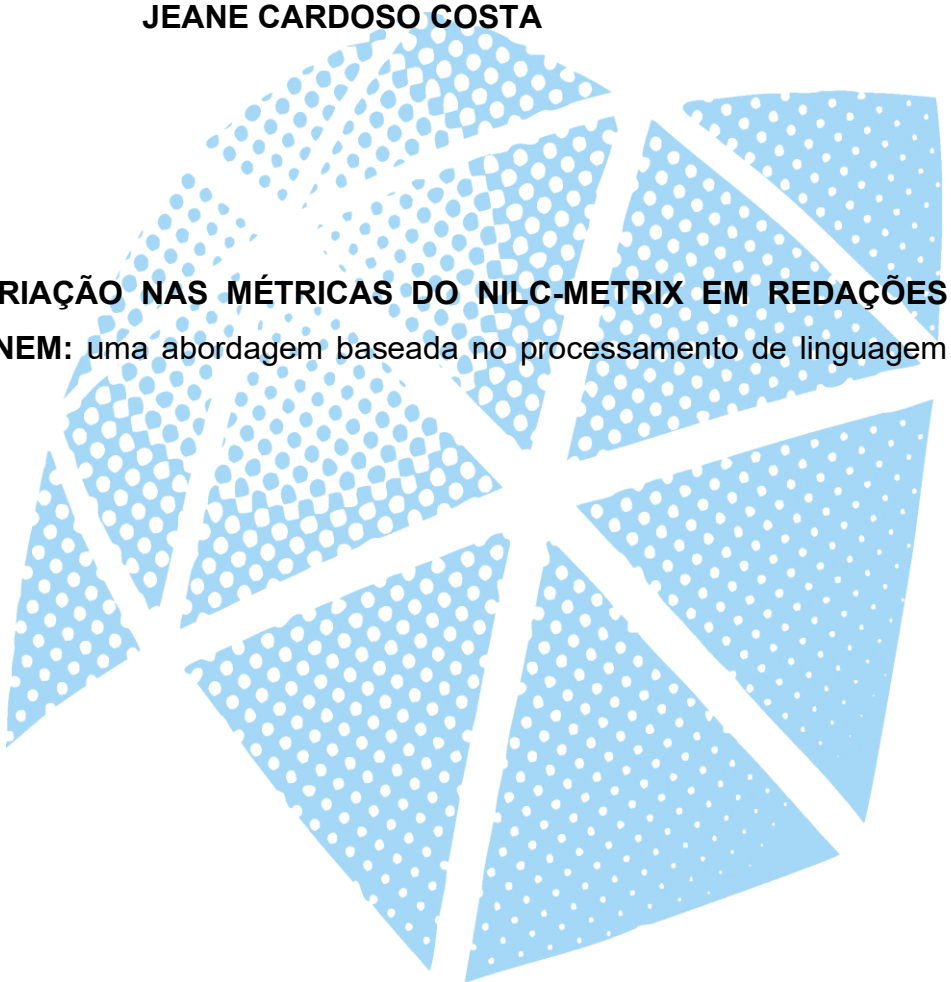
UNIVERSIDADE ESTADUAL PAULISTA – UNESP

Instituto de Biociências, Letras e Ciências Exatas - Campus de São José do Rio Preto

JEANE CARDOSO COSTA

ESTUDO DA VARIAÇÃO NAS MÉTRICAS DO NILC-METRIX EM REDAÇÕES NOTA MIL DO ENEM: uma abordagem baseada no processamento de linguagem natural



São José do Rio Preto - SP

2025

JEANE CARDOSO COSTA

ESTUDO DA VARIAÇÃO NAS MÉTRICAS DO NILC-METRIX EM REDAÇÕES NOTA MIL DO ENEM: uma abordagem baseada no processamento de linguagem natural

Tese apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, como parte dos requisitos para a obtenção do título de doutora em Estudos Linguísticos.

Área de Concentração: Linguística Aplicada

Orientador(a): Profa. Dra. Paula Tavares Pinto

Coorientador(a): Prof. Dr. Eduardo Batista da Silva

São José do Rio Preto - SP

2025

C837e

Costa, Jeane Cardoso

Estudo da variação nas métricas do Nilc-Metrix em redações nota mil do Enem : uma abordagem baseada no processamento de linguagem natural / Jeane Cardoso Costa. -- São José do Rio Preto, 2025

210 p. : tabs., fotos, mapas

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientadora: Paula Tavares Pinto

Coorientador: Eduardo Batista da Silva

1. Linguística de Corpus. 2. Linguística Processamento de dados. 3. Redação do Enem. 4. Ferramenta de análise textual. 5. Ensino de Escrita. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

A presente pesquisa possui um impacto significativo e multifacetado, abrangendo aspectos científicos, técnicos, sociais, inovadores, econômicos, educacionais e culturais. No campo científico, contribui para o avanço do entendimento sobre as métricas textuais e sua relação com práticas de escrita bem-sucedidas, fortalecendo a interseção entre linguística e processamento de linguagem natural (PLN).

Do ponto de vista técnico e inovador, a aplicação de ferramentas como o NILC-Metrix representa um aprimoramento no uso de tecnologias avançadas para análise textual, podendo ser integrado em plataformas educacionais automatizadas para avaliação e desenvolvimento de habilidades de escrita.

No âmbito social, a pesquisa promove a valorização da escrita como competência fundamental, impactando positivamente na formação de cidadãos críticos e reflexivos. A internacionalização é incentivada ao explorar métodos e métricas com potencial de aplicação em diferentes idiomas e contextos culturais.

Regionalmente, a pesquisa reforça o papel da academia no desenvolvimento de soluções para desafios locais, ao passo que, em nível nacional, contribui para o fortalecimento da educação pública e privada por meio de insights sobre as características de redações bem-sucedidas no ENEM. A nível econômico, as ferramentas desenvolvidas ou aprimoradas podem gerar soluções escaláveis e comercialmente viáveis para o setor educacional.

No contexto educacional, os resultados obtidos podem ser usados para orientar professores, alunos e instituições sobre práticas textuais mais eficazes, promovendo melhorias no ensino de redação e linguagem. Culturalmente, o estudo reforça a importância da redação como forma de expressão e como reflexo da diversidade linguística brasileira.

Por fim, os impactos ambientais associados à digitalização devem ser avaliados com cautela. Embora o uso de plataformas tecnológicas possa reduzir a demanda por materiais impressos, como o papel, cuja taxa de reciclagem no Brasil alcançou cerca de 85% em 2022, segundo dados da Associação Nacional de Aparistas de Papel (ANAP), o consumo energético e o descarte de resíduos eletrônicos configuram novos desafios ambientais. Assim, a sustentabilidade nesse contexto é relativa e depende de políticas integradas de uso consciente de tecnologias. Ainda assim, a temática da

pesquisa, ao se debruçar sobre textos de alto desempenho, contribui com subsídios teóricos e aplicáveis para o aprimoramento das competências linguísticas exigidas na sociedade contemporânea.

POTENTIAL IMPACT OF THIS RESEARCH

The present research has a significant and multifaceted impact, encompassing scientific, technical, social, innovative, economic, educational, and cultural aspects. In the scientific field, it contributes to advancing the understanding of textual metrics and their relationship with successful writing practices, strengthening the intersection between linguistics and natural language processing (NLP).

From a technical and innovative perspective, the application of tools such as NILC-Metrix represents an enhancement in the use of advanced technologies for textual analysis, with potential integration into automated educational platforms for assessing and developing writing skills.

On a social level, the research promotes the appreciation of writing as a fundamental competence, positively impacting the formation of critical and reflective citizens. Internationalization is encouraged by exploring methods and metrics with potential applications in different languages and cultural contexts.

Regionally, the research reinforces the role of academia in developing solutions to local challenges, while at the national level, it contributes to strengthening public and private education through insights into the characteristics of high-performing essays in the ENEM. Economically, the tools developed or enhanced can generate scalable and commercially viable solutions for the educational sector.

In the educational context, the results obtained can guide teachers, students, and institutions in adopting more effective textual practices, driving improvements in the teaching of writing and language. Culturally, the study underscores the importance of writing as a means of expression and as a reflection of Brazil's linguistic diversity.

Finally, the environmental impacts associated with digitalization must be assessed with caution. Although the use of technological platforms may reduce the demand for printed materials—such as paper, whose recycling rate in Brazil reached approximately 85% in 2022, according to data from the National Association of Paper Scrap Dealers (ANAP)—energy consumption and the disposal of electronic waste pose new environmental challenges. Thus, sustainability in this context is relative and depends on integrated policies for the conscious use of technology. Nevertheless, the theme of this research, by focusing on high-performing texts, provides theoretical and practical contributions to the development of linguistic competencies essential in contemporary society.

JEANE CARDOSO COSTA

ESTUDO DA VARIAÇÃO NAS MÉTRICAS DO NILC-METRIX EM REDAÇÕES NOTA MIL DO ENEM: uma abordagem baseada no processamento de linguagem natural

Tese apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Doutor(a) em Estudos Linguísticos.

Área de Concentração: Linguística Aplicada

Data da defesa: 24/06/2025

Banca Examinadora:

_____

Profa. Dra. Paula Tavares Pinto
UNESP/Câmpus de São José do Rio Preto

_____

Prof. Dr. Eduardo Batista da Silva
Universidade Estadual de Goiás

_____

Prof. Dr. Anderson R. Ávila
Universidade de Quebec - Canadá

_____

Profa. Dra. Maria José Borcony Finatto
Universidade Federal do Rio Grande do Sul

_____

Prof. Dr. Celso Fernando Rocha
UNESP/Câmpus de São José do Rio Preto

_____

Profa. Dra. Anise de Abreu Gonçalves D'Orange Ferreira
UNESP/Câmpus de Araraquara

_____

Dedico este trabalho a Deus, que por Sua graça, me alcançou.

AGRADECIMENTOS

A Deus, que esteve comigo em todos os momentos desta escrita, capacitando-me e me dando a força necessária para continuar.

À professora Dra. Paula Tavares Pinto, minha querida orientadora, que com sua maestria e singularidade especial, conduziu-me no doutorado e me ensinou para a vida. Aprendi a aprender, contigo! Obrigada por acreditar em mim!

Ao professor Dr. Eduardo Batista da Silva, meu coorientador, que esteve comigo no mestrado e no doutorado. Obrigada por seus ricos ensinamentos!

Ao meu namorado, futuro marido, Anderson Garcia. Obrigada, por ter chegado, meu amor, meu amigo; por tornar os meus dias mais leves e felizes, por me mostrar o quanto a vida é boa ao seu lado. Obrigada pelo auxílio com algumas questões de estatística e por toda a sua paciência comigo nos momentos de cansaço e de sobrecarga. Obrigada pelo seu amor, cuidado e proteção! Eu te amo!

À minha mãe, Joane, que nunca desistiu de mim; incansavelmente, esforçou-se para me ajudar a chegar até aqui e para que eu aprendesse a como viver, a como ser mulher. Obrigada por todo o seu amor, dedicação, sabedoria e companheirismo!

Ao meu pai, Luis Carlos, que nunca mediu esforços para a minha criação. Obrigada por fazer parte da minha vida!

À minha amada irmã, Jordane, que é a minha melhor amiga, minha conselheira. Obrigada por existir, por colorir os meus dias, por alegrar o meu coração apenas com a sua presença e com o seu sorriso.

Ao meu cunhado, Sérgio Felipe, que é também um grande amigo, um verdadeiro irmão.

Ao meu sobrinho, Jordan Felipe, o presente mais belo que Deus poderia nos dar. Jordan, a “tita” te ama mais do que o Universo. Você é o meu raio de luz!

Aos meus queridos cachorrinhos (Sansão, Sol, Salomé, Rubi e Jade). Às gatinhas “Holly” e “Léo”. Um agradecimento especial ao “Feijão”, que partiu há poucos dias, deixando muitas saudades no meu coração. Jamais te esquecerei!

Aos meus queridos colegas e amigos, que fazem parte da minha vida e que, de algum modo, torcem pelo meu sucesso.

À professora Cacilda *(in memorian),* que de um modo sensível e confiante, acreditou no meu futuro, disse-me isso, e me fez acreditar também.

Aos queridos colegas pesquisadores do Grupo de Pesquisa En-Corpora/Unesp. Como eu aprendi com vocês! Obrigada por me receberem tão bem e por compartilharem Ciência.

Aos ilustres professores, Sidney Leal e Maria Claudia Delfino, que compuseram a banca de avaliação da minha tese, na qualificação, contribuindo de forma imensurável para o meu aperfeiçoamento e acurácia, enquanto pesquisadora. Senti-me honrada!

À estimada banca examinadora composta pelo Prof. Dr. Anderson R. Ávila (Universidade de Quebec – Canadá), Profa. Dra. Maria José Bocorny Finatto (UFRGS), Prof. Dr. Celso Fernando Rocha (UNESP/São José do Rio Preto) e Profa. Dra. Anise de Abreu Gonçalves D'Orange Ferreira (UNESP/Araraquara). Agradeço a vocês com muito respeito, afeto e admiração. Foi um grande diferencial para o meu trabalho!

Ao Programa de Pós-Graduação em Estudos Linguísticos da Unesp de São José do Rio Preto; sou grata pelas aulas que assisti, pelos professores que encontrei, pelos colegas que conheci. Um agradecimento especial aos profissionais que trabalham em nossa secretaria, sempre dispostos a auxiliar em todos os processos burocráticos que também fazem parte do "fazer pesquisa".

À Universidade Estadual de Goiás, na qual me tornei mestre e pude exercer a docência. Neste mesmo momento, retorno à UEG como docente efetiva, o que me deixa muito feliz e realizada. Obrigada, UEG, por fazer parte da minha história.

À Capes, que investiu em minha formação e financiou a minha pesquisa. Sou eternamente grata. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – PROEX (Programa de Excelência Acadêmica) – Processo nº 88887.690313/2022-00.

“O nosso ir faz o caminho.”
(Lewis, 1956, p. 42, tradução minha).

RESUMO

A presente tese, intitulada "Estudo da variação nas métricas do NILC-Metrix em redações nota Mil do ENEM: uma abordagem baseada no processamento de linguagem natural", tem como objetivo principal analisar as métricas textuais que influenciam a qualidade das redações avaliadas com nota máxima no Exame Nacional do Ensino Médio (ENEM) entre os anos de 2014 e 2023. A pesquisa utiliza a Linguística de *Corpus* (doravante LC) como base metodológica e teórica, ancorada no Processamento de Linguagem Natural, para explorar padrões linguístico-textuais em textos autênticos de alto desempenho. Configuram parte do aporte teórico deste estudo, Rocco, 1981; Genouvrier; Peytard, 1985; Halliday, 1993; Biderman, 2003; Sinclair, 1991; Berber Sardinha, 2004, 2010; Antunes, 2012; Guerra; Andrade, 2012; Reppen, 2015. Embora o ENEM possua critérios bem definidos para a avaliação textual, a compreensão detalhada sobre quais características linguísticas contribuem significativamente para a nota máxima ainda é insuficiente. Com isso, a pesquisa busca responder às seguintes perguntas: 1. Nas redações nota mil do Enem, existe variação entre os valores mínimos e máximos nas métricas fornecidas pelo NILC-Metrix? 2. Quais métricas do NILC-Metrix mostram maior consistência ou variação nas redações nota mil do Enem? 3. As métricas do NILC-Metrix nas redações nota mil se alinham às competências avaliativas do Enem? A análise baseia-se no uso da ferramenta NILC-METRIX (NILC-USP, 2007), que atribui métricas de coesão, de coerência e de complexidade textual. A metodologia inclui a análise quantitativa (baseada em alguns cálculos estatísticos) e qualitativa de um *corpus* composto por 279 redações nota mil, processadas para a identificação de aspectos léxico-sintáticos e morfológicos presentes nos textos. Os resultados revelaram que determinadas variações de métricas textuais, como, densidade lexical, complexidade sintática, informações semânticas e morfossintáticas, medidas psicolinguísticas, conectivos, dentre outros, constituem padrões recorrentes em redações nota mil, destacando-se como influentes e alinhados aos critérios de avaliação do ENEM. Além disso, a pesquisa sugere que a integração de tecnologias como o NILC-METRIX no contexto educacional pode apoiar a formação de práticas pedagógicas mais eficazes, fornecendo subsídios para o ensino da escrita e da avaliação textual. Por fim, o estudo reforça a importância do uso de métodos automatizados no avanço da análise linguística e no aprimoramento das práticas de avaliação educacional no Brasil.

Palavras-chave: linguística de *corpus*; processamento de linguagem natural; NILC-Metrix; redações.

ABSTRACT

The present thesis, entitled "A Study on Variation in NILC-Metrix Metrics in ENEM Top-Scoring Essays: A Natural Language Processing-Based Approach," aims to analyze the textual metrics that influence the quality of essays awarded the highest score in the Exame Nacional do Ensino Médio (ENEM) between 2014 and 2023. The research employs *Corpus* Linguistics (henceforth CL) as its methodological and theoretical foundation, anchored in Natural Language Processing (NLP), to explore linguistic-textual patterns in authentic high-performance texts. Rocco, 1981; Genouvrier; Peytard, 1985; Halliday, 1993; Biderman, 2003; Sinclair, 1991; Berber Sardinha, 2004, 2010; Antunes, 2012; Guerra; Andrade, 2012; Reppen, 2015, constitute part of the theoretical framework of this study. Although ENEM has well-defined criteria for textual evaluation, a detailed understanding of which linguistic features significantly contribute to achieving the highest score remains insufficient. Thus, this research seeks to answer the following questions: In ENEM top-scoring essays, is there variation between the minimum and maximum values in the metrics provided by NILC-Metrix? Which NILC-Metrix metrics show the greatest consistency or variation in ENEM top-scoring essays? Do the NILC-Metrix metrics in ENEM top-scoring essays align with the exam's evaluative competencies? The analysis is based on the use of the NILC-METRIX tool (NILC-USP, 2007), which assigns metrics related to cohesion, coherence, and textual complexity. The methodology includes quantitative analysis (based on statistical calculations) and qualitative analysis of a *corpus* comprising 279 top-scoring essays, processed to identify lexico-syntactic and morphological aspects present in the texts. The results revealed certain variations in textual metrics—such as lexical density, syntactic complexity, semantic and morphosyntactic information, psycholinguistic measures, and the use of connectives, among others—constitute recurrent patterns in top-scoring essays, standing out as influential factors in ENEM's evaluation criteria. Furthermore, the study suggests that integrating technologies like NILC-METRIX into the educational context may support the development of more effective pedagogical practices, providing valuable insights for writing instruction and textual evaluation. Finally, this research reinforces the importance of automated methods in advancing linguistic analysis and improving educational assessment practices in Brazil.

Keywords: *corpus* linguistics; natural language processing; NILC-METRIX; essays.

LISTA DE FIGURAS

Figura 1 - Percentual e estimativa populacional de leitores no Brasil (2024) 18
Figura 2 - Distribuição Regional das Inscrições Confirmadas no ENEM 2024 31
Figura 3 - Cartilha do participante (2024) 32
Figura 4 - Panorama das cinco competências avaliativas do Enem 33
Figura 5 - Esquema estrutural cobrado no texto dissertativo-argumentativo.....34
Figura 6 - Níveis de desempenho - Competência I 35
Figura 7 - Proposta de redação – Enem 2023.....36
Figura 8 - Níveis de desempenho - Competência II 39
Figura 9 - Níveis de desempenho - Competência III 41
Figura 10 - Níveis de desempenho - Competência IV 42
Figura 11 - Quesitos necessários à proposta de intervenção.....43
Figura 12 - Níveis de desempenho - Competência V 44
Figura 13 - interface do *Corpus* of Contemporary American English 71
Figura 14: Os principais tipos de corpora (Berber Sardinha, 2004).....72
Figura 15 - Demonstração de subtokens.....81
Figura 16 - Modelo de entrada nos dicionários 85
Figura 17 - Divisão entre a morfologia e a morfossintaxe 88
Figura 18 - Exemplo de Anotação de PoS com *Coarse Tags* 89
Figura 19 - Tipos de aprendizado de máquina 90
Figura 20 - Tokenização.....91
Figura 21 - Exemplo de lematização 93
Figura 22 - *Deep Learning*.....94
Figura 23 - Subáreas do estudo da linguagem.....95
Figura 24 - *Rank Scale*.....96
Figura 25 – *Parsing* 96
Figura 26 - *Parser* de constituintes (árvore sintática) 99
Figura 27 - *Arced arrows* 100
Figura 28 - Processo de indentação.....100
Figura 29 - Hierarquia de constituintes.....101
Figura 30 - Sintaxe de dependência.....102
Figura 31 - Intersecção de categorias de palavras.....104
Figura 32 - Tipos de redes neurais.....114

Figura 33 - Interface do NILC-Metrix 124
Figura 34 - Projetos envolvidos na simplificação do português para a inclusão digital 126
Figura 35 - Interface do projeto PorSimples 127
Figura 36 - Fórmula do índice *Flesch Reading Ease* 128
Figura 37 - CohMetrix 129
Figura 38 - Ilustração gráfica da quantidade de redações avaliadas 131
Figura 39 - Cálculo amostral 133
Figura 40 - Redação 2/279 em formato PDF 135
Figura 41 - Texto limpo para o processamento; redação nota mil – 2014 135
Figura 42 - Bibliotecas *Python* aplicadas à pesquisa 136
Figura 43 - Etapas da Análise Estatística 137
Figura 44 - Estatísticas descritivas (recorte) 139
Figura 45 - Ilustração (teste Levene) 141
Figura 46 - Relação de métricas processadas pelo NILC-Metrix (recorte) 143
Figura 47 - Redação 50 do *corpus* (PDF) 149
Figura 48 - Métricas geradas no processamento – NILC-Metrix 150
Figura 49 - Contagem e resultado da métrica de exemplo (adj_arg_ovl) 151
Figura 50 - Exemplos de valores de variação no *corpus* 153
Figura 51 - Ilustração da análise no texto 8 (Competência I) 174
Figura 52 - Ilustração da análise no texto 48 (Competência II) 176
Figura 53 - Ilustração da análise no texto (Redação 72, 2018) 177
Figura 54 - Ilustração da análise no texto (Redação 147, 2019) 178
Figura 55 - Ilustração da análise no texto (Redação 230, 2022) 180

LISTA DE QUADROS

Quadro 1 - Formação de palavras 83
Quadro 2 - Estruturas (morfemas) 83
Quadro 3 - Tema lexical 84
Quadro 4 - Distinção entre lexema, lexia e lema 86
Quadro 5 - TTR – Type/token Ratio 87
Quadro 6 - Tarefas inseridas no pré-processamento 88
Quadro 7 - Separação entre palavras e sentenças 97
Quadro 8 - Constituintes da oração 98
Quadro 9 - Interpretações da mesma palavra 103
Quadro 10 - Etapas básicas da correção automática de redações 106
Quadro 11 - Definições de alguns gêneros textuais 107
Quadro 12 - Aspectos considerados nas correções de redações 107
Quadro 13 - Desvios mais comuns e suas categorias 110
Quadro 14 - Categorias presentes nas contagens de elementos textuais 112
Quadro 15 - Ferramentas utilizadas na extração de atributos 113
Quadro 16 - Grupos de referência e quantitativos de métricas existentes no NILC-Metrix 116
Quadro 17 - Caracterização dos grupos de referência do NILC-Metrix 124
Quadro 18 - Código de aplicação do teste de Levene 141
Quadro 19 - Métricas de variação mais significativa *(no corpus)* e seus respectivos grupos de referência 145
Quadro 20 - Visualização/caracterização das 90 métricas mais relevantes no *corpus* da pesquisa 147

LISTA DE TABELAS

Tabela 1 - Temas das redações do ENEM de 2014 a 2023.....45
Tabela 2 - Variação de pontuação atribuída aos critérios avaliativos.....108
Tabela 3 - Quantitativo de redações avaliadas com nota mil considerando a população da pesquisa.....131
Tabela 4 - Distribuição de redações coletadas por ano (*corpus* da pesquisa)..132
Tabela 5 - Medida da Riqueza lexical do *corpus*.....144
Tabela 6 - Exemplificação dos valores de algumas métricas.....151
Tabela 7 - Exemplo de resultados do teste de Levene.....153

LISTA DE GRÁFICOS

Gráfico 1 - Evolução dos níveis de alfabetismo ao longo dos anos (2001-2018).....17
Gráfico 2 - Marcos significativos do Enem (1988-2018).....29
Gráfico 3 - Desempenho médio regional – Redação do Enem (2023).....48
Gráfico 4 - Participantes do Enem por faixas proficiência e região (2023).....49
Gráfico 5 - Distribuição das redações coletadas, por ano, em % (Corpus da pesquisa) 132
Gráfico 6 - Distribuição de métricas por grupo de referência, no NILC-Metrix 146

LISTA DE ABREVIATURAS E SIGLAS

AAT	Avaliação Automática de Textos
ANAP	Associação Nacional de Aparistas de Papel
AT	Adaptação de Texto
AWL	*Academic Word List*
BNC	*British National Corpus*
CAR	Correção Automática de Redações
COCA	*Corpus of Contemporary American English*
DDL	Data-driven Learning
Enem	Exame Nacional do Ensino Médio
IFES	Instituições Federais de Ensino Superior
INAF	Indicador de Alfabetismo Funcional
GSL	*General Service List of English*
LC	Linguística de *Corpus*
LSTM	*Long Short-term Memory*
MICASE	*Michigan Corpus of Academic Spoken English*
NILC	Núcleo Interinstitucional de Linguística Computacional
PB	Português Brasileiro
PDF	Formato Portátil de Documento
POS	*Part of speech*
PLN	Processamento de Linguagem Natural
PUC-RJ.	Pontifícia Universidade Católica - Rio de Janeiro
PRO UNI	Programa Universidade para Todos
RLB	Retratos da Leitura no Brasil
RNN	Redes Neurais Recorrentes

SISU	Sistema de Seleção Unificada.
TIC	Tecnologia da Informação e da Comunicação
TRI	Teoria de Resposta ao Item
TTR	*TypeToken Ratio*
UNESP	Universidade Estadual de São Paulo
USP	Universidade de São Paulo

