas (seção 4.3). Com isso, segundo Bocorny e Welp (2021, p. 1596), o ensino e a aprendizagem de ELexs de língua inglesa para fins acadêmicos "ampliam o repertório de

práticas de letramento de pesquisadores através da familiarização com gêneros que circulam nesse contexto". Apesar de esta pesquisa não trabalhar com o conceito de gênero discursivo, concordamos com as autoras, quanto à necessidade de oferecer meios e oportunidades de prática de letramento em língua inglesa.

5.2 Implicações dos resultados para a área da Linguística Aplicada.

Os resultados desta pesquisa evidenciam o valor das ELexs para o ensino e a aprendizagem da língua inglesa (seção 3). No entanto, qual a relevância desta pesquisa para a área da Linguística Aplicada (LA)? Para responder a esta pergunta, precisamos, mesmo que brevemente, relembrar o propósito da LA.

Segundo Menezes (2009, p. 26), a LA surge no senário acadêmico com o propósito de ensino de línguas estrangeiras, porém, atualmente é uma área responsável pela investigação transdisciplinar, "de novas formas de pesquisa e de novos olhares sobre o que é ciência". Na homepage da AILA (*International Association of Applied Linguistics*),

> Linguística Aplicada é um campo interdisciplinar e transdisciplinar de pesquisa e prática que lida com problemas práticos de linguagem e comunicação que podem ser identificados, analisados ou resolvidos aplicando teorias, **métodos** e resultados disponíveis da Linguística ou desenvolvendo **novos quadros teóricos e metodológicos** na Linguística para trabalhar nesses problemas. A Linguística Aplicada difere da Linguística em geral principalmente com relação à sua orientação explícita para problemas práticos e cotidianos relacionados à linguagem e à comunicação[44](grifos e tradução meus).

Assim, esta pesquisa cumpre os propósitos da LA ao usar métodos específicos (seção 3) para investigar um fenômeno linguístico, as unidades lexicais que podem ser usadas no ensino no contexto universitário, por exemplo. A motivação inicial surgiu da leitura do estudo conduzido por Gray e Biber (2013), que sugere que a identificação das ELexs (unidades

[44] No original, "Applied Linguistics is an interdisciplinary and transdisciplinary field of research and practice dealing with practical problems of language and communication that can be identified, analysed or solved by applying available theories, methods and results of Linguistics or by developing new theoretical and methodological frameworks in Linguistics to work on these problems. Applied Linguistics differs from Linguistics in general mainly with respect to its explicit orientation towards practical, everyday problems related to language and communication". Acesso em: 26 abril 2024, disponível em: https://aila.info/

lexicais descontínuas) devem ser feitas a partir dos PLs (unidades lexicais contínuas). Nossa inquietação, portanto, nos levou a procurar artigos que buscavam identificar e compreender as ELexs. No entanto, percebemos que há mais ênfase concentrada em estudos relacionados aos PLs em comparação com as ELexs.

Além disso, todos os artigos pesquisados traziam dois parâmetros em comum: a frequência com que as sequências lexicais apareciam no corpus e o ponto de partida para a identificação das ELexs era os PLs (Cortes, 2002; Biber, 2007; Römer, 2009; Yoon e Casal. 2020). Dessa maneira, essa lacuna na literatura acadêmica nos estimulou a presente pesquisa.

Assim, partimos de um corpus especializado, por acreditar que nos proporcionaria condições para observar fenômenos específicos (seções 2.1 e 2.2), também optamos por uma abordagem pouco utilizada em estudos relacionados às ELexs, conhecida como abordagem dirigida por corpus (*Corpus-Driven approach*).

Ademais, com o propósito de identificar as ELexs independentemente dos Pacotes Lexicais, adicionamos ao parâmetro frequência com que as unidades lexicais aparecem no corpus, dois outros parâmetros: o de previsibilidade e variabilidade em conformidade com o estudo dirigido por Tan e Römer (2022). Com estes parâmetros pretendíamos verificar a construção interna das unidades lexicais, por haver uma hipótese que a diferença entre os PLs e as ELexs dependia da variante interna da unidade lexical.

A extração das unidades lexicais no corpus foi conduzida utilizando o software AntConc, especificamente através da ferramenta chamada N-gram, a qual quantifica e normaliza os índices de previsibilidade e variabilidade das unidades lexicais. Isso nos permitiu não apenas avaliar as similaridades entre as unidades lexicais, mas também agrupá-las aplicando uma técnica multivariada de dados para formar clusters. A Tabela 10, em anexo, nos possibilitou a formação do dendrograma (Figura 13) e no agrupamento da Figura 14.

Ao triangular as informações de frequência, previsibilidade e variabilidade,verificação das entropias internas das unidades lexicais, nos foi possível identificar unidades lexicais que compartilhavam características similares. Assim, após compará-las com o estudo dirigido com Biber *et al*. (1999), constatamos que as unidades lexicais da Tabela 06 não estavam incluídas na proposta taxonômica do autor. Além do mais, observamos que as unidades lexicais não se encaixavam nas características dos PLs, mas sim nas características das ELexs, conforme Quadro 07, seção 2.4.

Assim, pelo exposto nos parágrafos anteriores, os resultados da presente pesquisa contribuem com a área da Linguística Aplicada, por investigar de modo interdisciplinar e transdisciplinar a identificação de ELxes em um corpus linguístico.

5.3 As limitações do estudo e algumas sugestões de pesquisa futuras

Reconhecendo que toda pesquisa tem suas restrições, torna-se relevante destacar algumas limitações apresentadas neste estudo que, no entanto, não invalida os seus resultados.

Uma das limitações é quanto ao tamanho do corpus estudado. O corpus especializado (CorAAL) atendeu ao propósito desta pesquisa no que diz respeito à investigação das ELexs. No entanto, concordamos com Sinclair (2008, p. vii) que, se o objetivo é investigar padrões lexicais específicos, isto é, a frequência com certa repetição lexical ocorre em uma língua, um corpus em grande *escala* certamente seria mais apropriado.

Por outro lado, se a pesquisa tem um objetivo específico, um corpus pequeno pode servir a esse propósito. Neste caso, nossa hipótese inicial, era que as características internas de uma unidade lexical ofereciam características suficientes para qualificá-la como uma ELex independente de um PLs. Obviamente, apesar das suspeitas se confirmarem, é necessário aplicar a mesma metodologia e método a um corpus de grande *escala* para se verificar se as conclusões desta pesquisa se confirmam.

Outra limitação diz respeito à composição do corpus especializado. O Corpus de Artigos de Linguística Aplicada (CorAAL) é constituído por artigos acadêmicos da área da Linguística Aplicada (seção 3.1), datados entre 2014 e 2018. No entanto, uma questão importante deve ser considerada: será que em um corpus maior e com artigos relacionados à outras áreas do conhecimento, os resultados se confirmam?

Como dito anteriormente, há poucas pesquisas relacionadas à identificação e classificação das ELexs. O termo ELex foi cunhado por Fletcher a partir do British National Corpus (BNC) para descrever sequências lexicais descontínuas com lacunas variáveis, conhecidas, em inglês, como phrasal-frame ou “p-frame”, em 2003, tornando-se assim uma área de estudo ainda pouco explorada na academia. Portanto, não existe uma classificação taxonômica estabelecida para as ELexs, o que nos levou, nesta pesquisa, a utilizar a taxonomia dos Pacotes Lexicais propostos por Biber (1999) para fins comparativos.

Apesar das limitações aqui apresentadas, esperamos que os resultados desta pesquisa possam fornecer reflexões sobre as possibilidades de intensificar e explorar as ELexs a partir de outros corpora.

Muitas pesquisas podem ser realizadas com um viés diferente do que foi proposto nesta pesquisa. Por exemplo, além de utilizar um corpus de grande escala, as análise poderiam considerar as seções dos artigos individualmente, como o resumo, introdução, metodologia, resultados, discussão e conclusão.

Há, também, que se considerar a utilização de outro método estatístico. O método estatístico multivariado é apenas uma das muitas possibilidades de investigação que pode ser aplicada aos estudos linguísticos. Por fim, uma questão intrigante é se a Inteligência Artificial seria capaz de identificar e classificar as ELexs em corpus. Todas essas perspectivas podem abrir novos horizontes para novas pesquisas e ampliar o entendimento sobre as ELexs.

5.4 Considerações finais.

Os resultados desta pesquisa mostram que as ELexs são unidades lexicais que podem ser identificadas e analisadas independentemente dos PLs. No entanto, antes de uma nova proposta com o mesmo viés da presente pesquisa, estudos periféricos fazem-se necessários, por exemplo, estudos relacionados à classificação taxonômica das ELexs e a ampliação do corpus especializado.

Este estudo considera o estudo das ELexs útil para o curso de nível superior em Letras/Habilitação em inglês, pois proporciona a oportunidade de desenvolvimento da linguagem formulaica empregada na escrita acadêmica. Além disso, as pesquisas direcionadas por corpus são indutivas (Biber, 2012), sendo assim, com o auxilio computacional, os padrões linguísticos emergem do corpus independente de categorias linguísticas, dando ao pesquisador liberdade investigativa quanto ao uso real da língua em textos autênticos na criação de tarefas acadêmicas (Almeida *et al*., 2023).

Espero que esta pesquisa possa servir como ponto de partida para novas propostas e ideias investigativas na área do ensino e aprendizagem de escrita acadêmica, bem como, desperte o interesse por análises estatísticas aplicadas à Linguística.

6 REFERÊNCIAS

ALMEIDA, V.; ORFANÒ, B. M.; DUTRA. D. Is there a better choice? Verb–noun combinations in academic writing. In: Teaching English with Corpora. Abingdon, Oxon; New York, NY: Routledge, 2023, p. 228.

ANTHONY, L. AntConc. v. 4.2.0/2022 [Computer Software]. Tokyo, Japan: Waseda University, 2021. Disponível em: http://www.antlab.sci.waseda.ac.jp/. Acesso em: 03 out. 2022.

ASSIS, J. P.; SOUSA, R. P.; DIAS, C. T. S. Glossário de Estatística. Mossoró: EdUFERSA, 2019.

BERBER SARDINHA, T. Lingüística de corpus: histórico e problemática. Delta: documentação de estudos em linguística teórica e aplicada, 2000. v. 16, p.323 -367 Disponível em: encurtador.com.br/hlG35. Acesso em: 03 out. 2022.

BERBER SARDINHA, T. Linguística de corpus. São Paulo: Manole, 2004.

BIBER, D. Representativeness in corpus design. Literary and Linguistic Computing, 243-257. 1993.

BIBER, D; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. Longman grammar of spoken and written English. London: Pearson Eduction Limited, p. 990-991. 1999.

BIBER, D.; CONRAD, S. Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), Out of corpora Amsterdam: Rodopi. pp. 181-190. 1999.

BIBER, D. Using corpora to explore linguistic variation. Amsterdam: John Benjamins Publishing Company, 2002. p. 131-145.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at ...: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, v.25, n.3. 2004. p. 371–405.

BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. English for Specific Purposes, v. 26, p. 263-286, 2007.

BIBER, D. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. International Journal of Corpus Linguistics, v. 14, n. 3, p. 273-311, 2009.

BIBER, D. Corpus-based and corpus-driven analyses of language variation and use. In: Heine, B.; NARROG, H. (Ed.). The Oxford handbook of linguistic analysis. Oxford: Oxford University Press, 2010. p. 201-203.

BIBER, D. Register as a predictor of linguistic variation. In: Corpus Linguistics and Linguistic Theory, ed. 8, 2012, p. 9–37.