LISTA DE SÍMBOLOS

$$n_{ajustado} = \frac{n}{1 + \frac{n-1}{N}}$$

Cálculo Amostral

$$TTR = \frac{\text{types}}{\text{tokens}} \times 100$$

Medida de Riqueza Lexical

$$M_e = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Média Simples

$$DP = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M_A)^2}{n}}$$

Desvio-padrão

SUMÁRIO

1 INTRODUÇÃO 15
1.1 DIRECIONAMENTOS DESTA TESE 22
1.2 OBJETIVO PRIMÁRIO 23
1.3 OBJETIVOS SECUNDÁRIOS 23
1.4 PRESSUPOSTOS, QUESTÕES DE PESQUISA E HIPÓTESE 24
1.4.1 Questões de pesquisa: 26
1.4.2 Hipótese 26
2 A TESSITURA DA PESQUISA 28
2.1 O EXAME NACIONAL DO ENSINO MÉDIO (ENEM) 28
2.1.1 A redação no Enem 31
2.2 A ESCRITA DE UM TEXTO DISSERTATIVO-ARGUMENTATIVO 45
2.3 DADOS DO DESEMPENHO E DA PROFICIÊNCIA NO ENEM 47
3 FUNDAMENTAÇÃO TEÓRICA 51
3.1 TRABALHOS RELACIONADOS: LINGUÍSTICA DE *CORPUS, CORPORA* ESPECIALIZADOS E ESCRITA DE REDAÇÕES. 51
3.2 A LINGUÍSTICA DE *CORPUS* 57
3.2.1 O que é um Corpus? Abordagens e Conceitos na Linguística. 62
3.2.2 A questão da extensão 69
3.2.3 A questão da representatividade 70
3.2.4 Tipos de corpus 71
3.2.5 LEXICOLOGIA E LINGUÍSTICA DE *CORPUS* 72
3.3 PROCESSAMENTO DE LINGUAGEM NATURAL 75
3.3.1 Histórico e panorama da área do PLN 75
3.3.2 PLN e a sua relação com algumas áreas da LA 79
3.4 A MORFOLOGIA NO PLN 82
3.4.1 Etapas de processamento morfológico, morfossintático e sintático 87
3.5 LÉXICO E GRAMÁTICA NO PLN 102
3.5.1 Tipos de léxico: comum e especializado 104
3.6 CAMINHOS DA CORREÇÃO AUTOMÁTICA DE REDAÇÕES 105
3.6.1 A redação escolar e a correção automática de textos 106
3.6.2 Atribuição de notas aos textos 112

3.6.3 Atribuição de nota para Redações do Enem 115
3.7 ESTATÍSTICAS BÁSICAS DO TEXTO NA CORREÇÃO AUTOMÁTICA DE REDAÇÃO 115
3.7.1 Correção manual *versus* correção automática 118
4 MATERIAL E MÉTODO 122
4.1 A ABORDAGEM QUANTIQUALITATIVA DA PESQUISA 122
4.2 A FERRAMENTA NILC-METRIX 123
4.3 CONSTITUIÇÃO DO *CORPUS* DA PESQUISA 129
4.4 SOFTWARE ESTATÍSTICO 135
4.5 O CAMINHO METODOLÓGICO PERCORRIDO 136
4.5.1 Normalização dos dados com Z-score 137
4.5.2 Estatísticas descritivas e seleção de métricas 138
4.5.3 Teste de Levene: homogeneidade de variância 140
5 ANÁLISE DOS DADOS 143
5.1 PROCESSAMENTO DE DADOS NO NILC-METRIX E ORGANIZAÇÃO DAS MÉTRICAS OBTIDAS 144
5.1.1 Exemplo de processamento na NILC-Metrix no *Corpus* analisado 149
5.2 INTERPRETAÇÃO DAS VARIAÇÕES DAS MÉTRICAS EXTRAÍDAS DO CORPUS 154
5.2.1 Medidas descritivas 156
5.2. 2 Coesão Referencial 157
5.2.3 Coesão Semântica 158
5.2.4 Medidas Psicolinguísticas 160
5.2.5 Diversidade Lexical 161
5.2.6 Conectivos 162
5.2.7 Complexidade Sintática 163
5.2.8 Frequência de Palavras 164
5.2.9 Informações Semânticas de Palavras 166
5.2.10 Índices de Leiturabilidade 167
5.2.11 Informações Morfossintáticas de Palavras 168
5.2.12 Densidade de Padrões Sintáticos 170
5.3 ALINHAMENTO DAS MÉTRICAS OBTIDAS COM AS COMPETÊNCIAS AVALIATIVAS DO ENEM 172
5.3.1 Competência I – Domínio da norma culta da Língua Portuguesa 173

5.3.2 Competência II – Compreensão do tema, adequação ao tipo textual e o uso do repertório cultural 174
5.3.3 Competência III – Seleção, organização e interpretação de informações para sustentar um ponto de vista 176
5.3.4 Competência IV – Uso de conectivos e de operadores argumentativos 177
5.3.5 Competência V – Elaboração da proposta de intervenção.....178
6 CONSIDERAÇÕES FINAIS E ENCAMINHAMENTOS.....182
REFERÊNCIAS.....188

1 INTRODUÇÃO

> Escrever está na moda. As novas tecnologias de comunicação, quem diria, ressuscitaram o valor da escrita. Já não se escrevem cartas como antigamente, mas concisas mensagens eletrônicas. Já não se admitem relatórios longos e complexos. Tempo é dinheiro. Relatórios devem ser objetivos e contundentes. E os vestibulares? Estudante não entra na faculdade se falhar na redação. Nunca se precisou tanto da escrita quanto agora.
>
> *Dad Squarisi e Aríete Salvador (2012)*

O excerto “escrever é uma forma de existir”, retirado do livro “Escrever melhor”, de Dad Squarisi e Aríete Salvador (2012), dialoga profundamente com a minha trajetória como professora, sempre apaixonada pelo ensino de redação e pelo desenvolvimento da escrita. Na obra, os autores destacam como a escrita não é apenas uma ferramenta técnica, mas um espaço de expressão e construção de identidade. “Escrever é uma forma de existir porque é pelo texto que o autor traduz sentimentos e pensamentos, tornando-os palpáveis para os outros e para si mesmo.” Essa afirmação se entrelaça com a prática que tenho em sala de aula, especialmente, com o meu modo de refletir acerca da forma como a redação pode ser ensinada e de como os alunos podem ser bem-preparados para escrever.

Nesta tese, proponho uma análise dos critérios que têm definido a produção da redação no contexto do Enem (Exame Nacional do Ensino Médio), com especial atenção para os elementos considerados essenciais para uma "redação de sucesso".

A análise parte da minha experiência como professora de Língua Portuguesa e se estende à investigação estatística de redações nota mil do Enem, explorando as métricas indicadas por uma ferramenta computacional para compreender quais aspectos parecem ter sido determinantes nesse desempenho. Ao examinar a variação dessas métricas, busco identificar padrões que apontem não apenas as estratégias recorrentes, mas também os elementos que contribuem para elevar o potencial de escrita das redações. O gênero textual dissertativo-argumentativo (oriundo da tipologia dissertativa), cuja estrutura é estável e amplamente conhecida, configura-se como instrumento para alcançar o objetivo central dos estudantes: o ingresso no ensino superior.

Uma redação bem-sucedida neste contexto não se define por excelência criativa ou autoral, mas sim pelo atendimento rigoroso aos critérios técnicos estabelecidos pela banca examinadora. Contudo, seria possível desenvolver essa

competência textual em apenas um ano letivo? A experiência demonstra que sim - desde que a didática adote uma abordagem estratégica, com ênfase em técnicas comprovadas, modelos estruturais e uma sistematização criteriosa do processo de escrita.

O método em questão desenvolve competências específicas para os exames seletivos, treinando os estudantes na aplicação metódica de um modelo composicional que atenda às exigências avaliativas desses processos seletivos. Contudo, a preocupação com o que vem depois desse processo é fundamental. O desafio não termina com a aprovação: ele se estende à permanência desses estudantes nos cursos universitários. Apenas reproduzir o que é esperado, sem compreensão crítica ou autonomia na escrita, pode limitar o potencial de atuação acadêmica e profissional desses alunos, dificultando, em parte, seu papel como futuros pesquisadores, profissionais e cidadãos.

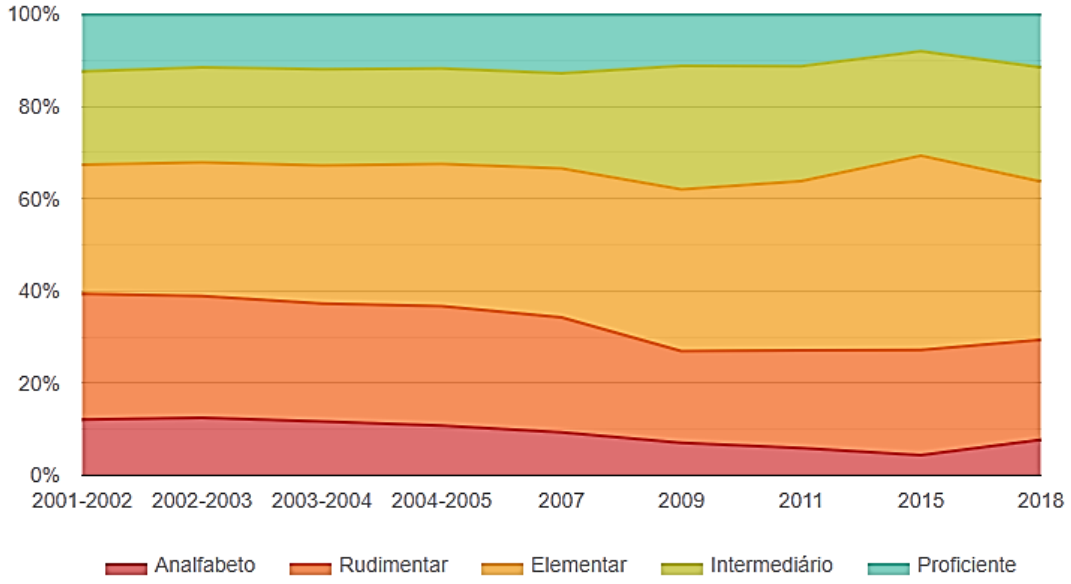
Nesse sentido, embora não seja o eixo central do meu trabalho, ao problematizar a produção de redações nos cursinhos e nas escolas de Ensino Médio, defendo que, não obstante o ensino de escrita ser direcionado por padrões preestabelecidos para os vestibulares e para o Enem, não implica total ausência de expressão ou criatividade. Ao contrário disso, nos cursinhos, muitos professores buscam promover o pensamento crítico e a livre expressão não somente nas aulas de linguagens, mas nas de ciências humanas, ciências da natureza e na matemática. O desafio está em equilibrar a preparação técnica com práticas que visem a autonomia discente e o pensamento crítico.

Indubitavelmente, existe um profundo equívoco quando discutimos acerca da escrita de textos, sem considerarmos questões que são cruciais, a saber: por que escrevemos? Para quem escrevemos? Para que escrevemos? A ênfase, muitas vezes, é dada ao formato, à estrutura e àquilo que se espera da escrita, relegando a reflexão crítica acerca de quem estabelece as normas, de qual o intuito de segui-las, de quem realmente espera alguma coisa da nossa escrita e do que espera. Na esteira desse raciocínio, consideramos que a escrita não é um ato isolado; é trabalho, é um ato de comunicação com um propósito definido, em um contexto específico. Contudo, a escrita, assim como a leitura, é um privilégio; não deveria ser. As vias de acesso à escrita e à leitura, deveriam ser universais, mas ao contrário disso, historicamente têm sido limitadas por questões estruturais, educacionais e culturais. Uma das barreiras que causa o distanciamento de muitos sujeitos da possibilidade de expressarem-se por

meio da escrita, é a propagação de concepções restritivas sobre o que significa escrever bem; o que transforma a escrita (que deveria ser um direito) em um privilégio excludente.

À luz de considerações sobre a alfabetização (aquisição de leitura e de escrita) no Brasil, menciono dados alarmantes oriundos do Indicador de Alfabetismo Funcional (INAF) e do Retratos da Leitura no Brasil (RLB). Os dados revelam que grande parte da população brasileira, inclusive a população universitária, não atinge o nível de proficiência de alfabetismo funcional. Mas a propósito, o que significa ser funcionalmente proficiente? Significa ser hábil para a compreensão, para a escrita e para a interpretação de texto em situações cotidianas, sem restrições significativas, como: ler e escrever textos mais complexos, analisar e fazer relação das partes, fazer comparações e avaliações de informações, bem como fazer a distinção entre fatos e opiniões.

Gráfico 1 - Evolução dos níveis de alfabetismo ao longo dos anos (2001-2018)



Fonte: autoria própria

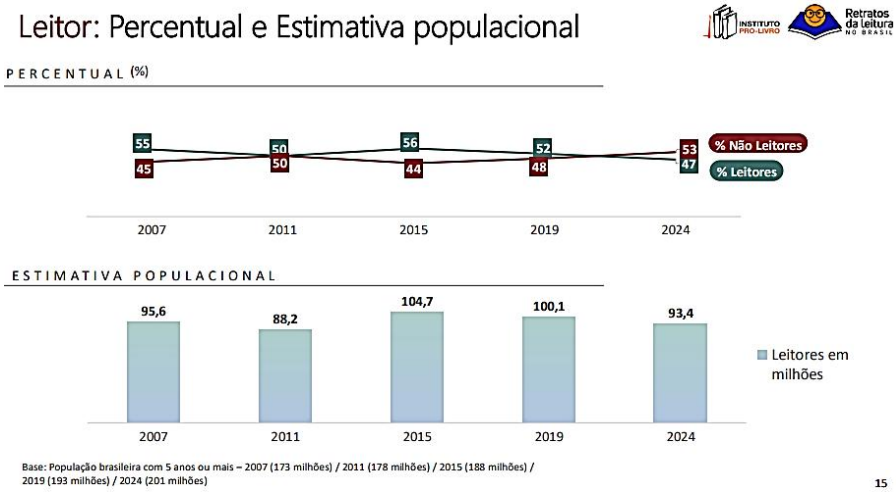
Legenda: gráfico de empilhamento com fundo colorido. As categorias de alfabetismo são representadas da seguinte forma: analfabeto (cor vermelha), rudimentar (cor alaranjada-forte), elementar (alaranjada-fraca), intermediário (cor amarela) e proficiente (cor verde-azulada). O eixo vertical traz as porcentagens e o eixo horizontal traz os anos.

A análise revela ao analisar o gráfico 01, uma redução significativa no número de analfabetos plenos na população brasileira em um período de 17 anos. De acordo com o INAF, houve uma queda de 12% (2001-2002) para 4% (2015) — apesar de

que, na edição de 2018, tenha se registrado um aumento não tão significativo dentro do limite da margem de erro, 6%. Em 2024, o percentual de analfabetos plenos foi de 7%. No decorrer desse período (2001-20024) também ocorreu uma diminuição proporcional de brasileiros classificados no nível rudimentar, isto é, indivíduos que utilizam a leitura, a escrita e as operações matemáticas de forma limítrofe em tarefas corriqueiras. O índice passou de 27% (2001-2002) para cerca de 20%, mantendo-se estável desde 2009.

Os dados supramencionados destacam que o Brasil (2018) tinha 14,5 milhões de analfabetos funcionais a menos do que teria caso essa redução não tivesse acontecido. Todavia, urge salientar que a proporção de alfabetizados proficientes permanece a mesma desde o início da série histórica, em torno de 12%, o que significa que somente cerca de 17,4 milhões dos 144,7 milhões de brasileiros (na época), [1]entre 15 e 64 anos, alcançaram esse nível de proficiência.

Figura 1 - Percentual e estimativa populacional de leitores no Brasil (2024)[2]



Fonte: Retratos da Leitura no Brasil (2024)

O gráfico apresentado na figura 1 mostra a evolução percentual e a estimativa da população de leitores no Brasil entre 2007 e 2024. Os dados revelam uma

[1] A população brasileira até 1º de julho de 2024 era de 212,6 milhões de habitantes, segundo estimativa divulgada pelo Instituto Brasileiro de Geografia e Estatística (IBGE).(Essa informação foi retirada do site do governo federal: www.gov.br).

[2] A Figura 1 foi elaborada com base nos dados divulgados pela pesquisa “Retratos da Leitura no Brasil” (2024), promovida pelo Instituto Pró-Livro em parceria com o IBOPE Inteligência. O gráfico foi construído a partir da compilação das informações percentuais e absolutas da população leitora brasileira desde 2007, com recortes nos anos em que a pesquisa foi aplicada: 2007, 2011, 2015 e 2024. Os dados originais estão disponíveis no relatório técnico da pesquisa, que apresenta os critérios para definição de “leitor”, bem como as estimativas populacionais em milhões, considerando os dados censitários do IBGE.

tendência de queda no hábito de leitura no período. No ano de 2007, 55% da população era considerada leitora, percentual que caiu para 50% em 2011, voltou a crescer para 56% em 2015, mas declinou nos anos seguintes, chegando a 47% em 2024. A estimativa populacional reflete essa mesma tendência: o número de leitores, que era de 95,6 milhões em 2007, diminuiu para 88,2 milhões em 2011, atingiu o pico de 104,7 milhões em 2015, e recuou para 93,4 milhões em 2024.

É notório que essas pesquisas expõem o cenário preocupante: enquanto sociedade, somos uma massa de não leitores e de não escritores proficientes, em um país de dimensões continentais. Há uma negação comum de acesso à escrita e à leitura. Essa negação não é apenas física ou material, está implícita na impossibilidade de compreensão de textos que garantam o exercício da plena cidadania, na escassez de uma educação de qualidade que contribua com a formação de cidadãos críticos e participativos, e, sobremaneira, no afastamento de práticas que proporcionem o pleno acesso à informação e ao conhecimento. Ter acesso à informação é ter poder; a capacidade de leitura, de interpretação e de produção de texto é transcendente ao domínio técnico e se configura como um direito crucial para a formação dos cidadãos.

A partir das considerações desses enquadramentos, busco tornar mais transparentes os processos nos quais a escrita de redações está inserida. Ademais, procuro realizar uma análise de dados linguísticos extraídos do desempenho escrito de candidatos do Enem. Sendo assim, descrevo e analiso nesta tese, um *corpus* de 279 redações nota mil do Enem que foram escritas entre os anos de 2014 e 2023. Trata-se de um *corpus* especializado que constitui o núcleo da pesquisa a partir do qual extraio dados e resultados que buscam elucidar aspectos preponderantes para o ensino da escrita, tanto no que se relaciona à preparação de estudantes para a obtenção de êxito no exame, quanto na promoção de uma reflexão crítica acerca dos processos que condicionaram a produção textual nesse contexto (encaminhamentos futuros).

A descrição e a análise das redações que constituem o *corpus* têm como objetivo mapear os padrões linguísticos que emergem nos textos de melhor desempenho. A partir dessa análise e descrição, procuro identificar marcas lexicais, gramaticais e sintáticos que ajudem a compreender como as práticas de ensino de escrita influenciam a produção textual dos alunos, o que poderá fornecer subsídios para o aprimoramento pedagógico nesse campo.

Como mencionado, o objeto de estudo deste trabalho é um *corpus* especializado. Em meu fazer de pesquisa, debruço-me sobre a Linguística de *Corpus* como metodologia e como aporte teórico, para analisar padrões de uso real da língua em conjuntos de textos autênticos, por meio de observações empíricas. A finalidade por mim buscada, detém-se no apontamento e na análise de aspectos léxico-sintáticos e gramaticais, além de estruturas possíveis e prováveis de serem acionadas com maior frequência pelos escritores do português em condições reais de produções específicas.

No capítulo um, elenco considerações acerca dos direcionamentos desta tese norteadores do meu percurso investigativo. De modo incipiente, delimito os objetivos da pesquisa, tanto o primário quanto os secundários. Neste mesmo capítulo, apresento as questões de pesquisa, que representam as indagações centrais que busco responder com base na literatura revisada, no *corpus* selecionado e nas ferramentas analíticas utilizadas. Em consonância com essas questões, apresento também a hipótese definida.

No capítulo dois, proponho um delineamento do contexto no qual se insere a pesquisa, situando o leitor quanto aos fundamentos institucionais e educacionais que justificam a escolha do objeto de estudo. Para isso, faço considerações gerais acerca do Exame Nacional do Ensino médio (Enem), destacando a sua relevância como política pública de avaliação em larga escala. No ensejo, destaco questões específicas no que diz respeito à redação exigida no Exame; ademais, abordo alguns dos desafios inseridos na escrita de um texto dissertativo-argumentativo.

No capítulo três, faço menção a importantes pesquisas envolvendo a linguística de *corpus*, o processamento de linguagem natural, os estudos linguísticos e a escrita de redações. Descrevo as teorias e conceitos que fundamentam as premissas e que orientam este estudo. Inicialmente, estabeleço discussões acerca da Linguística de *Corpus*; em seguida, abordo conceitos relacionados à Lexicologia e ao Processamento de Linguagem Natural com ênfase nos Estudos Linguísticos. Na sequência, faço considerações acerca da ferramenta de PLN utilizada para a análise e para a descrição textual, a NILC-Metrix. De modo pertinente, lanço luz ao estado da arte no que concerne à correção automática de redações e como elas têm sido apresentados por meio do PLN.

No que se refere à linguística de *corpus*, o capítulo a aborda como um campo concentrado na análise de grandes coleções de textos autênticos, os *corpora*,

objetivando a identificação de padrões linguísticos e de características da língua em uso. Autores como Biber, Conrad e Sinclair (1998) e Berber Sardinha (2004), são referências fundamentais ao descreverem como a análise baseada em *corpora* revela padrões de uso da linguagem que dificilmente seriam perceptíveis sem um exame sistemático e automatizado. A análise de *corpora*, como abordada por esses autores, é indispensável para a minha pesquisa, pois proporciona descobertas tanto quantitativas — sobre frequência de palavras, estruturas sintáticas, gramaticais e lexicais — quanto qualitativas, como a interpretação do uso linguístico em diferentes contextos específicos de produção. A aplicação dessa abordagem à análise de redações oferece um campo profícuo para estudarmos sobre quais aspectos influenciam a qualidade das produções escritas, especialmente no contexto de avaliações educacionais, como o Enem, o que me propus a pesquisar.

O capítulo, como dito, aborda ainda as ferramentas de Processamento de Linguagem Natural (PLN), uma área de estudo que tem como objetivo permitir que máquinas compreendam e processem a linguagem humana. No contexto desta pesquisa, a utilização de ferramentas automatizadas para a análise de textos, como a NILC-METRIX, tornou-se crucial. Ela é uma ferramenta desenvolvida especificamente para descrever, por meio de métricas, aspectos como coesão, coerência e complexidade textual, o que a torna uma aliada poderosa para a análise automatizada de redações. Suas funcionalidades incluem a análise de diferentes estruturas linguísticas; possui, além disso, capacidade de processar e de analisar o texto de forma objetiva, baseando-se em padrões estabelecidos pela linguística computacional. No decorrer do capítulo, a ferramenta será examinada com maior profundidade.

Em continuidade, o capítulo quatro, dedica-se à exposição minuciosa dos procedimentos metodológicos que orientaram a realização desta pesquisa. Nesse capítulo serão apresentados e justificados a natureza da abordagem adotada, o tipo de pesquisa desenvolvido, o delineamento metodológico escolhido, bem como os instrumentos e técnicas utilizados para a coleta e análise dos dados.

Na sequência, o capítulo cinco, apresenta os resultados e a análise dos dados obtidos acerca do estudo da variação das métricas do NILC-Metrix, com base no *corpus* de redações; os achados serão mencionados à luz das competências avaliativas do Enem, examinando como as métricas linguísticas automatizadas estão articuladas com os critérios de correção preestabelecidos. O capítulo será, portanto,

o núcleo empírico desta pesquisa, em consonância com os objetivos traçados e com as questões de pesquisa previamente formuladas.

No capítulo seis, dedicado às considerações finais, retomo os principais pontos discutidos ao longo da pesquisa, sintetizando os eixos teóricos, metodológicos e empíricos delineados nos capítulos anteriores; a ênfase será dada aos resultados e evidências coletados, a fim de validarmos efetivamente a análise realizada e as contribuições do estudo.

1.1 DIRECIONAMENTOS DESTA TESE

A análise descrita nesta tese responde a questões relevantes relacionadas ao ensino de redação sob a perspectiva de métricas textuais, o que me parece útil no que tange à reflexão acerca dos desafios de escrita enfrentados por estudantes do Ensino Médio, que farão a Redação do Enem. Esses desafios estão associados a uma gama de fatores, entre eles, o ensino instrumentalizado de fórmulas e de macetes para atenderem às demandas de avaliações. Conforme evidenciado na introdução deste trabalho, pesquisas como o Indicador de Alfabetismo Funcional (INAF) e o Retratos da Leitura no Brasil (RLB), revelam que grande parte da população brasileira não atinge níveis de alfabetismo funcional que lhes permitam compreender e produzir textos complexos, o que pode incluir, redações que atendam exigências acadêmicas.

A partir da análise dos dados oriundos de um *corpus* de 279 redações nota mil do Enem, e da identificação de padrões linguísticos que indiquem o que mais contribuiu para os desempenhos de destaque, esta tese poderá contribuir com os professores de Língua Portuguesa, especialmente do ensino de redação, oferecendo subsídios concretos para compreenderem os critérios que definem a produção de redações avaliativas e para aprimorarem as suas práticas pedagógicas, de modo a equilibrarem o ensino técnico necessário para a aprovação nas provas e a prática pedagógica intencional e reflexiva.

Ao longo desta tese, argumento que existem aspectos linguísticos fundamentais que são amplamente valorizados pelos avaliadores de redações, aspectos esses identificados e analisados quantitativamente. É sobre essas evidências que minha pesquisa se concentra, buscando trazer clareza sobre os elementos indispensáveis no ensino da escrita de redação. Ainda que o pensamento predominante esteja centrado nos resultados alcançados pelos estudantes, esta

investigação propõe um olhar mais profundo sobre as competências que realmente sustentam a produção textual de qualidade e como elas se relacionam ou estão sustentadas naquilo que as redações analisadas realmente demonstram linguisticamente.

1.2 OBJETIVO PRIMÁRIO

Esta tese tem como objetivo central investigar os padrões linguístico-textuais em redações nota mil **do Enem, utilizando a abordagem da Linguística de *Corpus* e ferramentas de Processamento de Linguagem Natural (PLN), como o NILC-Metrix,** para identificar elementos linguísticos que caracterizam produções de alto desempenho. Ao longo da pesquisa, pretendo destacar como escolhas lexicais e estruturais dialogam com as competências exigidas pelo exame, abrindo caminhos para reflexões pedagógicas que possam aprimorar o ensino de redação.

O foco da análise recai sobre as palavras que constroem os textos e sobre as configurações lexicais, sintáticas e gramaticais que emergem dessas produções. A pesquisa busca compreender em que medida as escolhas linguísticas refletem padrões de ensino baseados na reprodução de modelos preestabelecidos. Para isso, as redações serão analisadas em sua diversidade lexical, indicadores sintáticos, de coesão, de coerência e de complexidade textual, conforme métricas fornecidas pelo NILC-Metrix.

Além disso, um dos propósitos da pesquisa é fornecer subsídios concretos e verificados para professores de Língua Portuguesa, especialmente aqueles que atuam no ensino de redação, de modo que possam reavaliar práticas pedagógicas predominantes e direcioná-las intencionalmente ao que precisa ser ensinado, sem que a capacidade crítica e/ou a autoria sejam relegadas. O objetivo é propor um equilíbrio entre a preparação técnica para atender às exigências avaliativas do Enem e a promoção de uma escrita eficaz.

1.3 OBJETIVOS SECUNDÁRIOS

- Contribuir com a discussão do ensino de escrita de redações em Língua Portuguesa;

- Disponibilizar um *corpus* descrito e sistematizado que sirva como banco de dados linguísticos;
- Mapear, por meio de evidências, padrões linguísticos recorrentes nas redações nota mil.
- Estabelecer uma relação direta entre os dados obtidos da análise textual e as competências do Enem.
- Traçar encaminhamentos futuros que redirecionem algumas práticas pedagógicas de ensino de escrita.

O banco de dados gerado constitui um recurso essencial para investigações linguísticas sobre textos, servindo como referência para estudos sobre o ensino de escrita e como base para o desenvolvimento de ferramentas computacionais voltadas ao apoio à produção textual e à análise automatizada de textos. Todos os dados coletados e o *corpus* da pesquisa ficarão disponíveis publicamente.

1.4 PRESSUPOSTOS, QUESTÕES DE PESQUISA E HIPÓTESE

Esta tese, como mencionado anteriormente, está ancorada de modo essencial na abordagem da linguística de *corpus*, com ênfase na análise de componentes textuais de redações nota mil do Enem. Nesse contexto, a Linguística de *Corpus* é abordada tanto como referencial teórico quanto como metodologia, sustentando o tratamento dos textos e a sua análise estatística. Com base no objeto de estudo e nos objetivos traçados, os pressupostos teóricos e metodológicos desta tese são organizados nos seguintes eixos principais:

1) A Linguística de *Corpus* surge como uma nova perspectiva para a Linguística: para aqueles que se aproximam da linguística de *corpus*, ela se apresenta não só como uma metodologia, mas como uma abordagem de aprofundamento nos estudos de linguagem. Se há quem queira utilizar da Linguística de *Corpus* por conta de seu instrumental ou de seus procedimentos, não lhe será cobrado nenhuma filiação teórica ou epistemológica, embora defendamos que a LC tenha o seu modo próprio de compreender a língua. A Linguística de *Corpus* é central nesta tese, tanto como referencial teórico quanto como metodologia de análise. Ela possibilita uma abordagem empírica e probabilística da língua, permitindo identificar

padrões linguístico-textuais que caracterizam redações de alto desempenho. Por meio da análise de *corpora*, emergem fenômenos linguísticos que refletem escolhas estratégicas dos autores/escritores/candidatos, oferecendo uma base sólida para compreender como os textos dialogam com as competências avaliativas do Enem.

2) A língua como sistema probabilístico: a Linguística de *Corpus* define o seu objeto de estudo, que é a língua, como um sistema probabilístico de combinatórias, no qual uma unidade é definida pelas associações que faz com outras. Embora os *corpora* não nos tragam respostas para tudo, traz-nos questões novas, às quais não imaginávamos encontrar. “Um *corpus* não é mera ferramenta de análise, mas um relevante conceito teórico” (Sinclair, 1991, p. 18).

3) O Processamento de Linguagem Natural é uma área da computação que objetiva extrair representações e significados completos oriundos de textos de linguagem natural (Indurkhya and Damerau, 2010, p. 36): a linguagem natural é utilizada para comunicações cotidianas feitas por seres humanos; de modo contrário à linguagem artificial, com linguagens de programação e as matemáticas, as linguagens naturais evoluem à medida que passam de uma geração para outra, e são complexas de serem definidas com regras explícitas. Nesse sentido, o PLN pode ser definido como uma maneira de descobrir quem fez algo, a quem fez, quando fez, onde, como e por quê. Para tanto, o PLN geralmente faz uso de conceitos linguísticos, como por exemplo, as classes de palavras.

4) A relação entre competências avaliativas e padrões textuais: as competências avaliativas estabelecidas pelo Enem orientam as produções textuais dos estudantes e moldam suas escolhas lexicais e estruturais. Pensamos em uma escrita avaliativa como critério de acesso. Isto é, o ensino de redação, no contexto dos vestibulares, tornou-se uma prática fortemente influenciada pelas demandas de conformidade aos critérios avaliativos. Frente ao exposto, a escrita pode ser entendida como um ato de conformidade em que os estudantes necessitam alinhar aquilo que escrevem às expectativas das bancas avaliadoras. Todavia, essa submissão à avaliação não pode nem deve ser passiva; a partir de intervenções pedagógicas adequadas e fundamentadas é possível transformar o ensino da escrita em um espaço de reflexão crítica promovendo a apropriação consciente de critérios avaliativos não como ferramenta de alienação, mas de ascensão social.

Partindo das premissas anteriores, esta tese busca, essencialmente, compreender como os padrões linguísticos presentes nas redações nota mil refletem

essas competências (do Enem) e, ao mesmo tempo, como esses critérios podem ser interpretados pedagogicamente. A descrição e análise de métricas fornecidas pela NILC-Metrix no *corpus* (foco principal do trabalho) auxilia na identificação de elementos recorrentes, podendo culminar no desenvolvimento de estratégias que aprimorem o ensino de redação, alinhando-o às demandas avaliativas, e num ensino intencionalmente crítico.

Para isso, as questões de pesquisa e hipóteses a seguir delineiam o caminho investigativo.

1.4.1 Questões de pesquisa:

1. Nas redações nota mil do Enem, existe variação entre os valores mínimos e máximos nas métricas fornecidas pelo NILC-Metrix?
2. Quais métricas do NILC-Metrix mostram maior consistência ou variação nas redações nota mil do Enem?
3. As métricas do NILC-Metrix nas redações nota mil se alinham às competências avaliativas do Enem?

1.4.2 Hipótese

Esta pesquisa busca verificar, a partir de uma descrição e da análise do *corpus* de redações nota mil do Enem, a validade da seguinte hipótese de investigação: **as redações que compõem o *corpus* analisado oferecem pistas confiáveis sobre as características linguístico-textuais que distinguem produções de alto desempenho no exame.** Essas características, pautadas no estudo da variação de métricas textuais, indicam uma conformidade com modelos de escrita estáveis, frequentemente reproduzidos para atender às exigências das competências avaliativas do Enem. Embora a adequação a esses modelos favoreça o desempenho máximo na redação, ela não garante, por si só, uma proficiência acadêmica ou a capacidade de produzir textos que atendam às exigências de um ambiente universitário.

Nesse sentido, o ensino de redação voltado ao Enem deve não apenas instrumentalizar os estudantes para alcançarem os critérios avaliativos, mas também explicitar os processos de padronização e questioná-los de forma crítica ao longo do

aprendizado. Aspectos como o tamanho do texto, a riqueza lexical - *Type Token Ratio*- (medida lexical utilizada no estudo de Costa e Silva (2020) - que será mencionado na seção 3.3), o uso de conectivos e a organização argumentativa aparecem como elementos que, embora sejam necessários para alcançar uma boa nota no exame, não são, sozinhos, suficientes para atender às demandas da produção textual em outros contextos acadêmicos e sociais. Esses aspectos, portanto, devem ser trabalhados de forma que os estudantes, como cidadãos que são inseridos em diferentes situações comunicativas, desenvolvam autonomia discursiva e um repertório crítico para além do que o exame exige.

2 A TESSITURA DA PESQUISA

Neste capítulo, apresentarei o panorama que fundamenta esta pesquisa. Iniciarei com a descrição das competências avaliadas no ENEM e do impacto de seus critérios de correção na produção textual; a ênfase será dada ao que se espera de uma redação nota mil, por parte dos avaliadores. Essa descrição será feita com base na matriz de correção publicada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[3]. Trarei luz aos critérios que orientam os avaliadores e como eles se relacionam com as exigências do gênero dissertativo-argumentativo. Costurarei essas questões com pesquisas relacionadas a esta tese, esclarecendo conceitos e perspectivas teóricas relevantes para compreender a análise que empreendi por meio do estudo das redações que constituem o *corpus*.

2.1 O EXAME NACIONAL DO ENSINO MÉDIO (ENEM)

Esta subseção trará algumas informações gerais acerca do Enem.

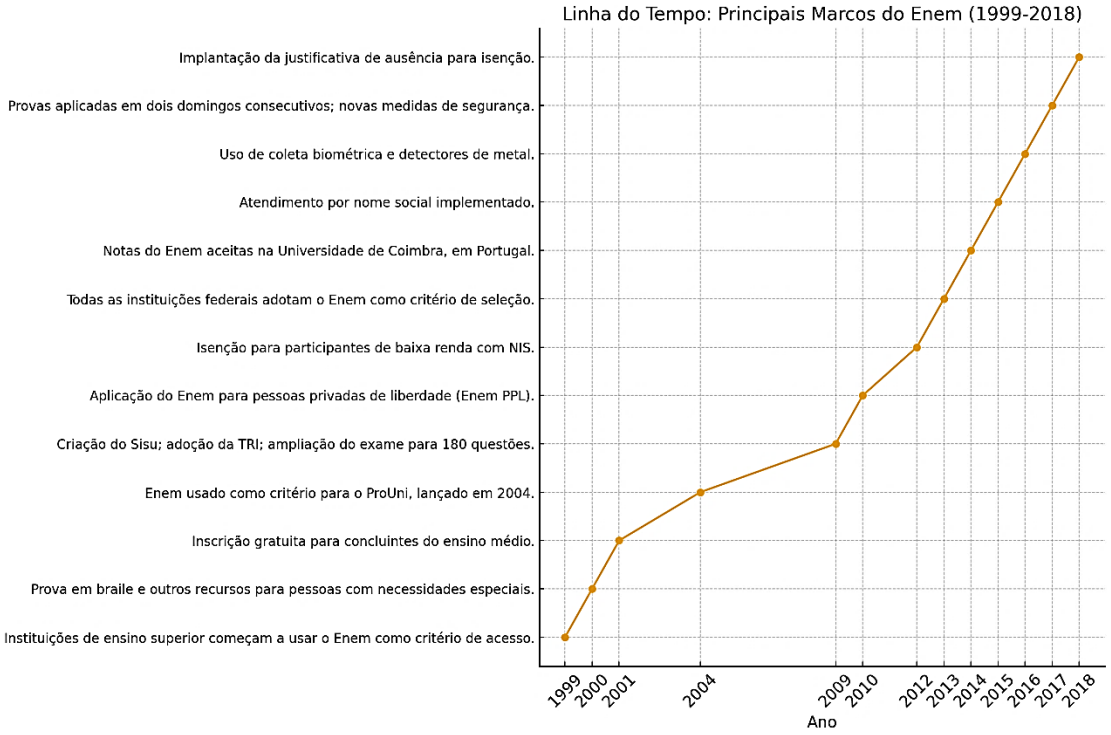
O Exame Nacional do Ensino Médio (ENEM) foi criado pelo governo federal, em 1998, com o objetivo inicial de avaliar o desempenho dos estudantes ao término da educação básica. Durante mais de dez anos, o ENEM foi utilizado unicamente como uma ferramenta de avaliação, medindo habilidades e competências adquiridas pelos concluintes do Ensino Médio, sem a finalidade de selecionar candidatos para o ensino superior. Nesse período, a inserção nas universidades brasileiras era realizada por meio dos tradicionais vestibulares, elaborados de forma descentralizada por equipes locais em cada instituição. Essa heterogeneidade resultava em processos seletivos com características distintas, contribuindo para a diversidade cultural e formativa dos ingressantes.

Todavia, a partir de 2009, o cenário foi transformado. Algumas medidas governamentais ampliaram a funcionalidade do ENEM, que passou a ser utilizado como um dos principais mecanismos de acesso ao ensino superior; o Sistema de

[3] O INEP é responsável por elaborar todo o planejamento do Enem, incluindo o formato da prova, conteúdo a ser avaliado, cronograma de aplicação e outros aspectos. O instituto gerencia o processo de inscrição dos participantes, assim como toda a logística relacionada à aplicação do exame.

Seleção Unificada (Sisu)[4] foi implementado, nesse cenário, centralizando o processo de alocação de vagas em Instituições Federais de Ensino Superior (IFES) em todo o país. A adoção do ENEM trouxe consigo vantagens significativas, por exemplo, a mobilidade acadêmica foi ampliada, permitindo que estudantes de diferentes regiões do Brasil tivessem a oportunidade de acessar instituições de ensino superior fora de seus estados de origem.

Gráfico 2 - Marcos significativos do Enem (1988-2018)



Fonte: Inep (2024)

Legenda: O gráfico apresenta a evolução histórica do Enem desde 1998 a 2018. Destacam-se mudanças relevantes, dentre elas, a adoção do exame por instituições federais, a criação do Sisu, a ampliação da prova e a implementação de medidas estratégicas, tanto de segurança quanto de inclusão.

Percebemos na linha do tempo do gráfico dois, um detalhamento dos principais marcos históricos do Exame Nacional do Ensino Médio (Enem) entre os anos de 1999 e 2018, destacando o desenvolvimento e as transformações ao longo de duas décadas. Cada ponto no gráfico representa um evento significativo que marcou a evolução do exame e que trouxe consolidação do seu papel como uma das principais ferramentas de avaliação e de acesso ao ensino superior no Brasil.

[4] A seleção é feita com base na média das notas obtidas no ENEM, respeitando o limite de vagas disponíveis para cada curso e modalidade de concorrência.

Em 1999 foi o marco inicial: a adoção do Enem como critério de acesso por instituições de ensino superior, começando com a PUC-RJ (Pontifícia Universidade Católica do Rio de Janeiro) e com a Universidade Federal de Ouro Preto. A partir de então, o exame passou por reajustes sucessivos, como: a introdução de provas adaptadas para pessoas com deficiência, em 2000; a política de inscrição gratuita para concluintes do ensino médio em 2001; e o uso do resultado obtido no exame, no Programa Universidade para Todos (ProUni) [5], a partir de 2004.

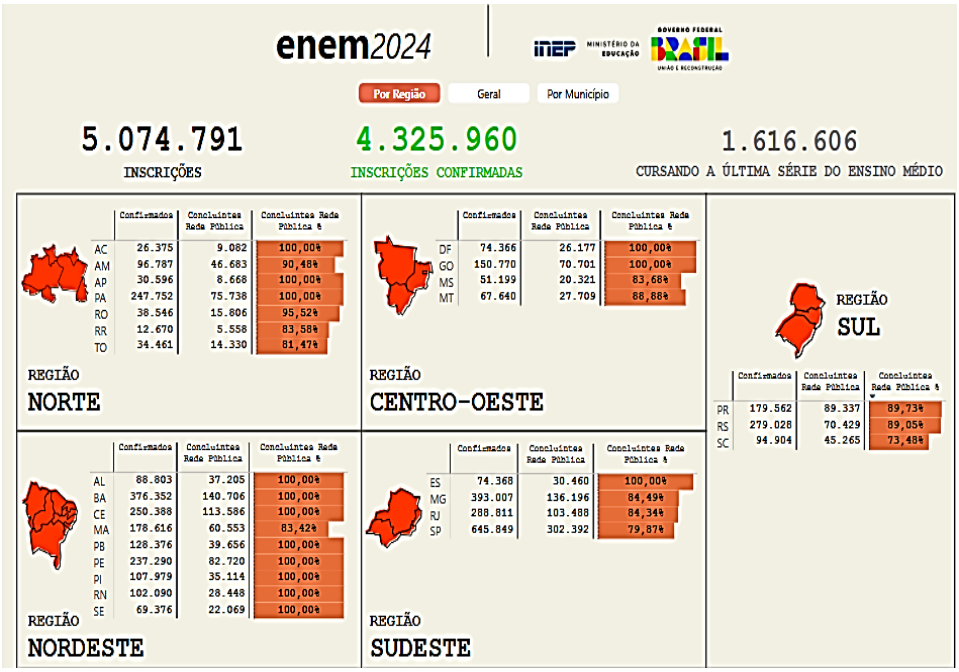
Em 2009, o Enem foi reformulado, adotando a Teoria de Resposta ao Item (TRI)[6], o que expandiu o número de questões para 180 e incorporou a certificação do ensino médio, além da criação do Sistema de Seleção Unificada (Sisu). Nos anos posteriores, o exame avançou em inclusão social e em segurança, com a implementação de atendimento por nome social, em 2015; com a coleta biométrica e com detectores de metal, em 2016; e com medidas de contenção de custos e justificativa de ausência para a isenção, em 2018.

Cumpre ressaltar que a realização do Enem traz em si, aspectos sociais, regionais e educacionais dos participantes. Considerando isto, ilustro abaixo, a título de informação e acréscimo, dados relacionados às inscrições confirmadas para o ENEM 2024, evidenciando a distribuição regional dos candidatos e a participação de concluintes de escolas públicas.

[5] ProUni é o Programa Universidade para Todos, criado pelo MEC (Ministério da Educação) em 2004 para selecionar candidatos e preencher vagas em universidades particulares de todo o Brasil.

[6] O sistema da TRI fornece um gráfico em que constam as questões com maiores e menores números de acertos, sendo possível classificar qual o nível de dificuldade de cada questão. Pela lógica do TRI, se sai melhor o candidato que acerta mais questões fáceis e médias do que difíceis, pois o sistema entende que o participante apresentou um comportamento coerente, ou seja, não “chutou”.

Figura 2 - Distribuição Regional das Inscrições Confirmadas no ENEM 2024



Fonte: Inep (2024)

Legenda: a imagem representa os números de inscritos no Enem, por regiões e estados, no ano de 2024; mostra também a taxa de confirmação das inscrições.

A imagem representada na figura 2, nos mostra um panorama estatístico das inscrições confirmadas no ENEM 2024, distribuídas por região do Brasil. O total de inscrições realizadas foi de 5.074.791, das quais 4.325.960 foram confirmadas. Além disso, 1.616.606 candidatos estavam cursando a última série do Ensino Médio. A segmentação da tabela foi por regiões (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), destacando o número de inscrições confirmadas e a porcentagem de candidatos provenientes de escolas públicas.

Vê-se que a região nordeste apresentou o maior número de confirmados (1.583.188), com destaque para estados como Bahia e Pernambuco. Na Região Norte, estados como Pará e Amazonas concentraram a maior parte das inscrições confirmadas. O Centro-Oeste, por sua vez, destacou-se no estado de Goiás, enquanto o Sudeste concentrou o maior volume em São Paulo. A Região Sul também trouxe números expressivos, especialmente no Paraná e no Rio Grande do Sul, com alta participação de estudantes de escolas públicas.

2.1.1 A redação no Enem

Nesta subseção, por conseguinte, apresentarei informações detalhadas que condizem com as competências avaliativas do Enem, bem como, aos critérios de avaliação.

Assim como ocorre em qualquer processo seletivo na universidade, que exija a produção textual como requisito, o Exame Nacional do Ensino Médio (Enem) apresenta especificidades relacionadas à escrita, que precisam ser rigorosamente atendidas pelos participantes. Tais características delineiam os critérios que guiam a elaboração da redação e que são fundamentais para o êxito no exame. Logo, para alcançar a nota máxima na redação no Enem, equivalente a mil pontos, é essencial que o texto atenda integralmente aos critérios estabelecidos na matriz de correção.

A matriz de correção é um documento orientador que traz as competências avaliadas e que tolera apenas desvios mínimos nas convenções da escrita. O documento é divulgado anualmente pelo governo federal, em colaboração com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), por meio da "Cartilha do Participante" e das redes sociais oficiais vinculadas ao exame.

Figura 3 - Cartilha do participante (2024)

Fonte: Cartilha do participante (2024)

Legenda: a imagem representa a capa da cartilha do participante, disponibilizada em 2024. O sumário interativo traz seções como a matriz de referência para a redação e as 5 competências avaliadas. Também inclui uma amostra de redações nota mil do Enem 2023.

Figura 4 - Panorama das cinco competências avaliativas do Enem

Competência I	Demonstrar domínio da modalidade escrita formal da língua portuguesa.
Competência II	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
Competência III	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
Competência IV	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
Competência V	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Fonte: Cartilha do participante (2024)
Legenda: a imagem traz uma visão geral cinco competências avaliativas exigidas pelo Enem.

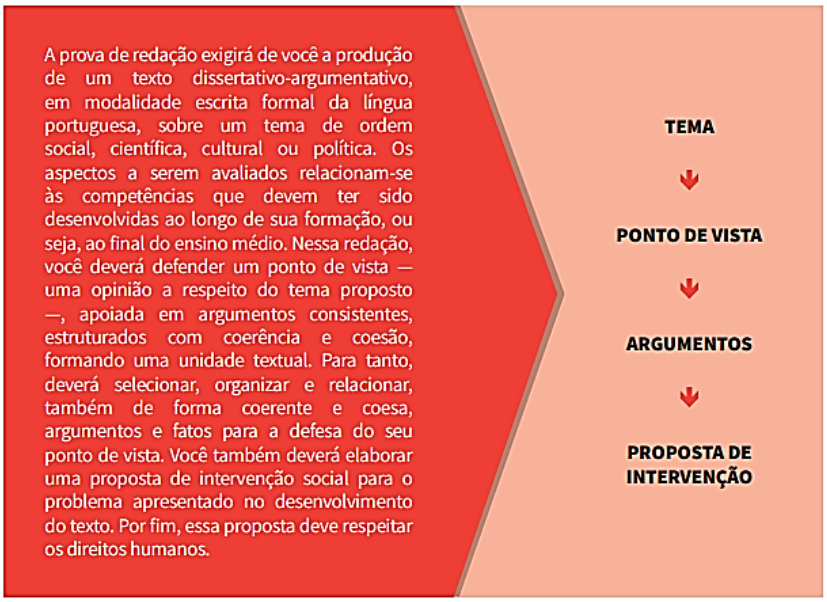
Diante das cinco competências apresentadas, percebe-se a complexidade da tarefa para os candidatos que buscam alcançar boas avaliações em todos os critérios da redação, no formato exigido pelo Enem. Cada competência exige desempenho máximo, o que requer atenção a detalhes linguísticos e domínio das habilidades propostas. Por outro lado, do ponto de vista do corretor/avaliador, há uma margem de subjetividade na interpretação da matriz de correção. Para lidar com esse aspecto, a banca do Enem adota um processo de dupla correção, garantindo maior transparência e justiça.

> Média de desempenho – A nota da redação, variando entre 0 (zero) e 1.000 (mil) pontos, é atribuída de acordo com os critérios disponibilizados pelo Inep nos editais do exame, e também estão previstos na Cartilha de Redação do Enem 2020. A redação é corrigida por, pelo menos, dois corretores, de forma independente, e cada um atribui uma nota entre 0 (zero) e 200 (duzentos) pontos para cada uma das cinco competências. (Brasil, 2023)

Nesse modelo, cada redação é avaliada por, no mínimo, dois corretores. Caso as notas atribuídas por eles apresentem diferenças pequenas, calcula-se a média aritmética entre as duas avaliações. Por exemplo, se um avaliador atribuir 960 pontos e outro mil pontos, a nota final será 980. Contudo, se houver discrepância superior a 100 pontos no total das avaliações ou de 80 pontos em qualquer uma das competências, o texto passa por um terceiro avaliador. Nesse caso, considera-se a média entre as avaliações mais próximas, conforme descrito na “Cartilha do Participante.” (Brasil, 2023).

À vista disso, é fundamental que tanto os candidatos quanto os professores que os preparam, compreendam as competências e as habilidades exigidas pelo exame. Entre os aspectos essenciais, estão: a abordagem completa do tema, o domínio da estrutura dissertativo-argumentativa, a utilização de repertórios socioculturais relevantes, a construção de argumentos sólidos e conectados a uma opinião bem colocada, o uso de mecanismos linguísticos que garantam coesão e coerência, e, por fim, a elaboração de uma proposta de intervenção que dialogue com o tema e com os argumentos desenvolvidos. Veja o esquema estrutural abaixo:

Figura 5 - Esquema estrutural cobrado no texto dissertativo-argumentativo



Fonte: Cartilha do participante (2024)
Legenda: a imagem traz uma visão geral cinco competências avaliativas exigidas pelo Enem.

Tomando por base a cartilha do participante, referência mais atual até o momento desta pesquisa, cumpre salientar que o documento apresenta informações detalhadas sobre cada subparte das competências avaliadas na redação do Enem. É um material que oferece diretrizes fundamentais, que podem e devem, ser estudadas tanto pelos professores de redação, responsáveis por orientar seus alunos no processo, quanto pelos próprios candidatos que participarão do exame. A seguir, estão dispostas as especificidades dos critérios, organizados de acordo com as menções anteriores.

Figura 6 - Níveis de desempenho - Competência I

200 pontos	Demonstra excelente domínio da modalidade escrita formal da língua portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizarem reincidência.
160 pontos	Demonstra bom domínio da modalidade escrita formal da língua portuguesa e de escolha de registro, com poucos desvios gramaticais e de convenções da escrita.
120 pontos	Demonstra domínio mediano da modalidade escrita formal da língua portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.
80 pontos	Demonstra domínio insuficiente da modalidade escrita formal da língua portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.
40 pontos	Demonstra domínio precário da modalidade escrita formal da língua portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.
0 ponto	Demonstra desconhecimento da modalidade escrita formal da língua portuguesa.

Fonte: Cartilha do participante (2024)
Legenda: a imagem traz uma descrição dos níveis de desempenho dentro da competência I. A pontuação segue uma escala de 0 a 200.