BOCORNY, A. E. P.; REBECHI, R.; REPPEN, R.; DELFINO, M. C. N.; LAMEIRA, V. .M. A produção de artigos da área das ciências da saúde com o auxílio de key lexical bundles: um estudo direcionado por corpus. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada [online]. 2021, v. 37, n. 1. Disponível em: <https://doi.org/10.1590/1678-460X2021370101>. Acesso em: 03 nov. 2022.

BOCORNY, A. E. P.; WELP, A. O desenho de tarefas pedagógicas para o ensino de Inglês para Fins Acadêmicos: conquistas e desafios da Linguística de Corpus. Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1529-1638, 2021.

BUTLER, C. S. Collocational Frameworks in Spanish. International Journal of Corpus Linguistics, v.3, n.1, 1998, p.1–32.

CASAL, J. E.; KESSLER, M. Form and rhetorical function of phrase-frames in promotional writing: A corpus- and genre-based analysis. 2020 Disponível em: https://doi.org/10.1016/j.system.2020.102370. 2020. Acesso em: 31 out. 2022.

CONRAD, S. M. The importance of corpus-based research for language teachers. In: System. v. 27, p. 1-18, 1999.

CORTES, V. Lexical bundles in Freshman composition. In: Reppen, R.; Fitzmaurice, S. M.;

ELLIS, N. C.; SIMPSON-VLACH, R.; MAYNARD, C. Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. TESOL Quarterly. 2008, v. 42, n.3, p. 375–396.

FLETCHER, W. Phrases in English. 2003. Disponível em: <http://phrasesinenglish\>. Acesso em: 30 out. 2022.

FLOWERDEW, L. The argument for using English specialized corpora to understand academic and professional language. In: Discourse in the Professions: Perspectives from corpus linguistics. edited by, Ulla Connor and Thomas A. Upton. 2004.

FRANCIS, G. A corpus-driven approach to grammar -Principles, methods and examples. In M. Baker, G. Francis & E. Tognini-Bonelli. Eds. Text and technology: In honour of John Sinclair. Philadelphia/Amsterdam: John Benjamins. 1993

GARNER, J. R. A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. International Journal of Learner Corpus Research, v. 2, 2016.

F. Preface. In: Phraseology: An interdisciplinary perspective. GRANGER, S.; MEUNIER, In:

GHADESSY, M. Criteria for English text types. Comunicação apresentada no 8º Euro-International Systemic Functional Workshop, Nottingham Trent University, Nottingham, Reino Unido, 24 de julho de 1996.

GHADESSY, M.; HENRY, A.; Roseberry, R. L. Small corpus studies and ELT : theory and practice. Amsterdam: John Benjamins Publishing Company amsterdam / Philadelphia. 2001.

GILQUIN, G.; GRANGER, S. How can data-driven learning be used in language teaching?. In: The Routledge Handbook of Corpus Linguistics. Publisher: Routledge, 2010.

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

GRAY, B.; BIBER, D. Lexical frames in academic prose and conversation. International Journal of Corpus Linguistics, v. 18, n. 1, p. 109–136, 2013.

GREEN, A. Exploring language assessment and testing: language in action. London: Routledge, 2020.

GRIES, S. T. Quantitative Linguistics with R: A Pratical Introduction. New York: Routledge. 2009.

HAIR, J. F. J.; BLACK, W. C.; Babin, B. J.; Anderson, R. E.; Tathan, R. L. Análise Multivariada de Dados. Trad., 6ª ed. Bookman Companhia Editora Ltda. Artmed Editora. 2009.

HASLWANTER, T. An Introduction to Statistics with Python: With Applications in the Life Sciences. Springer International Publishing Switzerland. 2016.

HOWARTH, P. Phraseology and Second Language Proficiency. In: Applied Linguistics. Oxford University Press. 24-44. 1998.

HUNSTON, S.; FRANCIS, G. Pattern Grammar. A corpus-driven approach to the lexical grammar of English. Amsterdam: John Benjamins. 1999.

HYLAND, K. Activity and Evaluation: Reporting Practices in Academic Writing**.** In: J. FLOWERDEW (ED) ACADEMIC DISCOURSE.LONDON, LONGMAN. PP 115-30. 2002.

HYLAND, K. Academic clusters: text patterning in published and postgraduate writing. International Journal of Applied Linguistics, v. 18, n. 1, p. 41–62, 2008.

LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. Análise de Dados Multivariados. São Paulo: Cengage Learning, 2011.

LEFFA, V. J.; IRALA, V. B. O ensino de outra(s) língua(s) na contemporaneidade: questões conceituais e metodológicas. 2014, p. 21-48. Disponível em: encurtador.com.br/DMOPW. Acessado em: 30 out. 2022.

MARTINEZ, R.; SCHMITT, N. A Phrasal Expressions List. Applied Linguistics. Oxford University Press. 2012. p. 299-320.

MENEZES, V.; SILVA, M. M.; GOMES, I.F. Sessenta anos de Lingüística Aplicada: de onde viemos e para onde vamos. In: PEREIRA, R.C.; ROCA, P. Linguistica aplicada: um caminho com diferentes acessos. São Paulo: Contexto, 2009.

MINGOTI, S. A. Análise de Dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: Editora da UFMG, 2023.

OUSHIRO, L. Introdução à Estatíticas para Linguístas [livro eletrônico]. 1ª ed. Campinas, SP: Editora da Abralin. Disponível em https://ead.abralin.org/. Acesso em 14 mai. 2022.

PRINA DUTRA, D., BERBER SARDINHA, T. A multi-dimensional typology of English research article sections. Paper presented online at the American Association for Applied Linguistics Conference (AAAL). 2021.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding Groups in Data : An Introduction to Cluster Analysis. A John Wiley & Sons, INC., Publication, 1990.

KUMAR, U.; KUMAR, V.; KAPUR, J. N. Normalized measures of entropy. International Journal of General Systems, v.12, n.1. 1986. p. 55–69.

O'KEEFFE, A.; MCCARTHY, M. Historical perspective: what are corpora and how have they evolved? In: The Routledge Handbook of Corpus Linguistics. 1ª ed: Routledge, 2010. p. 3.

RÖMER, U. English in Academia: Does Nativeness Matter? Anglistik: International Journal of English Studies. v. 20, n. 2, 2009. p. 89–100.

R CORE TEAM R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. Disponível em https://www.R-project.org/. Acessado em 20 abr. 2022.

RÖMER, U. The inseparability of lexis and grammar. In: Annual Review of Cognitive Linguistic: Corpus Linguistic perspective. John Benjamins Publishing Company. 2009.

RÖMER, U. Using General and Specialized Corpora in English Language Teaching: Present, Past and Future. In: Corpus-Based Approaches to English Language Teaching. London: Continuum International Publishing Group, 2010.

RENOUF, A.; SINCLAIR, J. Collocational frameworks in English. In: K. Aijmer & B. Altenherg (Ed.), English Corpus Linguistics. London: Longman, 1991. p. 128-143.

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

SINCLAIR, J. Preface. In: Small corpus studies and ELT : theory and practice. edited by Mohsen Ghadessy, Alex Henry, Robert L. Roseberry. 2001.

SINCLAIR, J. Trust the text: Language, Corpus and Discourse, London: Routledge. 2004.

SINCLAIR, J. Preface. In: Phraseology: An interdisciplinary perspective. Edited by Sylviane Granger; Fanny Meunier, John Benjamins Publishing Company, Amsterdam Philadelphia. 2008. p. xv.

STAPLES, S.; EGBERT, J.; BIBER, D.; McCLAIR, A. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. In: Journal of English for Academic Purposes. 2013.

STUBBS, M. An example of frequent English Phraseology: Distribution, structures and functions. In: FACCHINETTI, R. (Ed.). Corpus Linguistics 25 years on. Amsterdam: Rodopi, 2007. p. 89-105.

TAN, Y.; RÖMER, U. Using phrase-frames to trace the language development of L1 Chinise Learners of English, 2022.

TOGNINI-BONELLI, E. Corpus Linguistics at Work. Amsterdam: John Benjamins Publishing Company. 2001.

WELP, A.; DIDIO, A. R.; FINKLER, B. Questões contemporâneas no cinema e na literatura: o desenho de uma sequência didática para o ensino de inglês como língua adicional. 2019.

WOOD, D. Fundamentals of Formulaic Language – An Introduction. Bloomsbury Academic: London, 1ª ed., 2015.

WRAY, A.; PERKINS, M. R. The functions of formulaic language: An integrated model. Language & Communication, 2000. p. 01-28.

WRAY, A. Formulaic Language and the Lexicon, Cambridge: Cambridge University Press, 2002.

YOON, J.; CASAL, J. E. Rhetorical structure, sequence, and variation: A step-driven move analysis of applied linguistics conference abstracts. In: International Journal of Applied Linguistcs, v. 30, 2020. p. 462-478.

VILLALVA, A.; SILVESTRE, J. P. Introdução ao estudo do léxico: Descrição e análise do Português. Petrópolis, RJ: Vozes, 2014.

ANEXO A - Tabela 10 – Quadro geral das ULs encontradas no CorAAL.

Frame	Biber (1999)	Freq	TTR	Previsibilidade	
at the+of the (*)	√	143	0.15	0.69	5,7%
in the+of the (*)	√	143	0.48	0.87	5,7%
of the+of the	√	107	0.75	0.97	4,3%
the+of the present (*)	√	91	0.26	0.80	3,7%
the+of the study (*)	√	86	0.29	0.86	3,5%
english as a+language (*)	√	85	0.12	0.58	3,4%
for the+of the (*)		64	0.64	0.96	2,6%
that the+of the (*)	√	64	0.62	0.94	2,6%
with the+of the	√	58	0.69	0.93	2,3%
it is+that the	√	56	0.41	0.87	2,2%
based on the+of (*)		50	0.58	0.93	2,0%
the+in this study	√	46	0.52	0.87	1,8%
as a+of the	√	44	0.48	0.86	1,8%
the+of the participants (*)	√	43	0.72	0.94	1,7%
by the+of the	√	42	0.62	0.87	1,7%
in the+of a	√	40	0.52	0.85	1,6%
in terms of+and	√	39	0.90	0.99	1,6%
related to the+of (*)		37	0.68	0.95	1,5%
the use of+in	√	37	0.54	0.93	1,5%
the+of the current	√	35	0.57	0.93	1,4%
is important to+that	√	35	0.31	0.79	1,4%
the role of+in	√	34	0.68	0.94	1,4%
it is+to note (*)	√	33	0.15	0.63	1,3%
it can be+that	√	33	0.24	0.89	1,3%
the participants were+to (*)		33	0.48	0.83	1,3%
the+of the students	√	31	0.71	0.95	1,2%
the present study+the	√	31	0.81	0.97	1,2%
the+between the two (*)		30	0.37	0.80	1,2%
the+of the research	√	30	0.70	0.97	1,2%
from the+of the	√	30	0.77	0.96	1,2%
it should be+that (*)	√	30	0.20	0.45	1,2%
differences in the+of		29	0.59	0.86	1,2%
it+be noted that	√	28	0.11	0.45	1,1%
the+of the course	√	28	0.57	0.87	1,1%
the+were asked to	√	28	0.29	0.78	1,1%
may be+to the	√	28	0.25	0.78	1,1%
the+of the data	√	27	0.70	0.94	1,1%