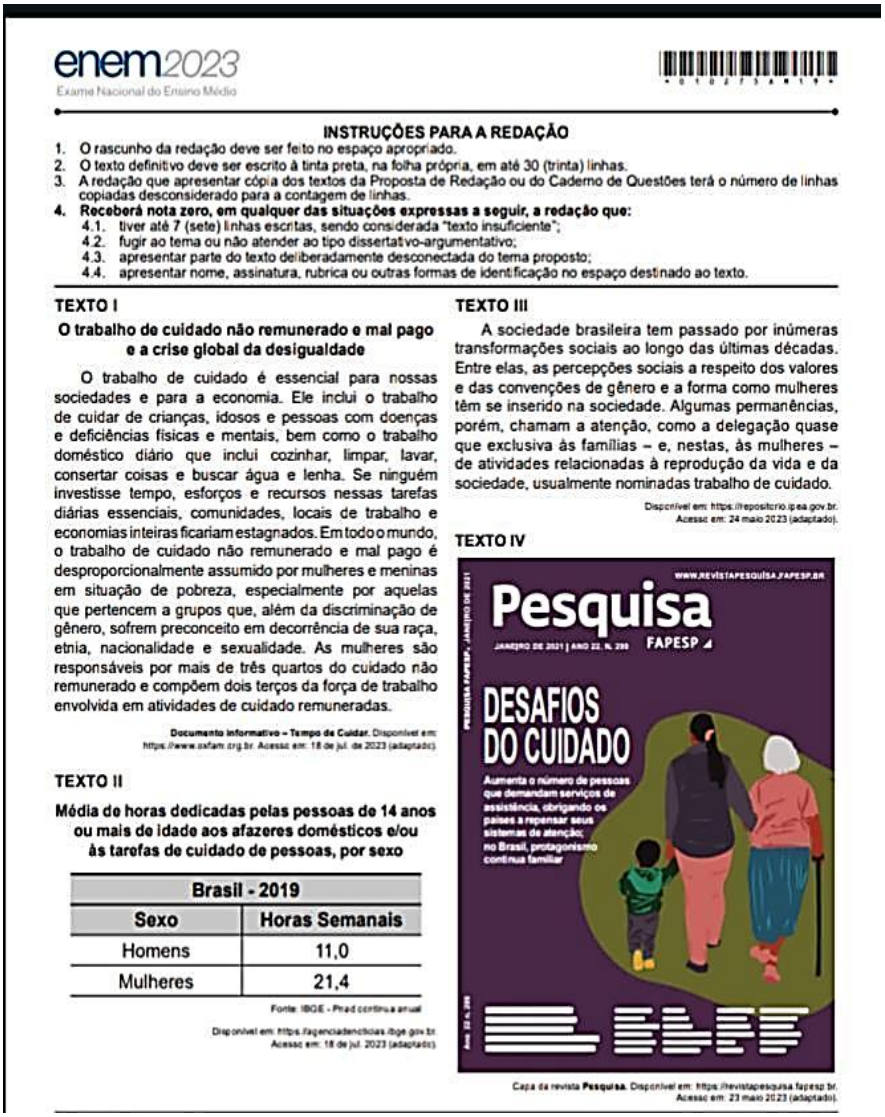
A Competência I tem como objetivo avaliar o domínio do candidato em relação à modalidade escrita formal da língua portuguesa. Espera-se que o candidato domine o conhecimento e a aplicação das convenções de escrita, como as regras de ortografia e a acentuação gráfica vigentes, estabelecidas pelo Acordo Ortográfico da Língua Portuguesa[7]. Além disso, a competência abrange a análise da adequação gramatical e da construção sintática do texto produzido.

Para esclarecer as expectativas quanto ao desempenho esperado de um concluinte do ensino médio nessa competência, destacam-se os principais aspectos avaliados pelos corretores ao definir o nível atribuído ao texto. A escrita formal, sendo

[7] Fruto de um longo trabalho da Academia Brasileira de Letras e da Academia das Ciências de Lisboa, os representantes oficiais dos então sete países de língua oficial portuguesa (além do Brasil e de Portugal, também Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe) assinaram em 1990 o Acordo Ortográfico da Língua Portuguesa, ratificado também, depois da sua independência em 2004, por Timor-Leste. O Acordo Ortográfico da Língua Portuguesa (1990) entrou em vigor no início de 2009 no Brasil e em 13 de maio de 2009 em Portugal. Em ambos os países foi estabelecido um período de transição em que tanto as normas anteriormente em vigor como a introduzida por esta nova reforma são válidas: esse período é de três anos no Brasil e de seis anos em Portugal. Com exceção de Angola e de Moçambique, todos os restantes países da CPLP já ratificaram todos os documentos conducentes à aplicação desta reforma. (Portal da Língua Portuguesa, 2025)

a modalidade associada ao gênero dissertativo-argumentativo, exige atenção especial a determinados critérios, os quais são ressaltados já na proposta de redação.

Figura 7 - Proposta de redação – Enem 2023



enem2023
Exame Nacional do Ensino Médio

INSTRUÇÕES PARA A REDAÇÃO

1. O rascunho da redação deve ser feito no espaço apropriado.
2. O texto definitivo deve ser escrito à tinta preta, na folha própria, em até 30 (trinta) linhas.
3. A redação que apresentar cópia dos textos da Proposta de Redação ou do Caderno de Questões terá o número de linhas copiadas desconsiderado para a contagem de linhas.
4. **Receberá nota zero, em qualquer das situações expressas a seguir, a redação que:**
   4.1. tiver até 7 (sete) linhas escritas, sendo considerada "texto insuficiente";
   4.2. fugir ao tema ou não atender ao tipo dissertativo-argumentativo;
   4.3. apresentar parte do texto deliberadamente desconectada do tema proposto;
   4.4. apresentar nome, assinatura, rubrica ou outras formas de identificação no espaço destinado ao texto.

TEXTO I

O trabalho de cuidado não remunerado e mal pago e a crise global da desigualdade

O trabalho de cuidado é essencial para nossas sociedades e para a economia. Ele inclui o trabalho de cuidar de crianças, idosos e pessoas com doenças e deficiências físicas e mentais, bem como o trabalho doméstico diário que inclui cozinhar, limpar, lavar, consertar coisas e buscar água e lenha. Se ninguém investisse tempo, esforços e recursos nessas tarefas diárias essenciais, comunidades, locais de trabalho e economias inteiras ficariam estagnados. Em todo o mundo, o trabalho de cuidado não remunerado e mal pago é desproporcionalmente assumido por mulheres e meninas em situação de pobreza, especialmente por aquelas que pertencem a grupos que, além da discriminação de gênero, sofrem preconceito em decorrência de sua raça, etnia, nacionalidade e sexualidade. As mulheres são responsáveis por mais de três quartos do cuidado não remunerado e compõem dois terços da força de trabalho envolvida em atividades de cuidado remuneradas.

Documento informativo – Tempo de Cuidar. Disponível em: https://www.oxfam.org.br. Acesso em: 18 de jul. de 2023 (adaptado).

TEXTO II

Média de horas dedicadas pelas pessoas de 14 anos ou mais de idade aos afazeres domésticos e/ou às tarefas de cuidado de pessoas, por sexo

Brasil - 2019	
Sexo	**Horas Semanais**
Homens	11,0
Mulheres	21,4

Fonte: IBGE - Pnad contínua anual

Disponível em: https://agenciadenoticias.ibge.gov.br. Acesso em: 18 de jul. 2023 (adaptado).

TEXTO III

A sociedade brasileira tem passado por inúmeras transformações sociais ao longo das últimas décadas. Entre elas, as percepções sociais a respeito dos valores e das convenções de gênero e a forma como mulheres têm se inserido na sociedade. Algumas permanências, porém, chamam a atenção, como a delegação quase que exclusiva às famílias – e, nestas, às mulheres – de atividades relacionadas à reprodução da vida e da sociedade, usualmente nominadas trabalho de cuidado.

Disponível em: https://repositorio.ipea.gov.br. Acesso em: 24 maio 2023 (adaptado).

TEXTO IV

Capa da revista **Pesquisa**. Disponível em: https://revistapesquisa.fapesp.br. Acesso em: 23 maio 2023 (adaptado).

Fonte: Inep (2023)

Legenda: a imagem traz uma descrição da proposta de redação do Enem 2023; a temática foi: "Desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil".

Nessa proposta, o participante é orientado a produzir um texto formal, conforme a instrução:

> A partir da leitura dos textos motivadores e com base nos conhecimentos construídos ao longo de sua formação, redija texto dissertativo-argumentativo em modalidade escrita formal da língua portuguesa sobre o tema "Desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil", apresentando uma proposta de intervenção que respeite

> os direitos humanos. Selecione, organize e relacione, de forma coerente e coesa, argumentos e fatos para a defesa de seu ponto de vista.(Enem, 2023)

Dessa forma, a avaliação nessa competência considera possíveis problemas de construção sintática, assim como desvios relacionados a quatro aspectos fundamentais:

- Convenções da escrita: incluem ortografia, acentuação, e outros elementos da norma-padrão.
- Regras gramaticais: englobam concordância, regência, tempos e modos verbais, entre outros.
- Escolha de registro: refere-se à necessidade de adequação à formalidade do gênero.
- Escolha vocabular: avalia o uso preciso e contextualizado do vocabulário.

A avaliação da Competência I na redação do Enem inclui a análise da estrutura sintática e dos desvios linguísticos, elementos que estão diretamente ligados às regras da língua portuguesa e à construção textual. Uma estrutura sintática bem elaborada pressupõe a presença de elementos oracionais organizados de maneira a garantir fluidez na leitura, clareza na exposição das ideias e períodos bem estruturados e completos. Como o gênero exigido é o texto dissertativo-argumentativo, escrito na modalidade formal da língua portuguesa, espera-se que as redações de nota máxima apresentem períodos com maior complexidade, utilizando orações subordinadas e intercaladas, de forma adequada.

Por outro lado, falhas na estrutura sintática comprometem a qualidade do texto. Erros comuns incluem: períodos truncados, no qual duas orações que deveriam fazer parte de um mesmo período são separadas por um ponto final, e a justaposição, em que vírgulas substituem inadequadamente o ponto final, comprometendo a segmentação correta das frases. A frequência desses problemas determina a pontuação atribuída na respectiva competência.

O segundo aspecto avaliado na redação do Enem é a compreensão da proposta de redação, que exige a elaboração de um texto dissertativo-argumentativo. Esse gênero textual vai além da simples exposição de ideias, pois requer que o candidato defenda um ponto de vista ou uma ideia por meio de argumentação sólida.

É fundamental que o ponto de vista apresentado esteja diretamente relacionado ao tema proposto.

O tema da redação é o núcleo em torno do qual as ideias do texto devem ser organizadas. Ele representa um recorte específico de um assunto mais amplo, sendo essencial que o candidato atenda a esse recorte. Desenvolver parcialmente o tema (tangenciamento) ou se desviar completamente dele pode comprometer seriamente a pontuação da redação.

Outro elemento essencial avaliado na Competência II é a utilização de repertório sociocultural produtivo[8]. Esse repertório consiste em informações, fatos, citações ou experiências que contribuam como argumentos para o desenvolvimento do texto. Para atender plenamente a essa competência, recomenda-se ao candidato seguir algumas práticas fundamentais:

- Ler atentamente a proposta de redação e os textos motivadores, compreendendo integralmente o que está sendo solicitado.
- Refletir sobre o tema para definir o foco da discussão, determinando o ponto de vista a ser adotado e como defendê-lo.
- Evitar copiar trechos dos textos motivadores, já que a recorrência de cópias é penalizada e pode até levar à anulação da redação.
- Utilizar as ideias dos textos motivadores como inspiração, mas não se limitar a elas, in*corpora*ndo informações e conhecimentos de outras áreas relacionadas ao tema.
- Selecionar repertórios socioculturais pertinentes e articulá-los de forma produtiva no texto, de tal modo que eles sirvam a um propósito determinado: validar o ponto de vista apresentado.
- Permanecer dentro dos limites do tema proposto, cuidando para não o tangenciar nem fugir completamente dele.

[8] Para a redação do Enem, o repertório sociocultural pode incluir referências a filmes, livros, obras de arte, citações e dados científicos ou históricos.

Figura 8 - Níveis de desempenho - Competência II

200 pontos	Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sociocultural produtivo, e apresenta excelente domínio do texto dissertativo-argumentativo.
160 pontos	Desenvolve o tema por meio de argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
120 pontos	Desenvolve o tema por meio de argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
80 pontos	Desenvolve o tema recorrendo à cópia de trechos dos textos motivadores ou apresenta domínio insuficiente do texto dissertativo-argumentativo, não atendendo à estrutura com proposição, argumentação e conclusão.
40 pontos	Apresenta o assunto, tangenciando o tema, ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais.
0 ponto	Fuga ao tema/não atendimento à estrutura dissertativo-argumentativa. Nestes casos a redação recebe nota zero e é anulada.

Fonte: Cartilha do participante (2024)
Legenda: a imagem traz uma descrição dos níveis de desempenho dentro da competência II. A pontuação segue uma escala de 0 a 200.

Não se pode olvidar que a fuga ao tema, deve ser uma preocupação por parte do candidato. Uma redação é considerada como tendo fugido totalmente ao tema quando não desenvolve nem o assunto mais amplo nem o tema específico proposto na proposta de redação. No Enem 2023, por exemplo, o tema abordava “os desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil”; para atender plenamente ao tema, era necessário reconhecer, inicialmente, a existência desse problema social e, em seguida, discutir os desafios para superá-lo. Redações que fugiram ao tema receberam essa classificação por desenvolverem assuntos que não se relacionavam com o recorte temático.

Exemplos incluem:

- Discussões sobre o trabalho no Brasil sem vínculo com o trabalho de cuidado ou com as mulheres;
- Abordagem da assistência oferecida por instituições públicas ou privadas sem menção à figura da pessoa que cuida;
- Enfoque exclusivo no trabalho de cuidado exercido por homens.

O terceiro aspecto analisado na redação do Enem está relacionado à capacidade de fazer seleções, organizações, relações e interpretações de fatos e opiniões para a defesa de um parecer. Para atender a essa exigência, é essencial que o texto apresente, objetivamente, a opinião defendida e a argumentação utilizada para justificar o posicionamento adotado frente à temática proposta. Essa análise é realizada no âmbito da Competência III, que avalia a inteligibilidade do texto, isto é, o nível de coerência e de pertinência das ideias apresentadas, que estão totalmente associadas ao plano inicial de escrita, também chamado de projeto de texto. Para uma redação ser inteligível, cabem os seguintes aspectos:

- Seleção de argumentos e a escolha acertada das ideias que sustentam o posicionamento.
- No que concerne à relação de sentido, espera-se uma nexos lógicos entre as partes do texto, garantindo a coesão.
- Pensando na progressão temática, requer-se a organização gradual das ideias, evidenciando um plano de estratégia.
- Referente ao desenvolvimento dos argumentos, o esperado é evidenciar a importância das informações escolhidas para sustentar o posicionamento.

A cartilha do participante define o projeto de texto da seguinte forma:

> Projeto de texto é o planejamento prévio à escrita da redação. É o esquema que se deixa perceber pela organização estratégica dos argumentos presentes no texto. Nesse projeto, são definidos os argumentos que serão mobilizados para a defesa do ponto de vista e qual a melhor ordem para apresentá-los, de modo a garantir que o texto final seja articulado, claro e coerente. (Brasil, 2024)

Figura 9 - Níveis de desempenho - Competência III

200 pontos	Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista.
160 pontos	Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista.
120 pontos	Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista.
80 pontos	Apresenta informações, fatos e opiniões relacionados ao tema, mas desorganizados ou contraditórios e limitados aos argumentos dos textos motivadores, em defesa de um ponto de vista.
40 pontos	Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista.
0 ponto	Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista.

Fonte: Cartilha do participante (2024)
Legenda: a imagem traz uma descrição dos níveis de desempenho dentro da competência III. A pontuação segue uma escala de 0 a 200.

Na competência 4, dando sequência, os aspectos avaliados fazem alusão à lógica de articulação entre as partes do texto, certificando a coesão necessária para que as ideias sejam compreendidas com fluidez. É necessário que frases e parágrafos estejam conectados de maneira coerente, de tal forma que as ideias apresentadas estejam inter-relacionadas e que contribuam com a progressão temática textual. Essa conexão é assegurada pelo uso de recursos coesivos, como operadores argumentativos: "assim como", "outrossim", entretanto", "porém", "por isso", "assim", "enfim", "portanto", etc. Ademais, elementos funcionais, como: preposições, conjunções, advérbios e pronomes.

Nesse sentido, a competência IV avalia a variabilidade e eficácia de recursos linguísticos que promovam a continuidade entre as partes textuais. Ela está relacionada à competência III, todavia, aquela foca a superfície do texto, levando em conta os recursos linguísticos que facilitarão o entendimento do leitor. Segundo os critérios de correção, para assegurar a coesão do texto, alguns princípios devem ser seguidos em diferentes níveis:

- Estruturação dos parágrafos: cada parágrafo deve apresentar uma ideia principal articulada a ideias secundárias, podendo ser desenvolvido por estratégias como causa e consequência, exemplificação, comparação ou detalhamento. Deve-se garantir a articulação explícita entre os parágrafos.

- Estruturação dos períodos: geralmente, no texto dissertativo-argumentativo os períodos possuem certa complexidade, por serem compostos por mais de uma oração, possibilitando que as relações de causa e efeito, bem como, de contradição e de temporalidade sejam expressas.
- Referenciação: a retomada de elementos no texto, bem como, a sua introdução, são realizadas por meio de pronomes, de advérbios, de artigo, de sinônimos, de antônimos, de hipônimos, de hiperônimos, de expressões resumitivas, metafóricas e metadiscursivas.

Figura 10 - Níveis de desempenho - Competência IV

200 pontos	Articula bem as partes do texto e apresenta repertório diversificado de recursos coesivos.
160 pontos	Articula as partes do texto, com poucas inadequações, e apresenta repertório diversificado de recursos coesivos.
120 pontos	Articula as partes do texto, de forma mediana, com inadequações, e apresenta repertório pouco diversificado de recursos coesivos.
80 pontos	Articula as partes do texto, de forma insuficiente, com muitas inadequações, e apresenta repertório limitado de recursos coesivos.
40 pontos	Articula as partes do texto de forma precária.
0 ponto	Não articula as informações.

Fonte: cartilha do participante (2024)
Legenda: a imagem traz uma descrição dos níveis de desempenho dentro da competência IV. A pontuação segue uma escala de 0 a 200.

O quinto aspecto avaliado na redação do Enem é a elaboração de uma proposta de intervenção para o problema apresentado no tema, respeitando os princípios dos direitos humanos. Para que a quinta competência seja obedecida, é necessário sugerir uma iniciativa possível e concreta que vise a solução do problema discutido no texto, o que chamamos de proposta de intervenção; essa proposta precisa demonstrar tanto a análise crítica do escritor, quanto a sua capacidade de

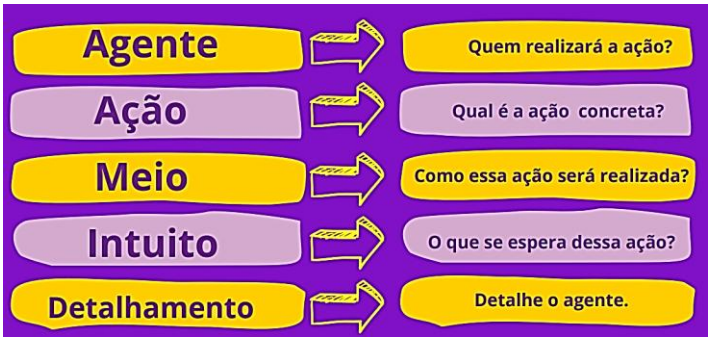
reflexão acerca de soluções que evoquem a cidadania e o respeito aos direitos fundamentais dos seres humanos[9].

A proposta de intervenção precisa estar intrinsecamente relacionada à temática proposta, bem como, alinhada ao projeto de texto desenvolvido. Em outras palavras, ela precisa estar em consonância com o posicionamento adotado e com os argumentos que foram apresentados ao longo do texto, enfatizando a autoria sobre as possíveis estratégias para o problema em questão. A

Para que a proposta seja bem avaliada, é essencial que ela seja concreta, específica e consistente com as ideias desenvolvidas. Além de propor uma ação interventiva, é necessário detalhá-la, indicando:

- Ação: o que deve ser feito como solução para o problema?
- Agente: quem é o responsável por executar essa ação? Pode ser um indivíduo, uma família, uma comunidade, uma entidade social, um governo ou outra instância competente.
- Meio de execução: Como essa ação será implementada?
- Efeito ou finalidade (intuito): Quais resultados podem ser alcançados com a intervenção?

Figura 11 - Quesitos necessários à proposta de intervenção



Fonte: autoria própria

Legenda: a imagem representa o passo a passo para a escrita de uma proposta de intervenção no formato exigido pelo Enem.

Uma proposta de intervenção plenamente desenvolvida deve responder às questões iniciais, trazendo uma conexão direta ao problema que foi apresentado e

[9] Em resumo, na prova de redação do Enem, quaisquer que sejam os temas propostos para o desenvolvimento do texto dissertativo-argumentativo, constituem desrespeito aos direitos humanos propostas que incitam as pessoas à violência, ou seja, aquelas em que transparece a ação de indivíduos na administração da punição — por exemplo, as que defendem a "justiça com as próprias mãos". (Inep, 2024)

endossando argumentação textual. Ao apresentar uma solução, o candidato deve demonstrar não apenas que domina o gênero textual proposto, mas também a habilidade de proposição de mudanças que sejam possíveis e relevantes contribuindo para que questões sociais (aliadas aos direitos humanos) sejam solucionadas.

Figura 12 - Níveis de desempenho - Competência V

200 pontos	Elabora muito bem proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.
160 pontos	Elabora bem proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
120 pontos	Elabora, de forma mediana, proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
80 pontos	Elabora, de forma insuficiente, proposta de intervenção relacionada ao tema, ou não articulada com a discussão desenvolvida no texto.
40 pontos	Apresenta proposta de intervenção vaga, precária ou relacionada apenas ao assunto.
0 ponto	Não apresenta proposta de intervenção ou apresenta proposta não relacionada ao tema ou ao assunto.

Fonte: cartilha do participante (2024)
Legenda: a imagem traz uma descrição dos níveis de desempenho dentro da competência V. A pontuação segue uma escala de 0 a 200.

Por fim, a cartilha do participante também informa o candidato situações que podem zerar uma redação. São elas:

> A redação receberá nota 0 (zero) se apresentar uma das características a seguir: • fuga total ao tema; • não obediência ao tipo dissertativo-argumentativo; • ausência de texto escrito na Folha de Redação, que será considerada “Em Branco”; • extensão de até 7 (sete) linhas manuscritas, qualquer que seja o conteúdo ou extensão de até 10 (dez) linhas escritas no sistema Braille, situações que configurarão “Texto insuficiente”; • impropérios, desenhos e outras formas propositais de anulação, o que configurará “Anulada”; • parte deliberadamente desconectada do tema proposto; • nome, assinatura, rubrica ou qualquer outra forma de identificação no espaço destinado exclusivamente ao texto da redação, o que configurará “Anulada”. (Cartilha do participante, 2024)

Panoramicamente, foram expostos até aqui, os critérios exigidos pelo Enem para a escrita de uma redação de alto desempenho; as competências foram descritas. Optei por trazer esse detalhamento, em virtude da natureza do *corpus* da pesquisa,

do tipo de texto analisado. Daqui para frente, portanto, em viés análogo, trarei questões importantes acerca dos desafios de escrita presentes no texto dissertativo-argumentativo, exigido pelo Enem.

2.2 A ESCRITA DE UM TEXTO DISSERTATIVO-ARGUMENTATIVO

A escrita de um texto requer habilidades específicas que nem sempre são plenamente desenvolvidas ao longo do processo educacional. No contexto escolar, a produção textual frequentemente se torna um desafio significativo para o estudante, especialmente quando sua habilidade de escrita não foi potencializada desde os primeiros anos de letramento e de alfabetização no ensino fundamental. Essa dificuldade tende a se agravar com o passar dos anos, culminando em momentos de grande pressão, como a preparação para o Enem e para os vestibulares, nos quais a escrita de um texto dissertativo-argumentativo é exigida.

Há que se ponderar que um texto bem elaborado, rigorosamente alinhado aos critérios de correção, incluindo a autoria argumentativa, tende a ser avaliado com nota máxima nos parâmetros exigidos, como vimos na subseção anterior (2.1.1). Isso ocorre porque tal texto apresenta conteúdo argumentativo sólido e completo, sem deixar espaços argumentativos a serem preenchidos por quem avalia, além de conter recursos linguísticos sintáticos e semânticos que estejam em harmonia com o contexto que foi abordado. Ademais, a abordagem completa do tema em questão é indispensável. A respeito do tema, no Enem, por exemplo, podemos dizer que, geralmente, evitam questões polêmicas, a prioridade é apresentar um caráter social e contextualizado de acordo com a realidade brasileira atual.

Tabela 1 - Temas das redações do ENEM de 2014 a 2023

Ano	**Tema:**
2014	Publicidade infantil em questão no Brasil
2015	A persistência da violência contra a mulher na sociedade brasileira
2016	Caminhos para combater a intolerância religiosa no Brasil
2017	Desafios para a formação educacional de surdos no Brasil
2018	Manipulação do comportamento do usuário pelo controle de dados na internet
2019	Democratização do acesso ao cinema no Brasil

2020	O estigma associado às doenças mentais na sociedade brasileira
2021	Invisibilidade e registro civil: garantia de acesso à cidadania no Brasil
2022	Desafios para a valorização de comunidades e povos tradicionais no Brasil
2023	Desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil
2024	Desafios para a valorização da herança africana no Brasil

Fonte: autoria própria

Embora outros aspectos também sejam considerados pelas bancas avaliadoras, independentemente da tipologia/gênero cobrados, para determinar um alto desempenho na produção textual, é crucial que os fatos apresentados sejam bem fundamentados, que tenham progressão temática e concatenação. Ou seja, a ausência de textualidade e de um conteúdo informativo consistente pode resultar em uma avaliação insatisfatória. No caso do texto dissertativo-argumentativo, entendemos assim:

> O texto do tipo dissertativo-argumentativo é aquele que se organiza com base na defesa de um ponto de vista sobre determinado assunto. É fundamentado com argumentos, a fim de influenciar a opinião da pessoa que lê, tentando convencê-la de que a ideia defendida é válida. É preciso, portanto, expor e explicar ideias. Por isso, a dupla natureza desse tipo textual: é argumentativo porque defende um ponto de vista, uma opinião, e é dissertativo porque utiliza explicações para justificá-lo. (INEP, 2024, p. 19)

Depreende-se, portanto, que a informação apresentada em uma produção escrita é um elemento crucial para sua validação pelo leitor ou ouvinte. Nesse contexto, a inclusão de fatos intertextuais na elaboração de uma opinião é o cerne, quanto maior o caráter informativo, mais alimentada se torna a ideia central do texto, alinhando-se ao tema proposto.

Para atingir a nota máxima em qualquer exame, o “poder da informação” exige, essencialmente, uma prática contínua de leitura eficiente — aquela em que o leitor compreende plenamente o conteúdo apresentado —, aliada à prática regular de escrita e reescrita, bem como à análise crítica e observação de aspectos que podem ser aprimorados. Assim (reafirmando a discussão inicial dessa tese no que condiz à expressão do escritor), seria inconsistente, por parte do candidato, atender a todas as regras gerais de estrutura e de adequação à norma-padrão da escrita sem demonstrar

“traços de autoria”, ou seja, a personalização da escrita, aspecto nuclear no tipo de texto exigido pelos principais exames nacionais.