the+of the two	√	27	0.74	0.97	1,1%
the+of the semester	√	26	0.15	0.70	1,0%
focused on the+of (*)		26	0.81	0.98	1,0%
the+nature of the	√	25	0.88	0.99	1,0%
play a+role in (*)		25	0.68	0.94	1,0%
according to the+of (*)		25	0.84	0.97	1,0%
in the first+of	√	25	0.64	0.97	1,0%
out of class+activities (*)		25	0.08	0.63	1,0%
is+to note that	√	24	0.17	0.68	1,0%
the+of english in (*)		24	0.42	0.90	1,0%
is the+of the	√	24	0.79	0.96	1,0%
it should+noted that	√	24	0.04	0.00	1,0%
the+of english as	√	23	0.61	0.91	0,9%
to+the meaning of (*)		23	0.61	0.87	0,9%
it is also+that	√	23	0.39	0.85	0,9%
the effects of+and	√	23	0.48	0.70	0,9%
the+used in the	√	22	0.77	0.97	0,9%
it is+important to	√	22	0.32	0.72	0,9%
that it is+to	√	22	0.64	0.92	0,9%
it+important to note	√	21	0.05	0.00	0,8%
in the+of english	√	21	0.48	0.86	0,8%
it is important+note	√	21	0.05	0.00	0,8%
there was a+difference	√	21	0.14	0.46	0,8%
the+of this article	√	20	0.55	0.91	0,8%
there+a significant difference	√	20	0.10	0.47	0,8%
in the+in the	√	20	0.90	0.99	0,8%
the current study+the	√	20	0.75	0.97	0,8%
the current study+to	√	20	0.65	0.90	0,8%
the use of+as	√	20	0.70	0.90	0,8%

Nota:

(*/√) Estruturas Lexicais, identificadas tanto em Biber *et al.* (1999) quanto nesta pesquisa, são derivadas de pacotes lexicais e encontram-se presentes em no mínimo cinco textos.

(√) Estruturas Lexicais encontradas somente na pesquisa conduzida por Biber *et. al* (1999).

(*) Estruturas Lexicais encontradas somente nesta pesquisa, ou seja, não presentes na pesquisa conduzida por Biber *et al.* (1999). Não são originadas em pacotes lexicais.

ANEXO B - Quadro 04 - Quadro sinóptico dos principais padrões lexicais segundo Hunston e Francis (1999)

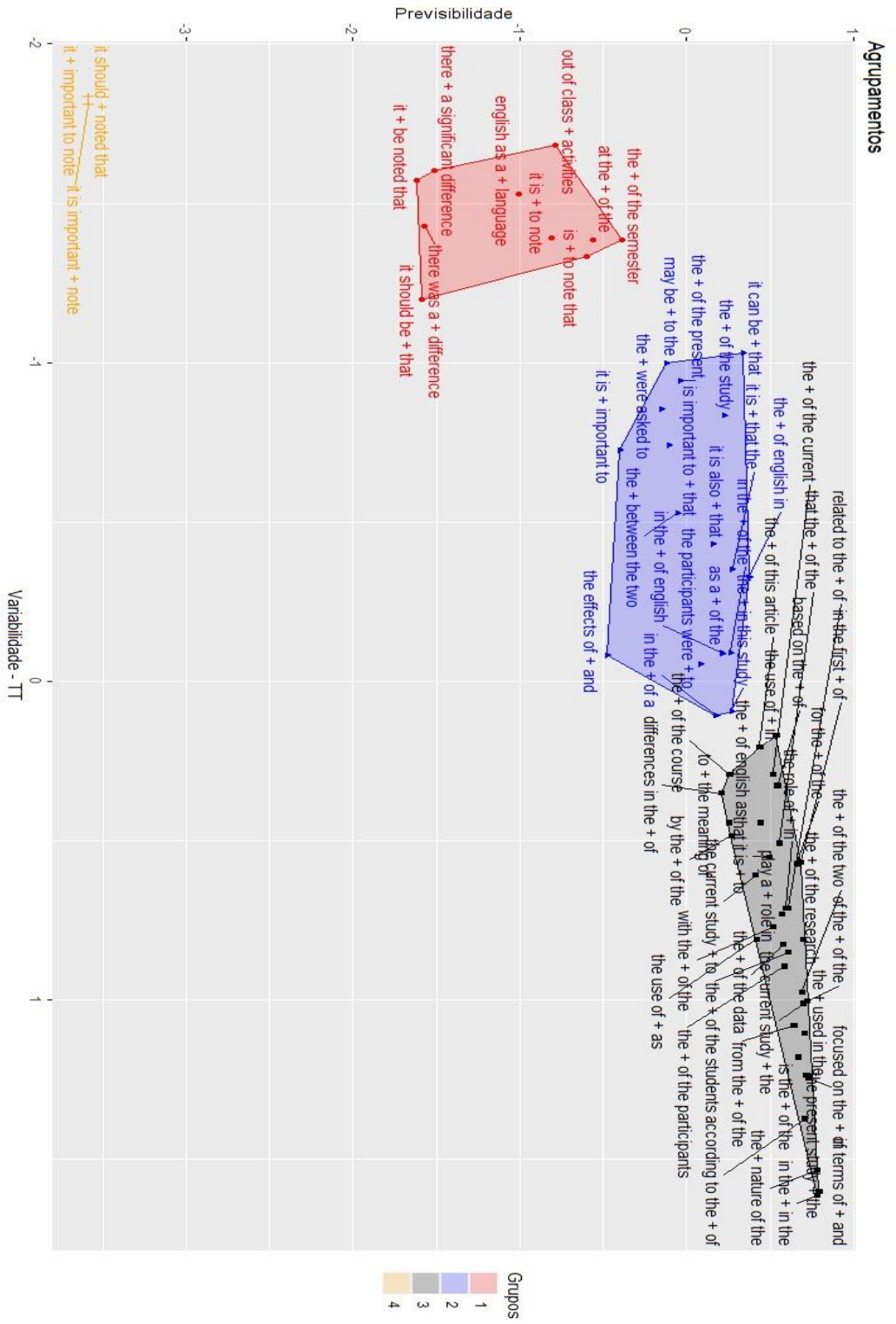
The patterns of verbs	The verb is followed by a single noun group, adjective group, or clause, yielding the following patterns:
V n	I broke my left leg
V pl-n	The research compares two drugs
V pron-refl (reflexive pronoun)	I enjoyed myself.
V amount	*Two and two make four.*
V adj	*He escaped unhurt.*
V -ing	*She started walking.*
V to-inf	*John began to laugh.*
V inf (bare nfinitive)	*I helped save these animals*.
V that	*We agreed that she was not to be told.*
V wh	*A passer-by inquired why the television cameras were there.*
V wh-to-inf (to infinitive clause introduced by a wh-word)	*I have forgotten what to say.*
V with quote	*'Hello', he said.*
V so/not	*I think so.*
V as if/as though	*You look as if you've seen a ghost.*
V *and* v	*I'll go and see him.*
	The verb is followed by a prepositional phrase or adverb group. In some cases, there is a wide range of adverbs and prepositions following the verb, and these cannot be specified.
V prep/adv	*He ran across the road.*
	Sometimes only an adverb can be used. This pattern is
V adv	*Sarah has fair skin that burns easily.*
	Sometimes only a prepositional phrase can be used. This pattern is
V prep	*She chewed on her pencil.*
	In other cases, the verb is followed by a noun group, adjective group, '-ing' clause or wh-clause introduced by a specific preposition. This pattern is **V *about* n, V *at* n, V *as* adj, V *by* -ing** etc., depending on the preposition. *He was grumbling about the weather*. *The rivals shouted at each other.*
	The prepositions which are used in patterns like this are as follows: ***about, across, after, against, around/round, as, as to, at, between, by, for, from, in, in favour of, into, like, of, offf, on, onto, out of, over, through, to, towards, under, with.*** Some times the adverb ***together*** is used in the pattern **pl-n V *together*** *The whole team must pull together.*
	The verb is followed by a noun group and another element such as another