> [...] ao considerá-la (a produção textual) imitação, isso implicaria um dado modelo de produção textual que se pautaria pela obediência a certos padrões linguísticos e formais. Por outro lado, considerá-la como original implica defender a primazia da criação entusiasta, da originalidade libertadora, o que apontaria para o produto de um sujeito criativo. (Baptista, 2005, p. 14)

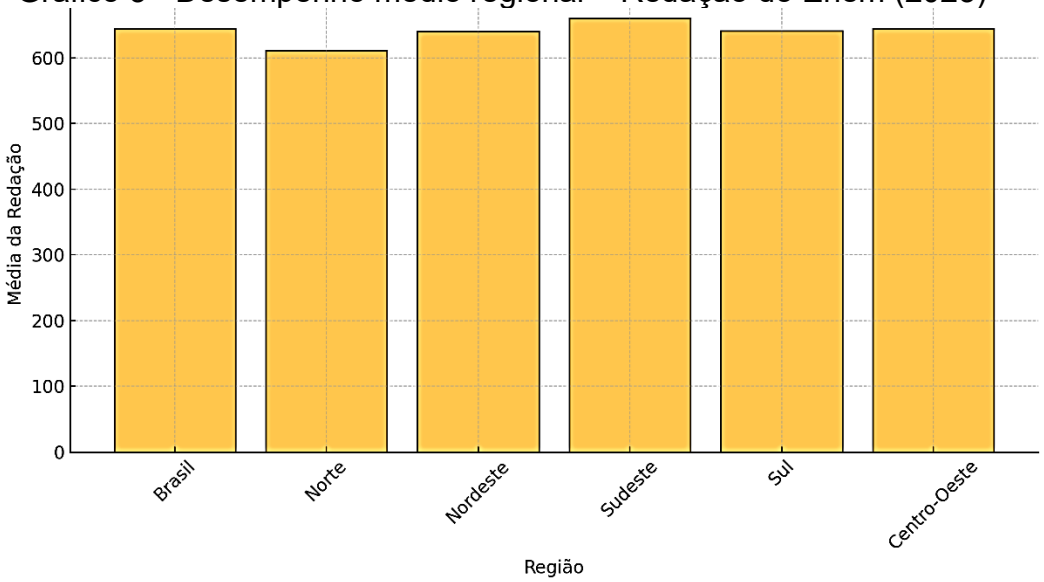
Frente ao exposto, *a priori,* no que condiz com a redação do Enem em si, *a posteriori,* com o que se espera de um texto dissertativo-argumentativo de modo geral, entendo que a redação do ENEM pode se configurar, desde que intencionalmente norteada, como um gênero textual prototípico, servindo de modelo normativo, mas que, ao mesmo tempo, abre espaço para a originalidade, permitindo que o candidato expresse sua opinião e articule conhecimentos interdisciplinares.

2.3 DADOS DO DESEMPENHO E DA PROFICIÊNCIA NO ENEM

Nesta subseção, representarei graficamente alguns dados relacionados ao desempenho e à proficiência dos participantes do Enem (2023), buscando evidenciar as tendências e diferenças observadas nas edições analisadas. A elaboração dos gráficos ocorreu a partir de dados retirados de fontes oficiais.

O primeiro gráfico (3) nos mostra as médias referentes ao desempenho em redação, por região brasileira, dando destaque às variações geográficas significativas. O segundo gráfico (4) por sua vez nos traz uma linha do tempo de proficiência, expondo a distribuição de participantes por faixas de pontuação, numa comparação entre regiões e grupos específicos.

Gráfico 3 - Desempenho médio regional – Redação do Enem (2023)



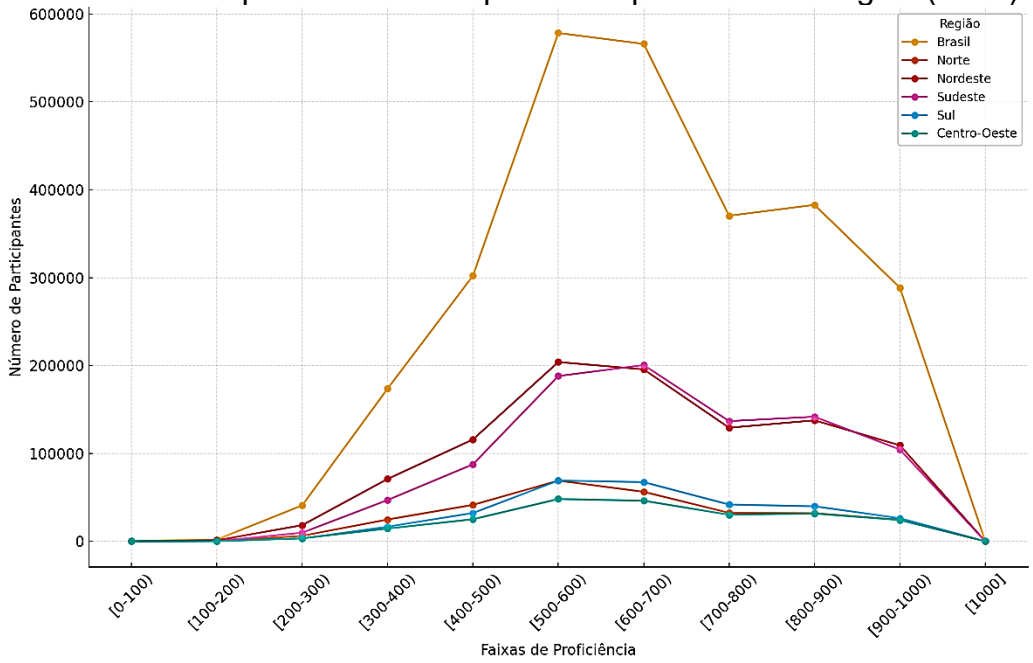
Fonte: autoria própria

Legenda: a imagem nos mostra os níveis de desempenho em redação, por regiões brasileiras.

Observa-se no gráfico 3, que o Sudeste lidera as médias, alcançando um desempenho médio de 660,71 pontos , seguido pelo Centro-Oeste e o Brasil como um todo, ambos com 644,71 pontos . A região Sul também se destaca com uma média de 641,43 pontos , enquanto o Nordeste apresenta um desempenho médio de 640,57 pontos . Por outro lado, a região Norte possui a menor média entre as regiões, com 611,18 pontos. Apesar da variação entre as regiões, a proximidade das médias indica um nível de desempenho relativamente uniforme entre boa parte do país, exceto pelo Norte, que aparece como um ponto de atenção.

.

Gráfico 4 - Participantes do Enem por faixas proficiência e região (2023)



Fonte: autoria própria

Legenda: a imagem nos mostra o quantitativo de participantes do Enem por faixa de proficiência.

Consoante às representações gráficas feitas até aqui, o gráfico 4, apresenta a quantidade de participantes do Enem (2023) distribuída entre diferentes faixas de proficiência em redação, categorizada por região geográfica do Brasil. Ao analisarmos o gráfico, é possível observar que as regiões Sudeste e Nordeste se destacam em participação nas faixas intermediárias e altas de proficiência, especialmente entre [500-600) e [700-800). A faixa [500-600) revela-nos a maior quantidade de candidatos em todas as regiões, seguida pelas faixas imediatamente superiores. Em contraste, a região Norte apresenta os menores números em todas as faixas, refletindo uma participação proporcionalmente mais baixa.

Embora o foco deste trabalho não esteja voltado para uma análise de cunho social, mas linguístico, considero pertinente incluir os dados de desempenho e de proficiência no Enem para um entendimento mais holístico da prática avaliativa, especialmente no que tange à redação. Os gráficos apresentados nesta subseção permitiram observar tendências regionais e variações na distribuição de proficiência dos participantes, oferecendo um contexto mais amplo para compreender o impacto e a abrangência do exame, o que pode ser extremamente útil para estudos futuros que contemplem fatores relacionados ao desempenho na escrita e que podem

dialogar com as práticas avaliativas e com as competências exigidas na produção textual do Enem.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem por objetivo apresentar o referencial teórico que fundamenta as premissas estabelecidas e que serve de base conceitual para o desenvolvimento desta pesquisa. De forma preliminar, referencio trabalhos acerca da Linguística de *Corpus,* de *corpora* especializados e da escrita de redações. Em seguida, elucido conceitos relacionados à Lexicologia e ao Processamento de Linguagem Natural com ênfase nos Estudos Linguísticos. No que se refere à linguística de *corpus*, este capítulo a abordará como um campo concentrado na análise de grandes coleções de textos autênticos, os *corpora*, objetivando a identificação de padrões linguísticos e de características da língua em uso.

3.1 TRABALHOS RELACIONADOS: LINGUÍSTICA DE *CORPUS, CORPORA* ESPECIALIZADOS E ESCRITA DE REDAÇÕES.

Como ressaltado em seção pretérita, neste estudo trabalho com um *corpus* especializado, ou seja, coletâneas de textos escritos. Conforme assegura Berber Sardinha (2000, p. 340), quanto ao conteúdo, os *corpora* podem ser:

- Especializado: os textos são de tipos específicos (em geral, gêneros ou registros definidos).
- Regional ou dialetal: os textos são provenientes de uma ou mais variedades sociolinguísticas específicas.
- Multilíngue: inclui idiomas diferentes.

Os *corpora* especializados consistem em textos que possuem uma especificidade de características de um domínio ou de um contexto particular, geralmente são resultados de produções acadêmicas, técnicas e profissionais, ao contrário da maioria dos *corpora* utilizados em estudos de Linguística de *Corpus*, que podem incluir textos revisados ou publicados, livros, artigos científicos, manuais técnicos ou notícias. Nesse sentido, o *corpus* especializado além de refletir alto grau de adequação linguística, também possibilita análises minuciosas acerca de padrões lexicais, sintáticos e gramaticais, em contextos específicos.

Na esteira dos estudos que consideram uma análise robusta de *corpora* de redações de vestibular, menciono o livro “Crise na Linguagem: a redação no

vestibular". Nele, Rocco (1981) tratou da relação entre a produção textual de vestibulandos e a estruturação mental dos candidatos.

A pesquisa utilizou um *corpus* composto por 1.500 redações da segunda fase do vestibular da FUVEST, realizadas em 1978, representando 5% do total de textos produzidos naquele ano. O objetivo central era examinar a existência de uma possível crise na linguagem escrita, com foco na coerência, na coesão e na criatividade dos textos. Rocco (1981, p. 54), afirma que "numa sociedade onde tudo nos é dado como pronto, são muitos os clichês percebidos na escrita, escritos sem lógica e coerência, *nonsense* e cheio de contradições."

A autora fundamentou a sua metodologia em teorias de linguística e lógica, integrando perspectivas de Piaget para analisar o nível de pensamento formal e hipotético-dedutivo dos vestibulandos. A pesquisa abordou tanto aspectos microtextuais (frásicos e interfrásicos) quanto macrotextuais, considerando fatores como relações lógicas, uso de conectivos, originalidade e correspondência ao tema proposto.

Os resultados apontaram para um baixo desempenho geral dos candidatos, com a maioria dos textos apresentando problemas de coerência, clichês e falta de correspondência ao tema. Apenas uma pequena parcela dos textos foi considerada adequada ou criativa. A pesquisa revelou que variáveis, como, o tipo de escola frequentada ou o acesso a cursinhos preparatórios não tinham impacto significativo na qualidade textual da maioria dos estudantes.

Tratando, mais uma vez, especificamente de *corpora* de redações, de textos produzidos por falantes nativos do português, Oliveira (2016), desenvolveu um estudo com o objetivo de caracterizar o gênero redação do ENEM sob perspectivas textuais e discursivas. A autora, cujo campo de pesquisa é o da Linguística Textual, analisou 100 redações do ano de 2013; ela buscou compreender tanto os aspectos externos quanto internos, inseridos nesse gênero textual. O trabalho foi guiado por uma abordagem metodológica robusta, baseada em concepções teóricas de gêneros como ações sociais (Miller, 2009 [1984]), nas diretrizes para a análise de gênero propostas por Bazerman (2016a) e em elementos argumentativos descritos por autores como Adam (2008) e Perelman e Olbrechts-Tyteca (2005).

As análises envolveram as seguintes etapas: a descrição das produções do gênero; a caracterização dos aspectos internos e externos; a sistematização da estrutura composicional textual e discursiva; e o levantamento dos tipos de acordo e

técnicas argumentativas empregadas. Os resultados apontaram que o gênero redação do ENEM apresenta uma hierarquia de sentido que reflete a proposta de Miller (2009), especialmente no que faz jus à situação sociodiscursiva retratada no *corpus* – neste caso, a Lei Seca (tema das redações de 2013).

No campo da análise retórica, a autora identificou a predominância de técnicas argumentativas baseadas no real, como fatos e argumentos de vínculo causal, o que reforça a estruturação sólida da tipologia, enquanto texto argumentativo. Com base nesses dados, Oliveira (2016) concluiu que a redação do ENEM, embora seja um gênero atrelado à tipologia argumentativa, apresenta características únicas que a diferenciam de outros textos dessa natureza, especialmente por sua especificidade enquanto prática situada no âmbito de uma política pública de avaliação educacional.

Dentro dos estudos baseados em *corpora* de redações, Grama (2016), apresenta uma metodologia fundamentada na Linguística de *Corpus*, empregada para a análise de elementos coesivos em redações do tipo dissertativo-argumentativo no estilo ENEM. O objetivo central do seu estudo piloto foi identificar quais elementos coesivos sequenciais eram utilizados de maneira inadequada pelos estudantes, investigando se tais aspectos poderiam ser critérios válidos para nortear o foco da pesquisa de mestrado do autor, ainda em desenvolvimento.

A abordagem da Linguística de *Corpus* permitiu a análise detalhada dos textos a partir de dados linguísticos reais. Para a análise lexical, a ferramenta utilizada (de análise lexical) foi a *WordSmith Tools,* que possibilitou a identificação e a categorização dos desvios relacionados ao uso de conectores e outros recursos de coesão textual. Os resultados obtidos no estudo piloto evidenciaram a eficácia da metodologia adotada, ao fornecer subsídios concretos para a compreensão dos problemas mais recorrentes no uso de elementos coesivos. Ademais, a análise revelou que os desvios trazem um impacto significativo na qualidade da redação.

Com base em abordagens semelhantes, destaca-se a pesquisa de Barreto (2016), intitulada “A redação de vestibular sob uma perspectiva multidimensional: uma abordagem da Linguística de Corpus”. A autora analisou 100 redações de vestibulares da UERJ, aplicando a metodologia de Análise Multidimensional proposta por Biber (1988), com o intuito de observar quais traços linguísticos caracterizam os textos de diferentes níveis de desempenho. O estudo partiu da premissa de que uma redação bem-sucedida deve apresentar uma série de traços textuais recorrentes, que podem ser sistematicamente descritos por meio da quantificação de *features* linguísticas.

Para isso, Barreto (2016) empregou categorias funcionais que englobavam, por exemplo, uso de substantivos, pronomes, voz passiva, conectores adversativos, entre outros. Os resultados revelaram que, nas redações com melhores notas, havia uma predominância de construções mais abstratas e informativas, com vocabulário mais técnico, típico de uma modalidade escrita mais formal. Além disso, a autora ressalta que o cruzamento entre os dados linguísticos e as avaliações humanas permitiu identificar regularidades na linguagem que poderiam ser utilizadas para fins pedagógicos, especialmente no ensino de produção textual.

Por conseguinte, no trabalho de Levandovski et al. (2020), as autoras abordaram o uso de ferramentas tecnológicas no ensino de textos dissertativo-argumentativos, com foco na potencialização da escrita de alunos do Ensino Médio em uma escola pública da região metropolitana de Porto Alegre. A pesquisa foi conduzida com base em dificuldades observadas nas práticas de escrita dos estudantes, incluindo questões relacionadas ao uso de conectores, estruturas frasais e construção semântica.

A proposta pedagógica central do estudo foi o uso do *Sketch Engine*, uma ferramenta que permite analisar textos comparativamente, identificar elementos linguísticos e construir conhecimento a partir da observação de padrões de uso da linguagem. O objetivo foi explorar como o *Sketch Engine* pode contribuir para o aprimoramento da escrita de estudantes, especialmente no contexto da redação do ENEM, que exige o domínio do gênero dissertativo-argumentativo.

No decorrer da pesquisa, foi elaborada uma sequência didática que associava a utilização dessa ferramenta para a promoção da autonomia e da análise crítica dos alunos. Os resultados obtidos revelaram que o uso do *Sketch Engine* foi eficaz para estimular a reflexão sobre práticas de escrita, ao possibilitar a identificação de padrões linguísticos e ao promover uma leitura mais aprofundada dos textos. A ferramenta contribuiu significativamente para o desenvolvimento da coesão e da clareza textual, elementos fundamentais na produção de redações de qualidade.

Seguindo a mesma linha de trabalhos, faço menção do estudo de Souza (2022), que apresenta uma análise detalhada das redações nota mil do Exame Nacional do Ensino Médio (ENEM), com foco na massificação de sua estrutura e de conteúdo. Com base na hipótese de que a estrutura e o conteúdo dessas redações seguem padrões preestabelecidos, a pesquisa buscou investigar como as escolhas

textuais, estruturais e conteudísticas, contribuem para o êxito das redações que alcançam a nota máxima.

Composto por 32 redações distribuídas igualmente entre os anos de 2014 a 2017, o *corpus* analisado revelou padrões recorrentes que sugerem uma padronização significativa na construção dos textos, descolada da aplicação integral das cinco competências da matriz de correção do ENEM. Souza (2022), recorreu a teóricos como Adorno (2009) para discutir o conceito de massificação, Authier-Revuz (1990) para abordar a heterogeneidade enunciativa mostrada, Pêcheux (1997a) para compreender a posição discursiva dos candidatos-autores e Charaudeau (2009) para discutir a organização e a lógica argumentativa dos textos.

Os resultados indicaram que há uma forte padronização na estrutura argumentativa e no conteúdo das redações nota mil, caracterizada pela utilização frequente de elementos estruturais previamente moldados. Essa padronização parece ter sido impulsionada pelos anos preparatórios dos candidatos, nos quais as práticas de ensino são direcionadas para que os textos atendam às exigências do exame de maneira mecânica e estratégica, como "redações modelo".

Dentro da mesma tessitura teórica, Buzato et al. (2021) realizaram um estudo aprofundado sobre o uso de operadores argumentativos em textos dissertativo-argumentativos. Esses operadores são elementos fundamentais para a construção da coesão textual (*vide* subseção 2.2) e para a orientação argumentativa. Entretanto, como apontado pelos autores, há dificuldades recorrentes no emprego adequado dessas estruturas por parte dos estudantes.

A pesquisa teve como objetivo central, fazer a análise do uso dos operadores argumentativos em redações do ENEM, baseando-se em um *corpus* e em ferramentas de Linguística Computacional. A abordagem metodológica adotada utilizou fundamentos da Linguística de *Corpus* e da Linguística Computacional. Os resultados indicaram que fatores relacionados ao ensino têm impacto direto na escolha e no uso dos operadores argumentativos pelos estudantes.

Dentre os trabalhos que dialogam diretamente com esta investigação, menciono, ainda, a tese de doutorado de Aline Evers, intitulada “A redação engaiolada: padrões lexicais e ensino de redação em cursos pré-vestibulares populares.” (Porto Alegre, 2018).

O estudo analisa um corpus de 341 redações do Concurso Vestibular da Universidade Federal do Rio Grande do Sul (CVUFRGS), edição 2014, sob os aportes

da Linguística de Corpus, da Linguística Computacional, da Linguística Textual e dos Estudos do Léxico. As redações foram organizadas em faixas de desempenho e, a partir da identificação de padrões léxico-sintáticos, descreveu-se o fenômeno denominado Engaiolamento, caracterizado pela produção textual fortemente moldada a estruturas formulaicas voltadas à aprovação, mas com redução da expressividade e do potencial argumentativo. A pesquisa demonstrou que a análise estatística e computacional de *corpora* avaliativos revela aspectos estruturais relevantes sobre a escrita em exames de larga escala, além de evidenciar os efeitos de práticas avaliativas sobre o ensino de redação, sobretudo em contextos de cursinhos populares.

Os estudos elucidados, até aqui, recorreram a *corpora* especializados e/ou de aprendizes, construídos com rigor metodológico e com foco em textos de redações dissertativo-argumentativas, como no caso do ENEM ou de vestibulares. Os *corpora* desses estudos tiveram como característica principal o recorte temático e a especificidade do contexto de produção. No entanto, ainda há desafios relacionados à construção e à manutenção de bases de dados textuais mais amplas e acessíveis, especialmente aquelas voltadas para o ensino de língua portuguesa no Brasil. Indubitavelmente, esses esforços demandam investimento acadêmico e institucional; todavia, a ampliação e a atualização de *corpora* deve ser contínua, pois além de facilitar futuras pesquisas na área, podem contribuir significativamente para o redirecionamento pedagógico do ensino da escrita.

Finalmente, cito uma importante obra de referência para os estudos contemporâneos em Processamento de Linguagem Natural, sobretudo em língua portuguesa, o livro “Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português” (2024), organizado por Caseli e Nunes. Trata-se de um trabalho coletivo, que reúne mais de sessenta autores da área, entre docentes, pesquisadores e profissionais; o livro apresenta uma abordagem didática, acessível e atualizada das principais técnicas, tarefas e aplicações de PLN, com foco no português do Brasil. A obra destaca-se por sistematizar o estado da arte da área e por sua proposta inovadora de atualização constante e de livre acesso digital, o que garante tanto a difusão ampla do conhecimento quanto a não obsolescência do conteúdo. No livro, encontramos uma articulação entre fundamentos históricos, perspectivas interdisciplinares e aplicações práticas; ressalto, desse modo, que o

trabalho em questão contribuiu com um embasamento teórico consistente para a discussão sobre PLN nesta tese.

3.2 A LINGUÍSTICA DE *CORPUS*

A Linguística de *Corpus* apresenta-se como uma abordagem interdisciplinar no tocante aos estudos da linguagem, caracteriza-se pelo uso de dados empíricos que são coletados de textos autênticos, de falantes reais. A relevância da Linguística de *Corpus* vai além de uma metodologia de pesquisa, mas apresenta-se como um campo de estudo que dialoga com diferentes teorias linguísticas e com diversas áreas do conhecimento. Nesta seção, faço questão de explorar alguns conceitos fundamentais que sustentam a LC, sob a ótica de diferentes autores; destacarei a sua aplicação, as suas características principais e as suas contribuições para a compreensão do uso real da língua.

.Recorrendo a Conrad e Reppen[10] (2015, p. 14), para conceituar a Linguística de *Corpus,* entendemos que se trata de uma abordagem de pesquisa que facilita as investigações empíricas relacionadas à variação e ao uso da língua. A Linguística de *Corpus* se ocupa do *corpus*, de sua análise e de sua compilação. Por *corpus*, entendemos, como sendo um conjunto de textos que podem ser escritos, falados ou multimodais. Esses textos são coletados de forma criteriosa e mantidos em formato eletrônico, constituindo o alvo da pesquisa linguística.

O termo Linguística de *Corpus*, segundo Sarmento[11] (2010, p. 88), pode ser compreendido como um estudo da linguagem que se baseia em exemplos do cotidiano real. É importante ressaltar que, a Linguística de *Corpus* não pode ser vista como a sintaxe, a semântica e/ou a pragmática, pois é um ramo da Linguística que se concentra na descrição e na explicação de algum aspecto linguístico em uso; ela é também uma metodologia que pode ser aplicada a uma gama de estudos linguísticos, até mesmo ao ensino de línguas. Ou seja, é uma das várias maneiras de se fazer Linguística.

---

[10] Susan Conrad e Randi Reppen foram alunas de Biber. Elas também participam como coautoras de outros trabalhos de Biber (Biber et al. 1994, 1996).

[11] Um dos trabalhos importantes de Simone Sarmento foi: linguística de *corpus*: histórico, metodologia, campos de aplicação (2010); disponível em: https://lume.ufrgs.br/bitstream/handle/10183/140032/000779555.pdf

Biber[12], Conrad e Reppen (1998, p. 4) trazem algumas características cruciais na Linguística baseada em *corpus*: ela é empírica, pois analisa padrões reais de uso em textos naturais, utiliza coletâneas de textos para a análise, faz uso de computadores, usando também técnicas automáticas e interativas, é dependente de técnicas analíticas, quantitativas e qualitativas para gerar conhecimento empírico.

Para Novodvorski e Finatto (2014, p. 7)[13], é importante dizer que a Linguística de *Corpus* se coloca como uma nova perspectiva para a Linguística, mas não como um novo tipo de Linguística. Mostra-se àqueles que se aproximam dela como uma metodologia ou como uma abordagem teórica diferenciada dos estudos da linguagem. Se alguém quiser se aproximar da Linguística de *Corpus* apenas por ter um interesse em seu instrumental ou em seus procedimentos, não será nada cobrado em termo de filiação teórica ou epistemológica, ainda que de modo insistente ressaltemos que a Linguística de *Corpus* é também um modo de compreender a língua, um modo próprio de defini-la como objeto de estudo. Ela é vista como um sistema de combinatórias, no qual as associações mantidas com outras unidades definem uma unidade em específico.

Dando destaque às explorações estatísticas de elementos lexicais e a observação de frequências e de combinações de palavras, temos uma trajetória de estudos realizados no Brasil. São estudos considerados amplos e que podem ser aproveitados em diferentes tipos de pesquisa, servindo hoje, minimamente, para que gêneros textuais sejam caracterizados. Ao longo do processo investigativo quase todos os gêneros textuais escritos foram objetos de alguns estudos em Linguística de *Corpus*.

---

[12] Em 1998, com o livro 'Variation Across Speech and Writing', Douglas Biber apresenta para o grande público a abordagem conhecida por Análise Multi-traço e Multidimensional de Variação de Registro (Multi-feature Multidimensional Analysis of Register Variation), ou simplesmente Análise Multidimensional.

[13] Considerando o percurso da Linguística de Corpus no nosso país desde 2004, os autores, na mesma época do lançamento do livro de Tony Berber Sardinha, refletem sobre a combinação de palavras dirigidas à Linguística de *Corpus*. O trabalho traz por título: Linguística de Corpus no Brasil: uma aventura mais do que adequada; está disponível em: https://www.academia.edu/26421707/Novodvorski_A_Finatto_M_J_2014_Lingu%C3%ADstica_de_*Corpus*_no_Brasil_uma_aventura_mais_do_que_adequada

Svartvik[14] (1996, p. 16) acredita que a Linguística de *Corpus* não é uma definição apenas de metodologia emergente para os estudos linguísticos, porém, é um novo modo de pesquisa e realmente uma nova abordagem filosófica para este assunto. O autor sueco acrescenta que, como uma ferramenta tecnológica de poder inestimável, o computador tornou esse novo tipo de Linguística possível, mas depende dos linguistas, com suas intuições acerca da língua, instruírem tais programas para que evidências linguísticas sejam extraídas.