	noun group, an adjective group, a that-clause, a wh-clause or an '-ing' clause, yielding the following patterns:
V n n	*I wrote him a letter.*
V n adj	*The darkness could drive a man mad.*
V n -ing	*I kept her waiting.*
V n to-inf	*My advisers counselled me to do nothing.*
V n inf	*She heard the man laugh.*
V n wh	*He showed me where I should go.*
V n wh-to-inf	*I'll show you how to do it.*
V n with quote	*'We'll do it', she promised him.*
V n -ed (the past participle form of another verb)	*I had three wisdom teeth extracted.*
	The verb is followed by a noun group and a prepositional phrase or adverb group. In some cases there is a wide range of adverbs and prepositions following the verb, and again these cannot be specified. This pattern is
V n prep/adv	*Andrew chained the boat to the bridge*
	Stir the sugar in.
	Sometimes only an adverb can be used. This pattern is **V n with adv**, where the adverb comes either before or after the noun group.
	He switched the television on.
	He switched on the television.
	Sometimes the pattern is formed with the word *way* and an adverb group or prepositional phrase. This pattern is
V *way* prep/adv	*She ate her way through a pound of chocolate.*
	In other cases, the verb is followed by a noun group and another noun group, adjective group or wh-clause introduced by a specific preposition. This pattern is **V n *about* n, V n *at* n, V n *as* adj** etc. depending on the preposition.
	I warned him about the danger. *I saw the question as crucial.*
	The prepositions which are used in patterns like this are almost but not quite the same as those in 2 above: ***about, against, as, as to, at, between/among, by, for, from, in, into, of, offff, on,onto, out of, over, to, towards, with***. Sometimes the adverb ***together*** is used in the pattern **pl-n V with *together***
	We stuck the pieces together.
	The verb pattern contains the word *it.* The main patterns are as follows. **Introductory *it:***
	It doesn't matter what you think.
***it* V *to* n clause**	*It sounds to me as if you don't want to help her.*
***it* V prep clause**	*It came to light that the plane had not been insured.*
***it be* V-ed clause**	*It is thought that the temple was used in the third century.*
***it* V n clause**	*It struck me that the story would make a good film.*
***it* V adj clause**	*It feels good to have finished a piece of work*.
V *it* clause	*I hate it when she's away.*