Stubbs (2001, p. 10)[15], por sua vez, diz que a Linguística de *Corpus* enxerga a linguagem como um sistema probabilístico. Isto é, ainda que possua várias combinações e características possíveis, nem todas são prováveis de ocorrer, a depender da adequação da análise estatística; os *corpora* fornecem informações sobre a frequência de muitos aspectos da língua, ainda que não traga resposta para tudo.

Em sintonia com as inferências precedentes, Berber Sardinha[16] (2004, p. 31), enfatiza que na diferença de frequências entre os traços, o mais importante é que não são aleatórios, se essas diferenças fossem aleatórias, a frequência não seria significativa e não acrescentaria informações acerca da língua e de suas estruturas. Porém, grupos de características linguísticas apresentam variações sistemáticas em textos específicos, variações provenientes de situações comunicativas determinadas e da avaliação sistemática, que pode ser compreendida como a recorrência de traços linguísticos (colocação, coligação, padrão sintático e outros) indicando a padronização da linguagem motivada por diferentes fatores, além das próprias necessidades comunicativas.

Segundo Oliveira (2009, p. 50) a Linguística de *Corpus* pode ser considerada como a face moderna da Linguística empírica, na medida em que enxerga a língua como um fenômeno social e a analisa a partir de atos concretos de comunicação, ou seja, por meio de textos reais, buscando os significados e as negociações que são feitas no discurso. É uma perspectiva própria sobre a linguagem, fenômeno que

---

[14] Jan é conhecido por seu desenvolvimento inicial e inovador de *corpora* legíveis por máquina em colaboração com a Pesquisa de Uso do Inglês na UCL, em particular o primeiro *corpus* falado do mundo, o London–Lund *Corpus* de inglês britânico falado, lançado em meados da década de 1970.

[15] O autor defende a ideia de que partir do momento em que o ser humano tem cada vez mais contato com 'máquinas', torna-se mais necessária ainda a convergência de disciplinas para o desenvolvimento de tecnologias que envolvem a linguagem.

[16] Tony Berber Sardinha possui uma produção profícua; é, sem dúvidas, um pesquisador exponencial da Linguística de Corpus no Brasil.

estuda, e uma maneira específica de fazer pesquisa, por meio do estudo de textos reais, auxiliados por programas de computador cujo objetivo principal é que evidências linguísticas do *Corpus* sejam extraídas; o que nos leva a crer que é um campo de estudos que constitui uma área de conhecimento de bases teóricas próprias e uma maneira de fazer análise linguística específica.

> Na Linguística de *Corpus* trabalha-se com dados reais tão exaustivos quanto possível e que, portanto, possam reproduzir com a máxima fidelidade a realidade linguística. Consequentemente, o volume de dados torna-se tão gigantesco que só uma codificação, ordenação e organização desses dados pode permitir seu uso sem que os pesquisadores fiquem naufragados em um mar de informações que não conseguem manipular. (BIDERMAN, 1999, p. 81).

A Área da Linguística de *Corpus*, acrescenta Oliveira (2009)[17], está em expansão e possui uma história ainda recente se comparada a outras subáreas linguísticas. Todavia, existem fatores que podem acelerar ou reter o seu desenvolvimento. A seu favor, temos o fato de que é uma área altamente relacionada ao uso computacional. Como a tecnologia vem se desenvolvendo aceleradamente, cada dia mais podemos contar com máquinas ainda mais robustas para armazenamento de quantidades ainda maiores, tornando os *corpora* cada vez mais complexos. Mas para analisá-los, programas cada vez mais sofisticados são necessários e para que sejam criados e desenvolvidos dependemos de pesquisa de diferentes áreas que trabalhem em colaboração.

Em continuidade, Halliday (1993, p. 33)[18] traz à tona a oposição criada por alguns pesquisadores entre a Linguística de *Corpus* e a Linguística Teórica, como se fossem divergentes. Há tempos, a Linguística de *Corpus* já era vista como altamente teórica no sentido de modificar os pensamentos sobre o léxico e sobre os padrões vocabulares, inclusive sobre a gramática. Com os novos dados surgidos a partir do *corpus*, pôde-se trazer problemas para as teorias; a dicotomia acaba sendo utilizada

---

[17] Um de seus trabalhos importantes é intitulado com o "Linguística de *Corpus*: teoria, interfaces e aplicações". Nele, a autora apresenta uma visão geral da Linguística de *Corpus*, caracterizando-a como uma área do conhecimento; levando em consideração sua interface com outras áreas; e ilustrando suas aplicações, com foco mais específico no português do Brasil. O trabalho está disponível em: https://www.e-publicacoes.uerj.br/matraga/article/view/27796

[18] Halliday (em consonância com alguns estudos de Vygotsky) concebe a linguagem como "semiótica social", inserida num contexto sociocultural, defendendo que a realidade social (ou uma cultura) é um edifício de significados, ou seja, um constructo semiótico.

entre teoria e dados quando seria mais adequado considerar a complementariedade entre eles, cada lado alimentando e redefinindo o outro de forma constante.

Kaplan (2002, p. 18)[19], reafirmando algumas assertivas, afirma que a Linguística de *Corpus* está estreitamente ligada ao desenvolvimento da Linguística Aplicada, cuja previsão é que possui uma ligação ainda maior com a Linguística Descritiva. Para ele, o desenvolvimento da Linguística de *Corpus* revela fatos sobre usos linguísticos e de variações entre registros, indispensáveis para lidar com questões práticas, muitas vezes incompatíveis com a maioria dos modelos teóricos da linguística.

Por outro lado, Oliveira (2009, p. 13) acrescenta que estudar a partir de *corpus* traz caracterizações pela busca de tendências, probabilidades ou padrões de ocorrência dentro de uma grande quantidade de dados. Os números nesses casos servem de base para que os padrões sejam identificados e para que possam ser interpretados pelos pesquisadores. Os resultados quantitativos produzidos com base em *corpus*, indicam numericamente o que deve ser discutido à luz dos distintos posicionamentos teórico-metodológicos para que sejam entendidos. Do mesmo modo que o *corpus* traz somente evidências linguísticas e não informações, os números que são extraídos dos dados linguísticos também não são informações em si mesmos, precisam ser interpretados pelos pesquisadores para que sirvam de base para outras descrições linguísticas ou para novas propostas de perspectivas teóricas.

No panorama mais atual da Linguística de Corpus, destaca-se a crescente sofisticação dos métodos estatísticos empregados na análise linguística. Gries (2024, p. 20) argumenta que a LC é, por natureza, estatística e distribucional, uma vez que seus pressupostos teóricos se ancoram na frequência e na coocorrência de elementos linguísticos em contextos autênticos de uso. Além da tradicional contagem de frequência de *tokens* e *types* (a explicação desses termos virá subsequentemente, na subseção 3.4), o autor discute a importância de medidas como dispersão, associação, modelos de regressão logística, *random forests* e análises de *cluster*, que ampliam significativamente a capacidade interpretativa dos dados do *corpus.* Gries (2004) também enfatiza a necessidade de considerar a estrutura hierárquica dos dados linguísticos e a quantificação da incerteza estatística, propondo o uso de técnicas de

[19] Kaplan parte da ideia de que a língua deve ser estudada em relação a um contexto, independentemente das escolhas teóricas e metodológicas.

reamostragem como *bootstrapping,* para estimar a variabilidade dos resultados. Esse conjunto metodológico, que vai além de testes simples como qui-quadrado ou correlações bivariadas, confere maiores desdobramentos às análises linguísticas e orienta a produção de inferências mais confiáveis e validadas.

Dessa forma, pode-se sugerir que a Linguística de *Corpus* seja concebida como uma teoria externa complementar, capaz de dialogar e estabelecer interfaces com os pressupostos de outras teorias linguísticas.

> Se considerarmos que uma teoria pode ser entendida como uma perspectiva sob a qual um fenômeno é observado, entenderemos facilmente o porquê de existirem múltiplas teorias de linguagem, que correspondem a diferentes maneiras de se olhar esse mesmo objeto de estudo. (OLIVEIRA, 2009, p. 51).

Como sustentado até aqui, podemos afirmar que a Linguística de *Corpus*, em suas múltiplas interfaces, encontra-se alicerçada na interdisciplinaridade e na complementaridade, estabelecendo relações com diversas áreas do conhecimento, teorias e abordagens linguísticas; corresponde a uma integração de saberes para promover uma compreensão mais aprofundada do objeto comum de estudo: a linguagem. Nesse contexto, é possível identificar pontos de convergência entre a Linguística de *Corpus* e campos como a Lexicologia, a Linguística Computacional, a Linguística Aplicada, entre outras áreas correlatas.

Torna-se relevante enfatizarmos a diferença marcante entre o que a Linguística defende como *corpus* e o que a Linguística de *Corpus* também define. Compreende-se que a disponibilização de *corpora* compilados para outras pesquisas é uma característica que já faz parte do *corpus*, de tal modo que todo o empenho gerado para que seja construído não seja útil apenas para uma pesquisa, já que se tem uma referência padrão de língua ou de variedade linguística que pode servir a outros pesquisadores. Vê-se assim, com base em todas as considerações conceituais feitas, que os dois grandes eixos diferenciadores entre Linguística e Linguística de *Corpus* são o fato de o *corpus* ser computadorizado e estar disponível para outros pesquisadores.

3.2.1 O que é um Corpus? Abordagens e Conceitos na Linguística.

Dando seguimento à análise dos fundamentos teóricos da Linguística de *Corpus*, esta subseção se dedica a uma abordagem detalhada sobre as definições e conceitos relacionados ao termo "*corpus*", tal como argumentou-se na seção anterior, pela perspectiva de diferentes autores. Evidenciarei aspectos fundamentais que delineiam sua composição, objetivos e relevância na pesquisa linguística. Além disso, destacarei como os *corpor* têm ampliado a capacidade de investigação e compreensão da linguagem em uso, promovendo novas possibilidades para o estudo empírico da língua.

O *corpus*, à vista do que já foi elucidado neste texto, é uma coletânea de textos disponíveis em formato eletrônico. Por meio dos *corpora*, identificamos algumas possibilidades de contribuição/vantagens para a prática docente do profissional de línguas, tais como: aprendizagem da frequência das palavras; colocações; frases e fragmentos lexicais; gramática; exemplos autênticos; observação, reflexão e conhecimento do léxico em uso; abordagens descritivas; palavras novas; *corpus* especializado/específico e revolução digital.

Para a constituição de um *corpus*, todavia, alguns aspectos devem ser levados em conta, dentre eles, a sistematização da coleta, a representatividade e a forma como os arquivos serão dispostos. Sabemos que os *corpora* escritos ainda dominam a produção na área, mas a compilação de *corpora* multimodais tem se ampliado de forma rápida. Há uma aplicabilidade crescente não somente nos estudos linguísticos canônicos, mas no desenvolvimento de tecnologias da fala, do reconhecimento e da síntese.

Berber Sardinha (2000, p. 383), traz uma definição completa, elucidando o que um *corpus* precisa conter:

> a) origem: Os dados devem ser autênticos; b) o propósito: o *corpus* deve ter a finalidade de ser um objeto de estudo linguístico; c) a composição: o conteúdo do *corpus* deve ser criteriosamente escolhido; (d) a formatação: os dados do *corpus* devem ser legíveis por computador; (e) a representatividade: o *corpus* deve ser representativo de uma língua ou variedade; (f) a extensão: o *corpus* deve ser vasto para ser representativo.

A afirmação de que todo *corpus* traz questões novas ou questões que não imaginaríamos encontrar é amplamente conhecida, ainda que nenhum *corpus* nos responda tudo. Desse modo, tanto as observações como os experimentos e hipóteses

inicialmente formulados no âmbito da investigação, nos orienta a uma revisão à luz das comprovações e dos resultados.

Consequentemente, a sistematização criteriosa de dados quantitativos e observações qualitativas apresenta-se como componente indispensável ao rigor metodológico da pesquisa; assim, toda teoria resultante de um trabalho consciente sobre os dados conduz o nosso olhar para uma compreensão relevante dos processos de observação, o que é indispensável nas pesquisas com *corpus* e nos diferentes níveis de descoberta.

> Um *corpus* é uma coleção de textos em idiomas que ocorrem naturalmente, escolhidos para caracterizar um estado ou variedade de um idioma. Na linguística computacional moderna, um *corpus* normalmente contém muitos milhões de palavras; isso ocorre porque é reconhecido que a criatividade da linguagem natural leva a uma variedade imensa de expressão. Essa é a dificuldade de isolar os padrões recorrentes que são as pistas para a estrutura lexical da linguagem. (SINCLAIR, 1991, p. 171, tradução minha)[20].

Analisar a língua a partir de *corpus* nos leva a compreender uma abordagem que privilegie o uso de dados; pode-se, assim, estudar o funcionamento de uma língua. Em vez de se recorrer ao conhecimento intuitivo, a observação é sobre como a língua é utilizada por seus usuários. Sendo assim, entendemos que uma linguagem não é inventada, ela é capturada, útil de ser estudada por base de dados textuais, *corpora* oriundos de falantes reais.

Para Dubois *et al.* (1998, p. 21), o *corpus* é considerado como um conjunto de enunciados a partir do qual se estabelece a gramática descritiva de uma língua, ele não pode ser reconhecido como a constituição da língua em si, mas como uma representação amostral dessa língua. Deve ser representativo e capaz de ilustrar uma gama de características estruturais. Se há um indefinido número de enunciados possíveis, não há exaustividade verdadeira no *corpus*. Ademais, a presença de volumosas quantidades de dados irrelevantes para análise não apenas dificulta o processo investigativo, mas também o torna excessivamente oneroso e ineficaz.

---

[20] A *corpus* is a collection of texts in naturally occurring languages, selected to characterize a state or variety of a language. In modern computational linguistics, a *corpus* typically contains many millions of words; this is because it is recognized that the creativity of natural language leads to an immense variety of expression. This poses the challenge of isolating recurring patterns, which serve as clues to the lexical structure of language." (SINCLAIR, 1991, p. 171)

Para Biber (1993, p. 33), o *corpus*, ao ser elaborado, constitui um processo de dois ciclos: inicialmente escolhe-se os textos baseados em critérios externos e culturalmente aceitos e em segundo lugar procede-se com investigações empíricas da língua ou da variedade linguística em análise; finalmente o projeto é todo revisado.

Para McEnery e Wilson (1996, p. 17), o termo *corpus* possui conotações específicas e a sua noção traz características fundamentais, dentre elas, a amostragem e a representatividade. Um *corpus* deve ter uma amostragem suficiente da língua ou variedade linguística que se quer analisar para que o máximo de representatividade dessa língua seja alcançado.

Para Biber *et al*. (1998, p. 18) uma amostragem proporcional não é adequada para um *corpus* linguístico, pois sua organização é demográfica. Contudo, esse tipo de *corpus* não representaria os gêneros textuais, pois poderia conter uma grande parte de conversação, cartas e notas de pedidos, entre tipos de textos jornalísticos, acadêmicos, literários e escritas não publicadas, considerando que a quantidade de pessoas que publicam ou falam para públicos é pequena. O que importa para o estudo de língua é um *corpus* com amostras representativas que trazem em si a variação linguística existente.

Com relação à diversidade, Biber *et al*. (1998, p. 19) dizem que não existe língua geral, cada gênero tem padrões de uso específico. Se um *corpus* serve ao estudo de variação ou representa uma língua, a preocupação deve ser com a diversidade de gêneros, tipos de textos, dialetos e tópicos.

Já para Ducrot e Todorov (2001, p. 12), *corpus* é um conjunto tão variado quanto possível de enunciados que são emitidos por falantes da língua em dada época. Para Trask (2004, p. 86), por sua vez, *corpus* é um conjunto de textos escritos ou falados numa língua com dados disponíveis.

Sinclair (2001, p. 21) propõe o *corpus* como sendo uma coleção de peças de linguagem e textos eletrônicos selecionados de acordo com critérios externos de representatividade e que na medida do possível atenda as variedades linguísticas. Já para Aluísio e Almeida (2006, p. 170), o que mudou com o passar dos anos foi a concepção de *corpus* e essa mudança pode estar atrelada à concepção que se adquiriu da Linguística de *Corpus.*

> Geralmente, aumentando-se o tamanho do *Corpus* aumenta-se também a sua representatividade, pois sua dimensão se aproxima à população, por outro lado, aumentar o tamanho apenas pode não significar um ganho de

> representatividade, caso o *Corpus* não respeite a distribuição da população. Por exemplo, se na população de textos jornalísticos existam 10% de editoriais e 20% de notícias e no *Corpus* constem apenas editoriais (100%), esse *Corpus* não será representativo daquela população, mesmo que cresça exponencialmente. Ele poderá chegar a ser representativo do gênero 'editorial', mas não de 'jornais' como um todo. (BERBER SARDINHA, 2010, p. 326).

É necessário admitir, pensando no formato computadorizado do *corpus*, que o surgimento do computador atingiu de modo direto não só a concepção que se tem de *corpus*, hoje em dia, como também a sua forma de armazenamento e manuseio, já que as ferramentas oferecidas pelo computador abrem caminho para que uma quantidade inimaginável de textos possa ser processada na tela em segundos, fazendo com que inúmeras hipóteses sobre alguns fenômenos linguísticos possam ser atestadas de modo rápido e eficiente. Essa nova forma de armazenar textos permitiu que fenômenos linguísticos antes imperceptíveis pudessem ser observados e descritos, o que não era possível antes, por contarem apenas com recursos manuais.

Outro aspecto importante é o formato eletrônico; quando se emprega o termo *corpus*, automaticamente se admite que os textos estejam no formato eletrônico, diferentemente da concepção tradicional de *corpus*, que se restringia exclusivamente a textos impressos. O formato eletrônico de textos possui vantagens consideráveis aos *corpora,* pois os textos podem ser manuseados de forma mais ágil e podem ter agregados em si, informações extras. Como referência padrão, acredita-se que o *corpus* deve estar disponível para outros pesquisadores, o que chamamos de reuso do *corpus*, por constituir uma referência padrão para variedade de língua que representa.

Para Trask (2004, p. 16), a partir do uso de *corpus* pode-se fazer observações precisas, trazer informações confiáveis, exemplos de opiniões e julgamentos acerca dos fatos de uma língua. Desse modo, aspectos morfológicos, sintáticos, semânticos e discursivos podem ser observados*;* o que é bastante relevante para uma pesquisa linguística. A produtividade e o emprego das palavras, as expressões e formas gramaticais também podem ser explicadas, novos fatos na língua, não perceptíveis, intuitivamente podem ser descobertos. Em suma, a língua pode ser descrita de forma objetiva.

No entanto, como afirma Berber Sardinha (2010) há um conjunto de requisitos que devem ser observados para o projeto de um *corpus* computadorizado. Esses requisitos impactaram de modo direto na validade e confiabilidade da pesquisa baseada em *corpus,* são eles: a autenticidade, a representatividade, o balanceamento, a amostragem e a diversidade de tamanho. Os textos devem ser autênticos, escritos em linguagem natural e não produzidos com alvo de serem pesquisados linguisticamente, devem ser escritos por falantes nativos a menos que se trate de *corpus* de aprendizes, aqueles *corpora* cujos textos são oriundos de falantes que estão aprendendo uma língua estrangeira.

Aluísio e Almeida (2006, p. 165) enfatizam que existem muitos *corpora* disponíveis de forma livre ou paga, a partir desses, subcorpos de estudo podem ser gerados, dependendo da questão de pesquisa, ainda pode ser necessário compilar um *corpus* próprio. Para que seja compilado existem três estágios: o projeto do *corpus* que inclui a seleção dos textos e os cuidados com os requisitos que já foram mencionados; compilação ou captura, manipulação e nomeação dos arquivos textuais, permissão de uso; e anotação.

Com relação ao projeto de *corpus*, inicialmente seleciona-se os textos pertinentes e relevantes para a pesquisa. É preciso definir que tipo de *corpus* será compilado, outras decisões também são importantes com relação à composição e aos gêneros que farão parte. Existem várias tipologias de *corpus* que indicam os parâmetros importantes a serem considerados. Para o processamento computacional, o *corpus* precisa ser preparado, precisa ser limpo e formatado, isso significa tirar imagens, gráficos, tabelas e demais anotações que não estão no texto propriamente dito. Essa limpeza e formatação possibilitam que o *corpus* seja processado por ferramentas computacionais.

Com relação à anotação, Aluísio e Almeida (2006, p. 178) dizem que os níveis de representação das informações presentes no *corpus* são estruturais. Em Linguística, a anotação estrutural refere-se à marcação de dados externos e internos dos textos. Os dados externos são a documentação do *corpus* em forma de cabeçalho que inclui os metadados textuais, bibliografia, catalogação, autoria, tipologia textual e a distribuição do *corpus.* Já os dados internos podem ser entendidos como a anotação de segmentação do texto cru que envolve a estrutura geral, os títulos e subtítulos, notas de rodapé, elementos gráficos com tabelas e figuras ou subparágrafos. A anotação linguística pode se dar em qualquer nível, morfossintático, sintático,

semântico e discursivo e se encerra em três formas: manual, automática ou semiautomática.

> Os *corpora* podem ter tamanhos variados, ser compilados para diferentes propósitos e ser compostos por textos de diferentes tipos. Todos os *corpora* são homogêneos até certo ponto, pois são formados por textos de uma única língua, de uma variedade específica de uma língua ou de um único registro, entre outros critérios. Ao mesmo tempo, também são heterogêneos em certa medida, uma vez que, no mínimo, são compostos por vários textos distintos. A maioria dos *corpora* contém informações adicionais além dos próprios textos que os compõem, como metadados sobre os textos, marcações de classe gramatical para cada palavra e informações de análise sintática." (FLOWERDEW, 2009, p. 402, tradução minha).

Um dos riscos associados ao uso de *corpus* é a possibilidade de os dados não serem selecionados com critérios suficientemente rigorosos, comprometendo a confiabilidade dos resultados obtidos e a sua interpretação como declarações precisas sobre uma língua ou variedade. A experiência do pesquisador torna-se essencial para a análise e a avaliação adequada dos resultados. Cabe à comunidade de linguistas, ao utilizar *corpus*, decidir se os méritos específicos de um dado *corpus* justificam sua preservação e o uso de terminologia tradicional, ou se as tecnologias de criação eletrônica de textos representam uma mudança tão significativa que a concepção de um *corpus* pode, em breve, se transformar em uma limitação, restringindo a busca por informações conclusivas.

Por um lado, como apontam Bonelli e Sinclair (1996, p. 216), há uma sensação de segurança e de estabilidade em se possuir "um *corpus* de língua X". Argumenta-se que a compilação de um dicionário baseado em *corpus*, por exemplo, gera diversos impactos positivos, considerando que os principais dicionários serão fundamentados em *corpus*. Por outro lado, o trabalho de compilação de um *corpus* continua sendo substancialmente desafiador, apesar dos avanços significativos na tecnologia ao longo das últimas décadas. Mesmo em sua concepção mais básica, ainda faltam estudos sobre aspectos fundamentais, como as proporções ideais de linguagem escrita em um *corpus* de amostra que busque representar o estado atual de uma língua.

De uma perspectiva intelectual global, então, um *corpus* parece ser uma entidade bastante frágil; fustigada por muitas forças mais poderosas de áreas como a 'engenharia da linguagem'. Uma vez que as línguas são centrais para a comunicação internacional, há um interesse contínuo e um apoio financeiro de algumas empresas de base política e de defesa de fontes, que podem atravessar as prioridades da

ciência. Em última análise, um fator que pode distinguir um *corpus* é o fato de que foi concebido para refletir o mais fielmente possível o tipo de linguagem a partir do qual é selecionado.

3.2.2 A questão da extensão

Ao abordar o tema da extensão de um *corpus*, é fundamental considerar que o tamanho ideal varia de acordo com os objetivos específicos de cada pesquisa. Esta subseção explora os parâmetros que influenciam a definição do tamanho de um *corpus*, destacando como diferentes abordagens e necessidades metodológicas determinam a sua amplitude. Além disso, serão discutidos os desafios relacionados à representatividade e à eficiência no uso de *corpora*, considerando as limitações impostas por ferramentas computacionais e a diversidade de conteúdos possíveis.

Segundo Hoffmann (1998, p. 25), o tamanho mínimo necessário para o *corpus* nas pesquisas linguísticas diverge de modo amplo. Na pesquisa de linguagem especializada, alguns resultados úteis já foram obtidos com amostras de 35 mil palavras, mas sugere-se 200 mil palavras como mínimo. O autor defende que esse tamanho vai depender dos objetivos da pesquisa e do tipo de *corpus*, assim, não se trata de uma fórmula matemática aceita que informa a quantidade e distribuição de palavras que um *corpus* deve ter para ser representativo, contudo, a maioria das palavras têm frequências de ocorrência muito baixas e para que elas apareçam em um *corpus*, precisa-se que ele tenha um grande número de palavras. Podemos dizer o mesmo sobre os diferentes sentidos ou significados de uma mesma palavra, há os mais e os menos frequentes, os sentidos mais raros terão uma probabilidade maior de aparecer em um *corpus* maior.

Segundo Sarmento (2010, p. 87), o tipo de conteúdo que um *corpus* deveria conter não é especificado. Um *corpus* pode conter desde a obra completa de Shakespeare até instruções expressas de uma caixa de sabão em pó, ou um texto jornalístico sobre time de futebol, quando foi campeão brasileiro. Com relação à sua dimensão, também não há um consenso no que se refere ao seu tamanho mínimo ou máximo.

Outro aspecto influenciador do tamanho de um *corpus*, diz Sarmento (2010, p. 85) é o objetivo da pesquisa. Para estudar gramática da linguagem falada, por

exemplo, *corpora* relativamente pequenos podem ser suficientes, pois as palavras gramaticais tendem a ser muito frequentes. Por outro lado, itens de baixa frequência vão exigir um *corpus* bem maior, geralmente os *corpora* coletados com base em um projeto de pesquisa linguística específica, tal como o fornecimento de informações sobre frequências para dicionários e para verbetes ou para produção de material didático para o ensino de língua. Muitas vezes, porém, os *corpora* coletados sem que haja um propósito específico, ficam disponíveis como recurso da língua geral para linguistas, professores de línguas, lexicógrafos, entre outros. Existem vários tipos de *corpora* dependendo do tamanho da proposta e da forma como foram compilados, assim como veremos na subseção 3.2.4.

3.2.3 A questão da representatividade

A representatividade de um *corpus* é um aspecto central na pesquisa linguística, pois determina o quão bem ele reflete as características de uma língua ou de uma variedade específica.

Existe o *corpus* geral contendo muitos tipos de texto. Por ser de cunho geral, provavelmente esse *corpus* não será representativo de nenhum todo e ainda um *corpus* geral precisa ser muito maior do que um *corpus* específico. Grande parte das vezes, ele é utilizado em contraste com relação aos *corpora* mais especializados. Há também o *corpus* monitor que é projetado para que mudanças atuais em uma língua sejam verificadas. Ele é alimentado anualmente, mensalmente ou até mesmo diariamente; cresce com muita rapidez, mas a proporção do tipo de texto é constante, de modo que cada período possa ser comparado com o ano anterior. O "*Corpus of Contemporary American English*" (COCA)[21], atualmente, possui mais de um bilhão de palavras.

[21] COCA (*Corpus* of Contemporary American English). Esse *corpus* possui mais de um bilhão de palavras em textos de oito gêneros: ficção, jornais, revistas populares, blogs, textos acadêmicos, legendas de TV e filmes, e outras páginas da web. Em seu layout estão dispostas as ferramentas chamadas search (busca), frequency (frequência) e context (contexto).

Figura 13 - interface do *Corpus* of Contemporary American English

Fonte: https://www.english-*corpora*.org/coca/
Legenda: a imagem nos mostra a interface do *Corpus* of Contemporary American English - COCA

Sarmento (2010, p. 83) ressalta que, quanto à representatividade, é essencial considerar a totalidade do fenômeno linguístico, embora, no caso da linguagem, essa totalidade, por vezes, permaneça desconhecida. Deve-se tentar dividir esse todo estimado, em partes. Por exemplo, um *corpus* de linguagem jornalística deve trazer vários tipos de jornais, desde os populares aos mais tradicionais, devem incluir também textos de diferentes seções, como esportes, negócios, editoriais, para que seja considerado como representativo e equilibrado. Deve-se incluir o número aproximado de palavras em cada categoria, negócios nos populares, negócio nos tradicionais, esporte nos populares e assim sucessivamente.

3.2.4 Tipos de corpus

Sarmento (2010, p. 85) acrescenta que há ainda o *corpus* comparável: dois ou mais *corpus* em línguas diferentes ou em variedades de uma língua, eles são compilados com as mesmas diretrizes, isto é, conterão a mesma diversidade de gêneros. O exemplo mais citado é ICE (International *Corpus* of English) que contém mais de um milhão de palavras. Há também o *Corpus* paralelo (dois ou mais *corpus* em línguas diferentes), contendo os textos que foram traduzidos de uma língua para outra ou que foram produzidos de forma simultânea de duas ou mais línguas.

Figura 14: Os principais tipos de corpora (Berber Sardinha, 2004)

Fonte: autoria própria

Além dos tipos de *corpus* citados acima, Hunston (1998, p.14) adiciona ainda os seguintes:

- *corpus* aprendiz: coletânea de textos e redações produzidos por aprendizes de uma língua cujo propósito é identificar em que aspectos esses aprendizes se diferem entre si e em relação à falantes nativos;
- *corpus* pedagógico: consiste na linguagem exposta ao aprendiz, podendo estar em livros didáticos e gravações;
- *corpus* histórico diacrônico: de diferentes períodos, utilizado para ver igualmente o desenvolvimento de diferentes aspectos de uma língua ao longo dos anos;
- *corpus* especializado: um *corpus* contém um tipo específico de texto ou gênero, como os resumos (*abstract*), artigos acadêmicos sobre um assunto específico, conversas telefônicas etc. É um tipo de *corpus* que tem por objetivo ser representativo de certo tipo textual de linguagem.

3.2.5 LEXICOLOGIA E LINGUÍSTICA DE *CORPUS*

No capítulo dois deste trabalho, fiz alusão à Linguística de *Corpus* e aos seus pontos de convergência com outros campos de estudo, dentre eles, a lexicologia. A menção a essa área também se justifica pelo fato de que a análise do *corpus,* realizada por meio da ferramenta adotada nesta pesquisa, oferece-nos indicadores consistentes acerca do léxico. Por conta disso, no encadeamento desse raciocínio,

e por considerar essencial os estudos do léxico aliados à análise de *corpus*, trarei uma tão breve consideração acerca da Lexicologia.

A lexicologia pode ser vista como uma área do conhecimento oriunda da própria linguística, a partir do entrelaçamento de conceitos teóricos advindos de diversas subáreas linguísticas. Quando entendida como disciplina, a lexicologia se ocupa do léxico das línguas, de forma completa e integrada (Lorente, 2004). O léxico, por sua vez, está situado na intersecção linguística que engloba informações advindas de diferentes caminhos, isto é, da fonética, da fonologia, da pragmática, da sintaxe, da semântica, da morfologia e das situações comunicativas.

Costa e Silva (2020, p. 36) nos trazem uma pesquisa acerca da riqueza lexical de redações escritas em língua portuguesa. Um trabalho baseado em cálculos estatístico-lexicais para possibilitarem uma análise quantiqualitativa. O principal intuito foi o de avaliar o conteúdo lexical, entendendo que o léxico mereça mais destaque nas aulas de língua portuguesa, não apenas nos anos iniciais de escolarização ou no Ensino Médio, mas, sim, ao longo da formação como uma prática sistemática.

Tal pesquisa, com o propósito de descrever, analisar e problematizar o repertório lexical em redações escritas em língua portuguesa, trouxe contribuições aos estudos vinculados à Lexicologia e à Linguística de *Corpus* não apenas para os professores da licenciatura em Letras, mas, também, para os professores já atuantes e para professores em formação inicial. O *corpus* da pesquisa em pauta foi constituído por 199 redações escritas em um processo seletivo para acesso ao Ensino Superior. Para a análise do *corpus* foi utilizado um programa de computador muito útil na operacionalização da descrição em Linguística de *corpus*, o *WordSmith Tools.*

Costa e Silva (2020, p. 37) utilizaram o método *type-token ratio*: a razão forma/item (ou vocábulo/ocorrência) expressa em porcentagem. O total de formas foi dividido pelo total de itens dividido por cem. Como por exemplo, na frase, "*o gato viu o rato*", o valor da razão fornecido pelo programa seria 8,0. Valor obtido assim: 4 ÷ (5 ÷100). Em qualquer *corpus* linguístico são perceptíveis algumas constantes na

distribuição das palavras. Dessa forma, observando-se os valores de riqueza lexical, analisou-se na pesquisa, o conteúdo presente nas redações que constituíram o *corpus.*

De acordo com os resultados obtidos por Costa e Silva (2020, p. 38), o aumento de insumo lexical que o aprendiz tem pode impactar em seu desempenho em uma redação, por exemplo – a depender da exposição e emprego consciente do léxico. Os autores afirmam, nesse sentido, que "é oportuno que esses indivíduos aumentem seu repertório lexical ao longo da formação no que se refere às habilidades de compreensão e produção porquanto sua produção linguística tende a ser avaliada, especialmente no contexto escrito." (Costa; Silva, 2020, p. 39)

Os resultados obtidos com a pesquisa mostraram que as escolhas lexicais das redações foram feitas para atender a um determinado propósito comunicativo. Costa e Sila (2020, p. 40) salientam que, ainda que o conhecimento lexical seja vasto, é necessário que o pensamento seja organizado de tal modo que ocorra um desencadeamento coerente das estruturas linguísticas na modalidade escrita do texto.

Segundo Biderman (2003, p. 64), é pelo léxico de uma língua natural que se registra o conhecimento do Universo, a partir de nomeações e classificações de objetos e seres. Sendo assim, nomear a realidade constitui o primeiro passo para a aquisição do conhecimento científico. Nos agrupamentos e categorizações feitas, por traços semelhantes ou distintos, alguns objetos foram sendo individualizados e o homem estruturou o mundo no qual vive.

De acordo com a mesma autora (1999, p. 179), qualquer sistema léxico é a somatória de toda a experiência acumulada de uma sociedade e do acervo da sua cultura através das idades. As mudanças sociais acarretam alterações nos usos vocabulares e o universo semântico se estrutura em torno de dois polos opostos: o indivíduo e a sociedade. O léxico das línguas naturais originou-se, portanto, desse processo de nomeação pelo qual o indivíduo apropria-se do real. Atos sucessivos de

cognição desse real, das categorizações e das experiências foram cristalizados em signos linguísticos, que podem ser chamados de palavras.

3.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Os conceitos basilares do conjunto de procedimentos desta pesquisa, como elucidado, situam-se também no processamento de linguagem natural (PLN). Por esse motivo, alguns apontamentos são essenciais.

A partir da década de 90 (*vide* seção 2) adquire um papel crucial nas pesquisas linguísticas, o uso dos *corpora*. É nessa época também que as contribuições da Linguística de *Corpus* e da linguística computacional começam a ser datadas. Podemos destacar de modo principal, o desenvolvimento de ferramentas computacionais para o processamento de língua natural (PLN), que tiveram um efeito substancial para o processamento de *Corpus* .

Antes de adentrar na análise dos dados e na aplicação das métricas textuais, considero fundamental oferecer um panorama histórico e conceitual do Processamento de Linguagem Natural (PLN), pois é esse campo que fornece a ferramenta computacional na presente pesquisa. O entendimento da trajetória do PLN, desde seus primórdios simbólicos até os modelos neurais contemporâneos, permitirão compreender as mudanças de paradigma que impactaram diretamente a forma como a linguagem tem sido tratada computacionalmente. Acredito que ao expor os princípios, os limites e as evoluções desses paradigmas, justifico o uso da NILC-Metrix como ferramenta de análise textual e o percurso metodológico adotado nesta investigação.

Ressalto também que, para garantir a compreensão dos mecanismos envolvidos no processamento linguístico-computacional, apresentarei, ao longo da seção, diversos conceitos fundamentais, tais como morfema, lexema, lema, léxico, gramática, palavras funcionais e lexicais, entre outros. Essa retomada teórico-conceitual servirá para demonstrar a pertinência e os fundamentos científicos da abordagem escolhida.

3.3.1 Histórico e panorama da área do PLN

Até meados de 1980, o paradigma simbólico era a base do PLN. Acreditava-se que todo o conhecimento acerca da língua deveria ser explicitado por formalismos lexicais, por regras e por linguagens óbvias. Isto é, por formas que são compreensíveis ao ser humano. A exemplo disso, temos a possibilidade de escrever regras que determinem a concordância entre o gênero gramatical de um substantivo e o gênero atribuído ao adjetivo que o acompanha. Desse modo, exemplos como “maçã madura” serão considerados corretos a partir da regra, ao passo que “maçã maduro”, não.

Nos anos 1990, incipientemente, o paradigma estatístico foi sendo acoplado às máquinas, que ganharam maior capacidade de memória e de processamento, a partir da proposta de diversos algoritmos de aprendizado. Grandes conjuntos de textos (*corpora*) passaram a ser utilizados como fonte de conhecimento para “ensinar” as máquinas. Para ilustrar isso, fenômenos como a concordância entre substantivo e adjetivo passaram a ser aprendidos por meio da ocorrência no *corpus*, tais como: “mamão estragado”, “morango maduro”, “alecrim amarelo”.

A luz desse viés, entende-se que a língua é, portanto, representada em modelos probabilísticos aprendidos pela frequência de ocorrência. Probabilidades calculadas a partir dos exemplos geram regras explícitas ou implícitas; a utilização desses modelos está na classificação, na tradução ou na geração de novos textos. Partindo do pressuposto de que esses modelos são aprendidos por dados reais, eles possuem uma significativa probabilidade de representarem bons modelos de língua.

Com o passar dos anos, as máquinas mantiveram-se num patamar de ganho de poder de memória e de processamento, resultando em maiores possibilidades do tratamento de grandes quantidades de dados, cada vez mais complexo; a exemplo disso, temos as redes neurais profundas *(deep learning),* que também se baseiam em uma quantidade exponencial de volume de dados para a aprendizagem de um modelo. Entretanto, a forma de aprendizado é diferente, haja vista que há o envolvimento de uma série de camadas de unidades de processamento para que padrões recorrentes sejam reconhecidos.

Nesse sentido, enquanto em técnicas de aprendizado de máquina *(machine learning)* tradicionais, os algoritmos determinam como o aprendizado deve acontecer; no *deep learning* , não se pode inferir com exatidão com que base o modelo foi aprendido; isso, em virtude da complexidade das arquiteturas existentes nas diversas camadas de processamento.

No paradigma neural, distintamente do paradigma simbólico, o conhecimento da língua pauta-se em valores numéricos, não por símbolos ou regras. Sendo assim, tanto o conhecimento linguístico quanto a parte do código que tenha produzido um comportamento específico, tornam-se praticamente irrecuperáveis, resultando em uma opacidade do código e de seu efeito, não determinístico.

Frente ao exposto, é possível notar que o PLN tem caminhado lado a lado com a evolução de paradigmas da IA (Inteligência Artificial): simbólico, estatístico e neural. Contudo, frente à realidade da insuficiência de uma única abordagem, os paradigmas híbridos (que combinam o simbólico aos demais) tem se expandido. É importante ressaltar que o hibridismo desse último paradigma é o que garante algum tipo de aplicabilidade das etapas seguidas pelos algoritmos.

Dando continuidade a algumas questões introdutórias, teço, a seguir, considerações importantes para os que se aplicam aos estudos acerca do PLN: é essencial esclarecer que no PLN podem ser aplicadas uma gama de abordagens, desde as que estão associadas ao paradigma tradicional (simbólico) quanto aos paradigmas mais recentes, o estatístico e o neural.

As estratégias automáticas de processamento da língua possuem limitações; nesse contexto, existe alguns fatores que definem a escolha da melhor estratégia, a saber: o suporte especializado, o poder computacional e a disponibilidade de uma quantidade significativa de recursos linguísticos (*corpora*). A maior parte das estratégias trabalha com processamento de caracteres e não de unidades linguísticas, ou seja, diversas estratégias podem gerar modelos a partir de coocorrências e de contextos de ocorrências de palavras e frases. Assim, é possível considerar que as estratégias utilizadas em grande parte das aplicações do PLN não aprendem a língua em suas nuances, mas aprendem a reproduzi-la, e por algumas vezes, a extrapolar aquilo que aprenderam com um pouco de treinamento (princípio de generalização). Consoante a isso, como temos visto, são diversos anos de pesquisa de desenvolvimento em PLN; algumas abordagens surgem, fundamentam-se, outras são relegadas; o crucial é que a linguagem natural possui uma complexidade de aprendizagem de compreensão que ultrapassa a contagem de frequências e de coocorrências.

Consequentemente, não obstante grande empenho que tem sido empregado nos métodos neurais, tanto na sua investigação quanto na sua evolução, ainda não houve uma completa representação/captura do conhecimento linguístico e da língua

em uso por nenhum dos métodos atuais. Uma aproximação disso talvez se deva ao formalismo de representações híbridas, quem integram estruturas semânticas explícitas mais robustas do que as utilizadas pelos métodos estatísticos e neurais.

Entendendo que a fala é uma modalidade de língua extremamente acessada pelo PLN, faz-se necessário ainda, nesse texto, adquirirmos um panorama acerca das habilidades e dos métodos recorrentes no universo do processamento computacional da língua falada. O processamento da língua falada depende de uma amplitude de conhecimentos nos quais estão inseridos: a acústica, a fonologia, a fonética, a linguística geral, a semântica, a sintaxe, a pragmática, as estruturas discursivas, entre outras. Ainda mais, existem outros conhecimentos mais comuns à ciência da computação que são necessários.

Desde o surgimento da interação falada entre os seres humanos, até os dias hodiernos, – e podemos afirmar com tranquilidade, que assim também será no futuro imaginável –, a fala tem sido um instrumento imprescindível para a conjuntura social. É por meio da fala que as emoções são expressas, a nossa atitude referente aos fatos e eventos, bem como as negociações de ideias e de ações. É a fala que nos identifica como seres humanos.

Há uma expansão, hoje, da língua falada que extrapola a interação direta para possibilidades telefônicas, televisivas, computacionais e multimodais. Em sua fase preliminar, havia uma limitação no processamento de língua falada em virtude da escassez de recursos computacionais e de técnicas adequadas. As abordagens iniciais baseavam se em regras gramaticais e em uma acústica superficial. Todavia, com progresso tecnológico e com a amplitude do poder computacional, houve o surgimento de abordagens e de inovações técnicas que culminaram em desenvolvimentos significativos na área, como dito.

É preciso ressaltar, contudo, um marco importante no processamento de língua falada em português: a introdução dos sistemas de síntese de fala. Tais sistemas permitem que uma máquina gere a fala humana, tendo como ponto de partida, o texto escrito em português. Embrionariamente, técnicas de concatenação (processo de união ou de combinação de diferentes partes ou segmentos de fala para a formação de uma sequência contínua ou mais extensa de palavras ou frases) eram a base da síntese de fala, elas envolviam a gravação de segmentos de fala de um locutor humano e a associação desses segmentos para que a fala sintetizada fosse criada.

Os desenvolvimentos mais atuais no processamento da fala em português estão associados à aplicação de modelos de linguagem neural, como os modelos de transformação sequencial e as redes neurais convolucionais recorrentes. Ademais, em decorrência do surgimento dos assistentes virtuais e dos sistemas de processamento de linguagem natural, tornou-se cada vez mais comum, a interação por meio da fala em português. São necessárias capacidades tanto de reconhecimento quanto de síntese de fala para o estabelecimento de um sistema computacional para a língua falada. No entanto, tais componentes são insuficientes para a construção de um sistema útil. Um componente de compreensão e diálogo é crucial para que haja a interação com o usuário; para que a interpretação da fala seja guiada e mapeada é necessário o conhecimento de domínio, e para todos esses componentes, existe uma gama de desafios, dos quais podemos destacar a flexibilidade, a facilidade integrativa e a eficiência de engenharia.

3.3.2 PLN e a sua relação com algumas áreas da LA

Dando continuidade à trajetória do processamento de linguagem natural, há outro ponto importante sobre o qual temos de tratar: a identificação da unidade mínima, quando a língua é tratada computacionalmente. Não há um consenso entre pesquisadores e profissionais de PLN no que condiz com essa delimitação. Até mesmo nas áreas da linguística existem impasses relacionados aos parâmetros de definição, como exemplo disso, a definição de palavra ou de frase. Para subáreas especializadas dos estudos linguísticos, a unidade mínima de processamento é entendida como elementos distintos relacionados aos seus focos e aos pontos de análise. A fonologia, por exemplo, considera que o fonema é a menor unidade sonora e distintiva da língua. A morfologia , por sua vez, considera o morfema como a menor unidade na língua que é dotada de significado.

De modo paralelo, podemos compreender que geralmente o PLN serve-se de modelos que trabalham as palavras como unidade primária de processamento. A depender do critério utilizado e da finalidade com a qual as referências de unidades de partes são buscadas, a análise é traçada. É importante salientar que, o texto a ser tratado computacionalmente parte de um processo que envolve a delimitação de unidades linguísticas essenciais para a tarefa para a qual se propôs. Por exemplo, ao definirmos basicamente o termo “palavra” em português, tanto os espaços quanto a

pontuação serão considerados. Entretanto, considerando o desenvolvimento de modelos computacionais, é importante refletir sobre o conceito “palavra”, as variações que possuem e como o pré-processamento pode influenciar nos resultados alcançados.

Sabemos que os vocábulos configuram elementos essenciais da linguagem, eles articulam as representações mentais do sujeito com o universo que o circunda. Isso nos remete a definição do léxico que, por sua vez, corresponde ao conjunto de termos de uma língua que permite a expressão de ideias. No contexto computacional, nesse sentido, podem ser realizadas adaptações necessárias para que palavras ou expressões sejam processadas eficientemente pelos algoritmos. À luz disso, temos o fato de que a complexidade da linguagem natural desafia as máquinas devido à ambiguidade e às diversas interpretações possíveis. É por esse motivo que uma gama de estratégias surge a cada dia, tais como: a criação de expressões unitárias para frases que são úteis para análises e tarefas específicas, como tradução ou classificação gramatical.

Um conceito importante que não devemos deixar de lado é o de “palavra computacional”, que segundo Finatto *et al.* (2024, p. 2):

> se refere a uma unidade linguística que foi adaptada ou criada especificamente para facilitar seu processamento por máquinas. Isso pode envolver a manipulação de palavras, frases ou até mesmo caracteres de maneira que seja mais conveniente para algoritmos e sistemas de PLN lidarem com elas.

Vê-se, assim, que a palavra computacional se refere a uma unidade linguística tratada de forma padronizada para facilitar a análise automatizada de linguagem por ferramentas de Processamento de Linguagem Natural (PLN). Nesse contexto, o termo frequentemente envolve procedimentos como a tokenização, que é o processo de segmentar o texto em unidades mínimas significativas. Uma consequência direta disso é a geração dos chamados *subtokens,* que são frações ainda menores derivadas de uma palavra, produzidas quando sistemas de PLN, como os baseados em modelos de linguagem pré-treinados, realizam decomposição de formas complexas ou raras em unidades mais frequentes. Por exemplo, uma palavra composta ou um termo com sufixos pode ser segmentado em dois ou mais subtokens.

Figura 15 - Demonstração de subtokens

	Forma	Descrição	Subtokens
1	ferro	Palavra-base, substantivo comum	['ferro']
2	ferreiro	Derivado de 'ferro' com sufixo '-eiro', indica profissão	['ferro', 'eiro']
3	ferragem	Derivado de 'ferro' com sufixo '-agem', indica local ou conjunto	['ferro', 'agem']

Fonte: autoria própria

A escolha por esse tipo de representação linguística busca otimizar a aprendizagem estatística e reduzir a escassez de dados linguísticos raros. Assim, palavras computacionais e *subtokens* constituem a base da estrutura lexicográfica utilizada por sistemas computacionais contemporâneos, permitindo que eles operem sobre a linguagem de modo mais eficaz e escalável, especialmente em análises voltadas à frequência, coocorrência e padrões sintático-discursivos de grandes *corpora.* Mas por que essas palavras computacionais precisam existir? Elas existem em decorrência das complexidades do PLN por máquinas.

Sabemos que a linguagem humana possui nuances difíceis de serem capturadas e analisadas automaticamente. Nesse sentido, quando há uma simplificação ou padronização de palavras, as tarefas podem ser executadas pelos sistemas de PLN; a análise gramatical, a extração de informações e a tradução eficaz são algumas dessas tarefas.

A definição das rotinas de pré-processamento é dependente do objetivo da tarefa ou da aplicação de PLN. São os objetivos da tarefa que indicarão quais são as rotinas mais produtivas para que as palavras computacionais sejam geradas, isto é, pode ocorrer a remoção ou não de acentos ortográficos, de espaços em branco, como em “fim de ano” ou em “guarda-roupa”. Nesse mesmo sentido, a depender da intencionalidade da tarefa, pode ser que o modelo de aprendizado de máquina precise da definição de expressões multipalavras , como por exemplo: “quem me dera”,

hashtags, URLs e outros compostos como palavras (ou unidades lexicais) únicas, como se fossem representadas sem espaços.

A partir de então, apresentarei conceitos que estão relacionados às unidades mínimas de processamento e que devem ser reputados quando nos propomos a lidar com a análise de textos. Subsequentemente, apresentarei algumas tarefas relacionadas ao processamento morfológico dos textos. Considero pertinente, antes de discutirmos acerca de como ocorre a identificação computacional das unidades mínimas de processamento, a definição de alguns conceitos linguísticos básicos e indispensáveis, que estão de todo modo atrelados a ferramenta computacional em si, que está sendo utilizada no estudo.

Aqui, como mencionei na subseção 3.3, definirei os conceitos de morfema, lexema, lexia, lema, léxico e gramática, léxico comum e especializado; ademais, abordarei acerca de palavras funcionais e lexicais do processo de formação de palavras, da morfologia e da morfossintaxe. Ainda que, à primeira vista, a quantidade de conceitos, paradigmas e aplicações que serão apresentados sobre o Processamento de Linguagem Natural possa parecer extensa, tal escolha é intencional e justificada pela complexidade da própria área. Considero que, por esta tese se propor a analisar um conjunto de métricas linguísticas geradas automaticamente, é imprescindível tratar, com a devida profundidade, as bases teóricas, metodológicas e tecnológicas que sustentam esse campo interdisciplinar.

3.4 A MORFOLOGIA NO PLN

A morfologia tem como objeto de estudo central, o morfema, que é definido linguisticamente como a menor unidade da língua, dotada de significado. Todavia, há de se considerar que existem outras unidades linguísticas que possuem significado, a saber: a palavra, o sintagma, a frase, a oração, o período, o texto etc. Ademais, considerar o morfema como dotado de significado é uma oposição ao fonema, que corresponde ao objeto de estudo da fonética e da fonologia, correspondendo a menor unidade de análise linguística, entretanto não possui significado em si mesmo; a função do fonema é estabelecer a distinção de significado entre uma palavra e outra.

Dito de outra forma, o fonema é somente uma unidade linguística distintiva (sem significado), uma vez que distingue as palavras a partir de seus sons. Em uma explicação simplificada é possível dizer que os morfemas são os pedaços de

estruturas (desinência, raiz, radical, afixo, vogal temática e tema) que juntos formam as palavras. Por exemplo, podemos juntar o radical “ compr ” com a vogal temática “a” com a desinência verbal “ri” e com a desinência verbal “a” para formar a palavra “compraria”.

Quadro 1 - Formação de palavras

Símbolo	Significado
Compr	Representa o conceito lexical de ‘aquisição, apropriação mediante pagamento’.
A (1)	Indica que o verbo pertence à primeira conjugação.
Ri	Indica que o verbo está flexionado no futuro do pretérito do modo indicativo.
A (2)	Indica que o verbo está flexionado na terceira pessoa do singular.

Fonte: autoria própria

Na sequência, no quadro 2, veremos uma breve apresentação dos tipos de morfemas em português: desinência, raiz e radical, afixo e vogal temática.

Quadro 2 - Estruturas (morfemas)

Termo	Definição	Exemplo
Desinência	Cada palavra possui um final que marca gênero e número (para substantivos e adjetivos) ou número, pessoa, tempo e modo (para verbos). Podem ser nominais ou verbais.	**Em 'garotas':** 'a' indica feminino e 's' indica plural. **Em 'falássemos':** 'sse' indica modo/tempo (pretérito imperfeito do subjuntivo) e 'mos' indica número/pessoa (1ª pessoa do plural).
Raiz e Radical	A raiz é o constituinte da palavra que possui significado lexical, sem incluir afixos derivacionais ou flexionais. O radical também é o constituinte da palavra, mas pode incluir afixos derivacionais.	**Raiz:** 'com' em 'comer', 'comemos', 'comendo'. **Radical:** 'com' em 'comer', mas 'comid' em 'comida', 'comidinhas'.

Afixo	Elemento que se junta ao radical para formar uma nova palavra, alterando seu sentido básico. Podem ser prefixos (antes do radical), sufixos (depois do radical) ou infixos (dentro do radical).	**Prefixo:** 'des-' em 'desfazer'. **Sufixo:** '-mente' em 'necessariamente'. **Infixo (raro):** 'o' em 'gasômetro'.
Vogal Temática	Elemento que liga o radical às desinências, formando o tema. Pode ser nominal (‘a’, ‘e’, ‘o’) ou verbal (‘a’, ‘e’, ‘i’), indicando a conjugação dos verbos.	**Nominal:** 'a' em 'artista', 'e' em 'cliente', 'o' em 'prato'. **Verbal:** 'a' em 'falar' (1ª conjugação), 'e' em 'ler' (2ª conjugação), 'i' em 'sorrir' (3ª conjugação).