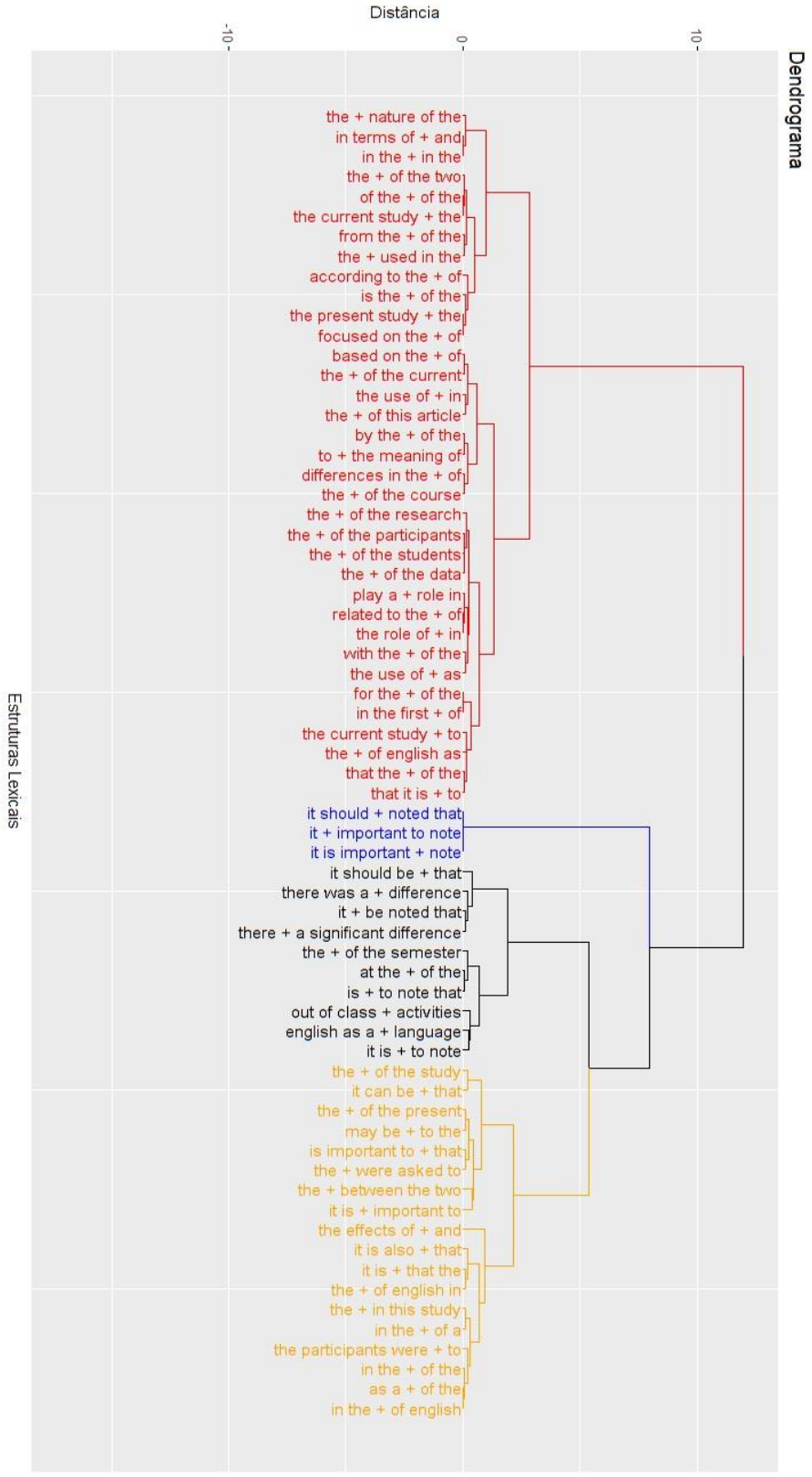
V *it to* n clause	*I owe it to my parents to work hard.*
V *it as* n/adj clause	*He would take it as an insult if I left.* *He regards it as significant that the Government is suggesting cuts.*
V *it* n clause	*They felt it their duty to visit her in hospital.*
V *it* adj clause	*I think it best if you tell him the truth.*
'General' *it:*	
***it* V**	*It snowed all afternoon.*
***it* V adj**	*It was very windy.*
***it* V adj prep/adv**	*It's nice here.*
***it* V n**	*It's blowing a gale.*
***it* V *to* n**	*It got to the point where we couldn't bear to be in the same room as each other.*
***it* V prep/adv that**	*It says here that they have live music.*
V *it*	*They didn't make it.*
V *it prep/adv*	*My family hated it in Southampton.*
The patterns of nouns	Patterns with elements preceding the noun
***a* N; *the* N -** The noun is preceded by an indefinite or definite article:	*a <u>cinch</u>, a <u>standstill</u>; the <u>blues</u>, the <u>bourgeoisie</u>.*
poss N - The noun is typically preceded by a possessive determiner like *my* or *your*, or a possessive formed from a noun group:	*She had tidied away her possessions.* *I give you my word* *My husband's sister came to stay.*
adj N The noun is preceded by an adjective:	*He was a tough customer.* *She's a smart dresser.*
n N The noun is preceded by another noun:	*A window cleaner was arrested.*
***from* N, *on* N, *to* N** etc. The noun is preceded by a specific preposition:	*I've been blind in my right eye from birth.* *The film was shot on location in Washington.* *They went to school together every day.* The prepositions most frequently used in patterns like this are as follows: ***at, by, from, in, into, on, out of, under, with***. **supp N** The noun is preceded by a range of the elements given above: determiner, possessive determiner or possessive noun group, adjective or noun.
	Patterns with elements following the noun
N to-inf	*All four teams have shown a desire to win.*
N that	*There was a suggestion that the whole thing was a joke.*
N prep The noun is followed by a prepositional phrase	

introduced by a wide range of prepositions. **N *of* n, N *for* n, N *from* n** etc. The noun is followed by a prepositional phrase introduced by a specific reposition.	*It was the latest in a series of acts of violence.* *Their hatred for one another is legendary.* *The threat from terrorists is at its highest for two years.*
	The prepositions most frequently used in patterns like this are as follows: ***about, against, among, as, at, behind, between, for, from, in favour of, in, into, of, on, over, to, towards, with***. In addition there is the pattern **N with supp**, which means that the noun is both preceded by a range of the elements mentioned above, and followed by them.
The patterns of adjectives	
ADJ -ing	*I felt uncomfortable watching him.*
ADJ to-inf	*The print was easy to read.*
ADJ that	*I am absolutely horrified that this has happened.*
	ADJ prep The adjective is followed by a prepositional phrase introduced by a wide range of prepositions. **ADJ *as* n, ADJ *of* n, ADJ *on* n** etc. The adjective is followed by a prepositional phrase introduced by a specific preposition.
	We felt inadequate as parents.
	I think he's fully aware of those dangers.
	He's always been very dependent on me.
	The prepositions most frequently used in patterns like this are as follows: ***about, against, as, as to, at, between, by, for, from, in, into, of, offff, on, with.*** It must be stressed that the patterns listed above are only the major ones. There are many more; full lists can be found in Francis et al. (1996; 1998).

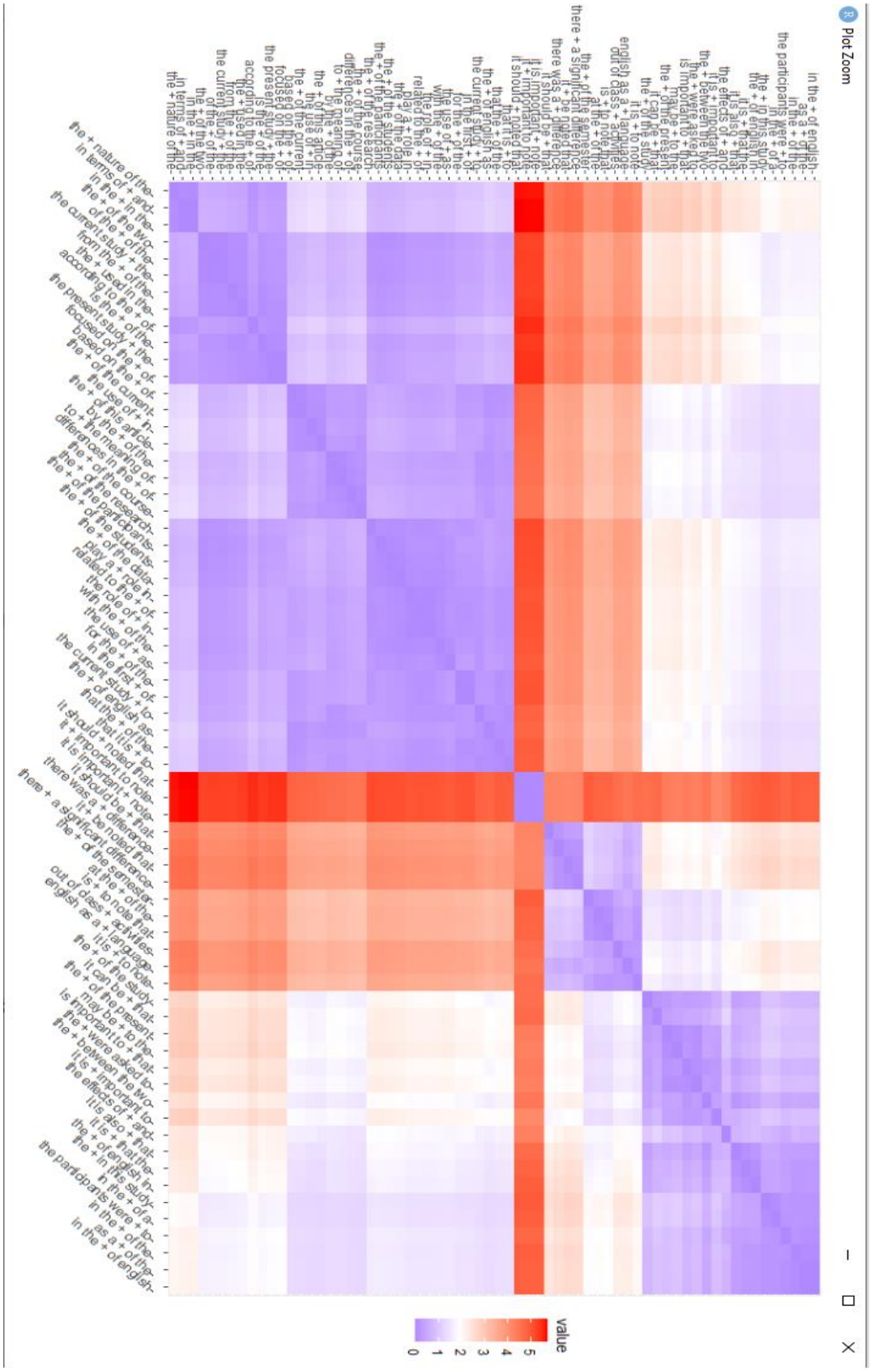
ANEXO C - Figura 14 – Agrupamento a partir do Método Hierárquico



ANEXO D - Figura 13 – Dendrograma a partir da Matriz de Distância



ANEXO E – Figura 07 - Matriz de Distância



ANEXO F – Scripts desenvolvido em Linguagem R

```
install.packages('readxl', dependencies = T)
library(readxl)
library(factoextra)
#Importar dados da planilha
setwd('C:/Users/Edilson/OneDrive/Área de Trabalho/CorAAL')
#Manipulação dos dados
arquivo <- read_excel("ELexs_66.xlsx", sheet = 1,
                      col_types = c('text', rep('numeric',3)))
arquivo <- as.data.frame(arquivo)
row.names(arquivo) <- arquivo$Elxs
arquivo$Elxs = NULL
arquivo$Freq = NULL
#normalização dos dados (arquivo)
arquivo.n <- scale(arquivo)
#Cálculo da medida de Distância - (Distância Euclidiana)
dista <- dist(arquivo.n)
p.dista <- fviz_dist(dista,gradient = list(low='blue',mid='white',high='red'))
ggsave(filename = 'D.euclidiana.png')
#Dendrogrma - (Árvore Hierárquica)
arvore.h <- hclust(dista,method = 'ward.D2')
fviz_dend(arvore.h, k= 4,labels_track_height = 17,
          ylab = 'Distância',
          xlab = 'Estruturas Lexicais',
          main = 'Dendrograma',
          palette = c('red','blue','black','orange'),
          horiz = F, ggtheme = theme_gray())
#Agrupamentos (elíptico)
arquivo$grupos <- cutree(arvore.h,h=4)
grupos <- cutree(arvore.h,h=4)
fviz_cluster(list(data=arquivo,cluster=grupos),
             choose.vars = c('TT','Prev'), repel = T,
             palet = c('red','blue','black','orange'),
```

```
        ggtheme = theme_gray(),
        show.clust.cent = F, main = 'Agrupamentos',
        xlab = 'Variabilidade - TT',
        ylab = 'Previsibilidade', c('red','blue','black','orange'))+
 labs(fill ='Grupos', c('red','blue','black','orange'))+
 guides(col=F,shape=F)
ggsave(filename = 'Agrupamentos.png')
```