Fonte: autoria própria

Quando há a junção de dois morfemas: o radical e a vogal temática, temos a criação de uma forma lexical, a saber: o tema. Vejamos o exemplo a seguir: a combinação do radical “ atlet ” com a vogal temática nominal “a”, forma-se o tema “atleta”. Ainda que seja a junção de dois tipos de morfemas, o tema também é considerado como um morfema; a forma lexical assumida por ele é coincidente com as do lexema, da lexia e do lema, que são entradas dos verbetes nos dicionários. Aqui, cabe uma ressalva importante: na linguística textual, tema e lema são distintos, o que não acontece nas abordagens morfológicas.

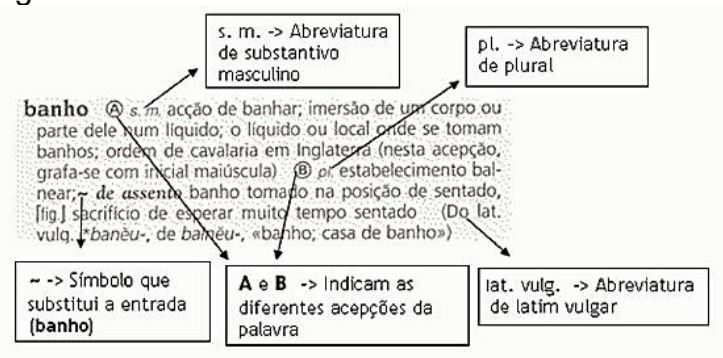
Quadro 3 - Tema lexical

Radical	**Vogal Temática**	**Tema**
atlet	A	atleta
poet	A	poeta
artist	A	artista
jornal	ista	jornalista

Fonte: autoria própria

O tema lexical é oriundo da fusão entre dois elementos basilares da estrutura das palavras: o radical e a vogal temática , como se vê no quadro (3) acima. A configuração lexical do tema é semelhante às formas. A configuração lexical do tema se assemelha às formas de lexema, lexia e lema , que correspondem às palavras-base registradas nos dicionários. Contudo, vale ressaltar que, no âmbito da Linguística Textual, os conceitos de tema e lema possuem interpretações distintas das empregadas na morfologia.

Figura 16 - Modelo de entrada nos dicionários



Fonte: autoria própria

Nos dicionários, a organização dos verbetes segue um padrão que facilita a consulta e a compreensão dos significados das palavras. A imagem apresentada (figura 16) ilustra essa estrutura, destacando os principais elementos utilizados para categorizar e para definir os termos. Primeiramente, vê-se que a palavra de entrada, a que aparece em negrito, é uma representação do vocábulo que será explicado. As abreviações indicam a classe gramatical da palavra, como "s. m.", que informa que o termo é um substantivo masculino, ou "pl.", que informa ser um plural. Ademais, para evitar redundâncias, utiliza-se o símbolo "~", que substitui a palavra de entrada dentro da própria definição. A diferenciação de significados que o mesmo termo pode ter, é outro aspecto importante na estruturação dos verbetes são empregadas letras que facilitam a organização das diferentes acepções da palavra; a etimologia do termo, normalmente, é indicada, como no exemplo "lat. vulg.".

Como sinônimo de unidade lexical, o termo lexema é utilizado, trazendo a representação de aspectos sonoros, gráficos e de significado. Por exemplo, a palavra “lembrei” é um lexema cuja pronúncia fonética corresponde a [lẽ'brej]. Sob o ponto de vista morfológico, trata-se de um verbo conjugado na primeira pessoa do singular

do pretérito perfeito do indicativo. Em relação ao significado, encontramos no dicionário o sentido de que algo foi trazido à memória, que foi recordado. Nos estudos lexicais no Brasil, há também o conceito de lexia, que se refere à realização concreta do lexema. Enquanto o lexema é uma forma abstrata – como, por exemplo, “planta”, a lexia condiz com as variações concretas dessa forma no discurso, como “plantas”.

Nesse sentido, a lexia se manifesta diretamente no texto ou na fala, ao passo que o lexema pertence ao nível abstrato da língua. De maneira simplificada, podemos dizer que o lexema é uma representação conceitual abstrata, enquanto a lexia corresponde à forma materializada dessa unidade em um enunciado. Vale lembrar que, na linguística, frequentemente encontramos diferentes termos que expressam conceitos próximos ou semelhantes. Por exemplo, palavra pode ser considerada sinônima de vocábulo.

Outro termo relevante é o lema, a representação sintático-semântica de um item lexical. Por meio do lema, conseguimos identificar os argumentos exigidos em uma estrutura gramatical. Retomando o verbo “lembrar”, por exemplo, vemos que ele necessita de dois argumentos: um como sujeito e outro como objeto. Outrossim, o lema serve como forma padrão de entrada nos dicionários. Nesse contexto, ele representa uma categoria de palavra e de suas possíveis flexões. Por exemplo, “geógrafo” é o lema correspondente às formas “geógrafo”, “geógrafos”, “geógrafa” e “geógrafas”.

Quadro 4 - Distinção entre lexema, lexia e lema

Termo	Definição	Exemplo
Lexema	Unidade lexical abstrata que representa aspectos de som, forma e significado.	O termo 'lembrei' é um lexema, cuja pronúncia fonética é [lẽ'brej] e representa a ação de recordar.
Lexia	Realização concreta do lexema no discurso, manifestando-se no texto ou na fala.	Enquanto 'planta' é um lexema, 'plantas' é uma lexia, pois se manifesta concretamente no discurso.

Lema	Representação sintático-semântica de um item lexical, servindo como forma padrão de entrada nos dicionários.	'Geógrafo' é o lema que agrupa as formas flexionadas 'geógrafo', 'geógrafos', 'geógrafa' e 'geógrafas'.

Fonte: autoria própria

Em suma, o lexema trata-se de uma unidade abstrata, enquanto a lexia, de uma unidade concreta no discurso. Já o lema, traz a organização das propriedades sintático-semânticas, agindo como referência principal para o agrupamento das distintas formas de uma palavra.

Outras definições que não podem ser deixadas de lado são as de *“token”* e “ *types*”. Qualquer segmento de caracteres que possui valor é chamado de *token.* Nas línguas europeias, a sequência é formada por caracteres delimitados por espaços gráficos, e a tokenização é ajustada para separar os sinais de pontuação. No entanto, na maioria das outras línguas, a tokenização não se baseia em espaços gráficos. Sendo assim, a somatória da quantidade de palavras e dos sinais de pontuação de uma sentença estabelecem a quantidade de *tokens*. *Type*, por sua vez, está relacionado à unicidade de tokens encontrados em uma frase ou em um texto. A proporção *token/type* (divisão da quantidade de tokens pela quantidade de *types)* indica a riqueza lexical de um texto; isto é, indica qual a diversidade de palavras existentes em um *corpus*, não considerando as suas repetições.

Quadro 5 - TTR – Type/token Ratio

Categoria:	**Descrição:**	**Exemplo na Frase:**
***Tokens* (Total)**	Quantidade total de palavras e sinais de pontuação no texto.	["O", "cachorro", "branco", "pulou", "sobre", "o", "tapete", "."]
***Types* (Únicos)**	Número de palavras únicas, desconsiderando repetições.	["O", "cachorro", "branco", "pulou", "sobre", "tapete", "."]
TTR *(type-token ratio)*	Razão entre tokens e types: 1.17 (indica riqueza lexical).	7 / 6 = 1.17

Fonte: autoria própria

3.4.1 Etapas de processamento morfológico, morfossintático e sintático

Em continuidade, no processamento de linguagem natural (PLN), a morfologia aborda também a caracterização de traços linguísticos, como os de gênero, de número, de tempo, de modo, de pessoa etc. Ao passo que a morfossintaxe analisa as

escolhas morfológicas como as flexões e concordâncias, e como elas afetam a organização em uma sentença. Resumidamente, a morfologia se dedica ao estudo da composição interna das palavras e de seus componentes básicos; a morfossintaxe, por sua vez, investiga o impacto dessas estruturas na organização das sentenças.

Figura 17 - Divisão entre a morfologia e a morfossintaxe

Fonte: autoria própria

Legenda: a imagem nos mostra que a morfologia se dedica ao estudo da composição interna das palavras e de seus componentes básicos, enquanto a morfossintaxe investiga o impacto dessas estruturas na organização das sentenças.

Com esses conceitos estabelecidos, explorarei como esse nível de análise linguística está presente no âmbito do Processamento de Linguagem Natural (PLN). Para que uma aplicação seja criada é indispensável realizar o pré-processamento. Durante esse processo, algumas tarefas comuns incluem:

Quadro 6 - Tarefas inseridas no pré-processamento

Processo	Definição	Exemplo
Sentenciação	Divisão do texto em sentenças, identificando limites de frase.	'O gato dormiu. Acordou cedo.' → ['O gato dormiu.', 'Acordou cedo.']
Tokenização	Segmentação do texto em palavras ou sequências de caracteres.	'O gato dormiu.' → ['O', 'gato', 'dormiu', '.']

Vetorização de Subtokens	Subdivisão das palavras em unidades menores, como subpalavras.	'inacreditável' → ['in', 'acredit', 'ável']
Normalização Lexical	Processo de padronização lexical, incluindo lematização e derivação de radicais.	'correr', 'correu', 'correndo' → 'correr' (lematização); 'correr' → 'corr-' (radical)

Fonte: autoria própria

Além das atividades de pré-processamento, também é possível realizar tarefas relacionadas ao conteúdo textual, como a etiquetagem morfossintática, que classifica as palavras de acordo com suas categorias gramaticais (PoS tagging) , e a anotação automática de atributos morfológicos (anotação de *features* morfológicas). Cada etapa é fundamental para assegurar a qualidade e a funcionalidade das aplicações baseadas em PLN. Vejamos um exemplo de etiquetagem morfossintática:

Figura 18 - Exemplo de Anotação de PoS com *Coarse Tags*

Fonte: autoria própria

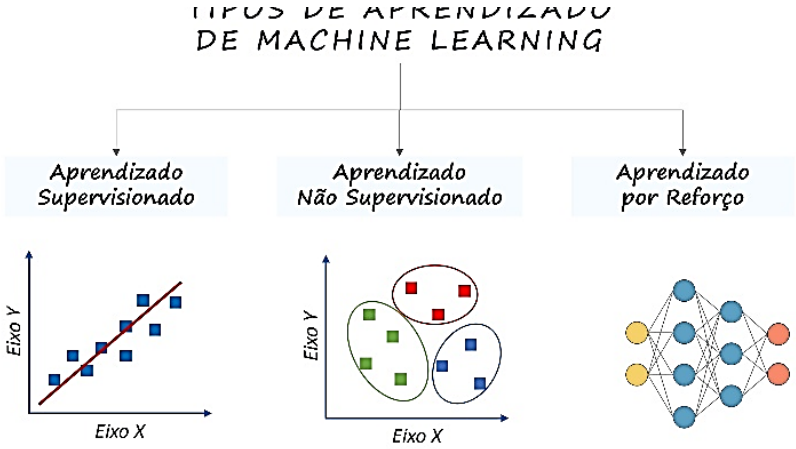
Por conseguinte, a segmentação em sentenças, também chamada de sentenciação ou detecção de limites de sentença, consiste em dividir um texto em unidades menores, correspondentes às sentenças, delimitando onde cada uma começa e termina. Essa tarefa, amplamente conhecida como detecção de limite de sentença, está essencialmente focada em identificar os pontos de finalização das sentenças. Trata-se de um processo intrinsecamente complexo devido à ambiguidade inerente às línguas naturais, o que impede uma determinação infalível dos limites das sentenças. No português escrito, as abordagens mais comuns utilizam sinais de pontuação como delimitadores. Entretanto, no caso de textos falados ou em línguas que não possuem uma pontuação evidente para indicar o término das sentenças, o desafio de segmentação se torna significativamente maior.

Embora na língua portuguesa o reconhecimento desses sinais de pontuação não seja tão complexo, dada a limitação de seu conjunto, a dificuldade

surge na desambiguação dos caracteres em contextos que apresentam outras funções. A segmentação automática de textos traz-nos um problema que ainda não está plenamente resolvido: as abordagens computacionais desenvolvidas até aqui, podem ser categorizadas das seguintes formas: a primeira delas baseia-se em regras predefinidas, que utilizam padrões específicos para identificar os limites das sentenças (heurísticas, listas de abreviações etc.).

A segunda abordagem fundamenta-se no aprendizado supervisionado, em que modelos computacionais são treinados a partir de conjuntos de dados anotados, conhecidos como *gold standards*. A eficácia dessa técnica está inteiramente ligada à quantidade e a relevância das amostras que são utilizadas no treinamento; essas devem estar alinhadas com as características dos textos que serão segmentados. Já a terceira categoria, conhecida como aprendizado não supervisionado, faz uso de grandes volumes de dados não anotados, mas representativos, possibilitando que os modelos construam padrões de segmentação autônomos, a partir das regularidades da língua.

Figura 19 - Tipos de aprendizado de máquina



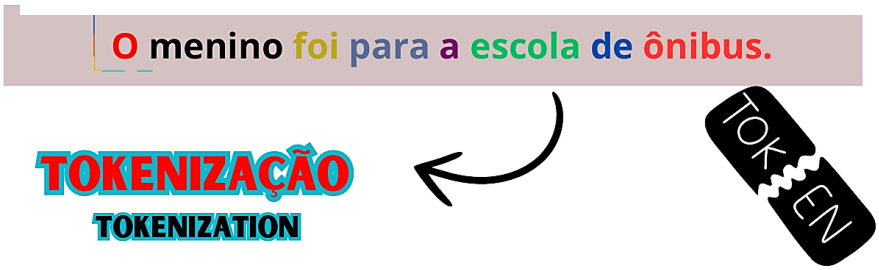
Legenda: a imagem apresenta três tipos de aprendizado de máquina: o aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço.

No PLN a tarefa de segmentação de sentenças é crucial, pois trata se de uma etapa preliminar do pré-processamento; se houver falhas nessa etapa, há um grande risco do comprometimento das fases posteriores, o que pode impactar negativamente

o resultado das aplicações mais complexas. No entanto, apesar de sua relevância, a segmentação de sentenças nem sempre recebe a devida atenção tanto nas pesquisas acadêmicas quanto nas práticas implementadas, em comparação com outras tarefas no campo do PLN. Ademais, embora a segmentação trabalhe com sentenças como unidade de análise, mantém uma relação intrínseca com a morfologia. Muitos dos casos de ambiguidade que dificultam a segmentação estão relacionados à delimitação e interpretação das palavras, que constituem o foco central da análise morfológica.

Assim, é evidente que questões morfológicas influenciam diretamente o sucesso ou a falha das técnicas de segmentação automática. A tokenização (laconicamente mencionada), ou tokenization em inglês, refere-se à divisão de um texto em suas menores unidades linguísticas, denominadas tokens. No caso da língua portuguesa, esse processo geralmente baseia-se na separação de palavras utilizando delimitadores, como espaços em branco ou sinais de pontuação (vírgulas, dois pontos, ponto e vírgula, hífens e pontos finais). Contudo, é necessário considerar casos específicos nos quais certos delimitadores não devem ser separados dos caracteres adjacentes. Vejamos um exemplo de tokenização:

Figura 20 - Tokenização



Fonte: autoria própria

Outra função relevante da tokenização é a descontração de palavras contraídas, como “do”, que é decomposta em “de” + “o”, ou “neles”, que é separada em “em” + “eles”. Em muitas situações esse processo não é tão complexo, uma vez que a palavra contraída não apresenta ambiguidade. Adicionalmente, a tokenização enfrenta dilemas quanto à separação de palavras hifenizadas. Substantivos compostos como “segunda-feira”, por exemplo, não costumam ser divididos, pois representam uma única unidade semântica. Por outro lado, formas como “acho-me” são frequentemente tokenizadas em três unidades: o verbo “acho”, o hífen e o

pronome “me”. A escolha por separar ou não esses elementos dependerá da aplicação específica para a qual a tokenização está sendo realizada. A implementação da tokenização pode se dar por meio de abordagens baseadas em regras, também por técnicas de aprendizado de máquina, supervisionadas.

A tokenização (que será detalhada e ilustrada, posteriormente), tal como o processo de segmentação em sentenças, pode ser implementada tanto por meio de abordagens baseadas em regras quanto por técnicas de aprendizado de máquina, supervisionadas ou não. Entretanto, a tokenização apresenta menor complexidade em comparação à segmentação, especialmente devido à existência de recursos léxicos que auxiliam na delimitação dos tokens de maneira mais eficiente.

Por conseguinte, temos a normalização, que consiste em um processo de conversão das palavras em formas padronizadas, resultando na eficiência do processamento. Entre os exemplos mais recorrentes de normalização, destacam-se: a conversão de versões abreviadas de palavras, como transformar “ tbm ” em “também”; a conversão de caracteres para minúsculas, como no caso de “Ela” para “ela”. É a lematização, que estabelece que “temos” é uma conjugação do verbo “ter”; e é a radicalização, que identifica, por exemplo, que “recomeço” tem o radical “começo” precedido do prefixo “ re ”.

Existem diferentes tipos de normalização que podem ser aplicados em diferentes abordagens e para diferentes objetivos da aplicação computacional. Nesse sentido, o intuito do processamento textual influencia a tomada de decisão sobre quais formas de normalização são necessárias. Por exemplo, para que o sentido geral de um texto seja identificado, a conversão de abreviaturas é importante, uma vez que permite tratar a abreviação e sua equivalência. Contudo, em tarefas como a identificação de entidades nomeadas e a substituição de caracteres minúsculos podem ser empecilhos, haja vista que essa transformação tende a obscurecer informações essenciais como os nomes próprios e as siglas. A desambiguação sintática é o principal desafio, considerando que palavras podem ter diferentes lemas, a depender do contexto. Observem a imagem:

Figura 21 - Exemplo de lematização

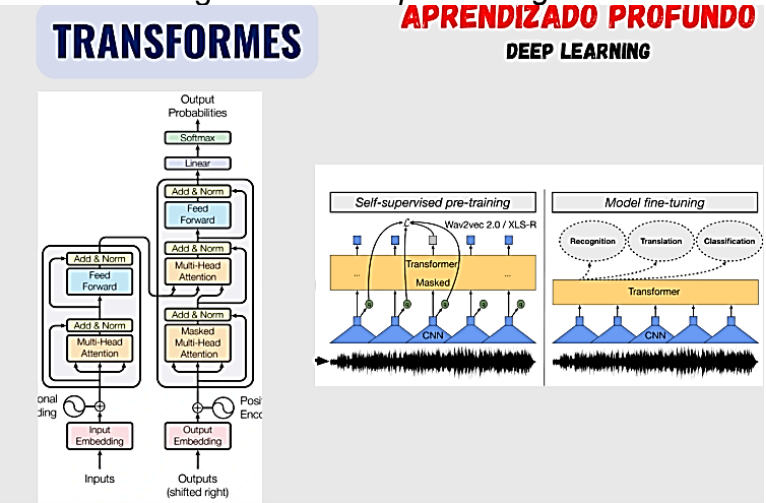
Fonte: autoria própria

Finalmente, temos o aprendizado profundo *(deep learning)*[22] que tem ganhado destaque como uma abordagem altamente eficaz, sobretudo em cenários que dispõem de grandes volumes de dados. Essa metodologia utiliza redes neurais artificiais com múltiplas camadas, integrando diferentes níveis de abstração. O processo começa com o pré-processamento dos dados cuja saída alimenta a primeira camada da rede. A partir dessa entrada inicial, as redes neurais utilizam modelos de linguagem pré-treinados para atribuir a etiqueta PoS [23](*Part of speech*) mais provável a cada palavra. Durante o treinamento, o modelo ajusta iterativamente os pesos das conexões entre as camadas, refinando sua capacidade de predição.

[22] O aprendizado profundo é um subconjunto do machine learning (ML), em que redes neurais artificiais - algoritmos modelados para funcionar como o cérebro humano - aprendem com grandes quantidades de dados.

[23] PoS é a sigla para Part of Speech, que em português significa classe gramatical ou categoria morfossintática de uma palavra — como substantivo, verbo, adjetivo, advérbio etc. No contexto do Processamento de Linguagem Natural (PLN), a etiqueta PoS é uma marca atribuída a cada palavra (ou *token*) por algoritmos treinados, indicando qual a sua função gramatical naquela frase.

Figura 22 - *Deep Learning*



Fonte: autoria própria

Como visto, as áreas de Morfologia e Morfossintaxe estão intimamente relacionadas e se complementam. A identificação dos traços morfológicos de uma palavra depende da compreensão de sua classificação gramatical (PoS), enquanto a determinação do PoS , por sua vez, frequentemente exige uma análise dos seus traços morfológicos. De modo geral, temos um paralelo entre a linguística e o PLN: na linguística, o morfema é a menor unidade significativa, estudado pela Morfologia; no PLN, a menor unidade de processamento é a palavra, associada à Morfossintaxe.

O foco, a partir de agora, recairá sobre a sintaxe, subárea responsável por estudar como as palavras se organizam em estruturas que desempenham diferentes funções gramaticais no âmbito das frases ou sentenças. No modelo de estratos do sistema, a sintaxe se posiciona como uma camada central, intermediando a organização das funções provenientes da morfologia e fornecendo estruturas fundamentais para os estratos superiores, como a semântica e a pragmática.

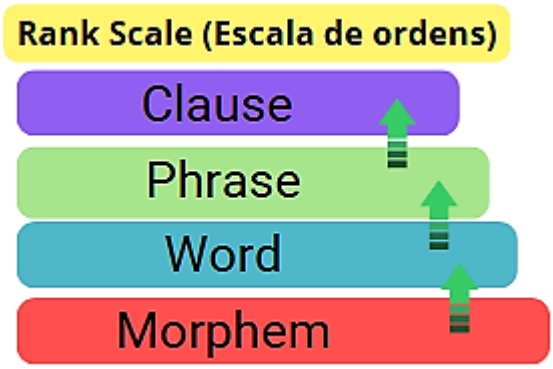
Figura 23 - Subáreas do estudo da linguagem

Fonte: autoria própria

Legenda: a imagem nos mostra a subáreas referentes ao estudo da linguagem.

Além dessa organização em estratos, a linguagem apresenta uma escala hierárquica de ordens (*rank scale*), que abrange diferentes unidades de análise. Nesse modelo, os morfemas constituem a menor unidade, combinando-se para formar palavras. Essas, por sua vez, se organizam em sintagmas, que desempenham funções específicas nas orações, como sujeito, objetos e complementos. A oração, unidade superior de análise, é composta por palavras organizadas de forma a construir significados relacionados a eventos. Consequentemente, na língua escrita, a sentença, também chamada de frase ou período, é delimitada por convenções gráficas, como a inicial em maiúscula e a pontuação final (ponto, interrogação ou exclamação).

Figura 24 - *Rank Scale*



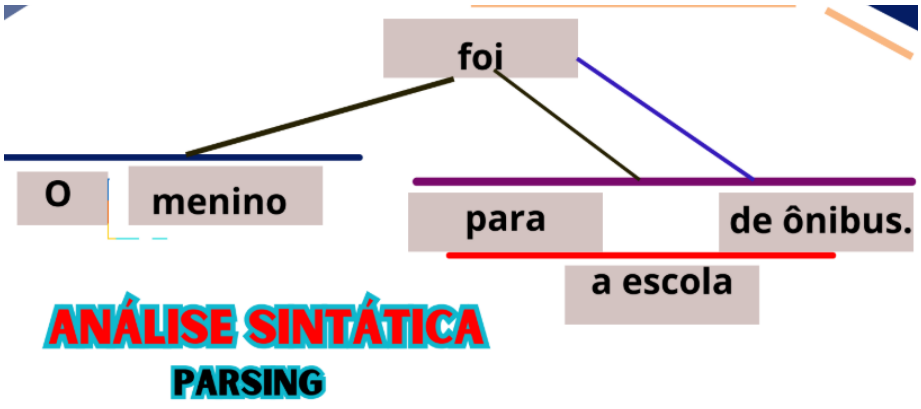
Fonte: autoria própria

Legenda: a imagem acima representa uma *Rank Scale* (Escala de ordens)

Distinguir “oração” de “sentença” é crucial para o PLN, pois uma sentença pode conter mais de uma oração, representando múltiplos eventos. É importante nos atentarmos aos níveis da escala de ordens e a sua relevância para a análise sintática, examinando como palavras e sintagmas interagem nas orações.

No que condiz à sintaxe, ela é o nível de análise linguística que investiga os padrões de estruturação das sentenças. Nesse âmbito, analisamos como as palavras se organizam em unidades que constroem significados dentro de uma sentença. Levando em conta as etapas de um processamento sintático, consideramos a classe gramatical de cada palavra, sua posição na sentença e sua relação com as demais. No contexto do Processamento de Linguagem Natural (PLN), a análise computacional no nível sintático é denominada *parsing*.

Figura 25 – *Parsing*



Fonte: autoria própria
Legenda: a imagem acima representa uma *Rank Scale* (Escala de ordens)

A ferramenta responsável por essa tarefa é denominada *parser*, e o recurso produzido a partir da análise sintática é conhecido como *treebank*. Embora hoje a separação gráfica entre palavras e sentenças pareça natural, esse é um fenômeno relativamente recente na história da escrita. Na Antiguidade, por exemplo, era comum o uso da *scriptio continua*, uma forma de escrita sem espaços ou pontuação, como se observa nos textos do latim clássico. Assim, o que hoje entendemos como segmentação textual resulta de um processo histórico de refinamento da linguagem escrita.

Quadro 7 - Separação entre palavras e sentenças

Análise	**Explicação**	**Exemplo**
Sentença Original	A frase base, sem ajustes ortográficos: 'o garoto tirou um cochilo.'	'o garoto tirou um cochilo.'
Versão Impressa	Com capitalização e pontuação adequadas: 'O garoto tirou um cochilo.'	'O garoto tirou um cochilo.'
Agrupamentos Intermediários	Unidades intermediárias que formam blocos de significado: 'o garoto' e 'um cochilo'.	['o garoto'], ['um cochilo']

Fonte: autoria própria

O reconhecimento das palavras individuais e da sentença completa como unidade, no entanto, não é suficiente para explicar por que o texto faz sentido para nós. Entre palavras isoladas e a sentença como um todo, identificamos agrupamentos intermediários que obedecem às estruturas da língua. Esses agrupamentos, como "o garoto" e "um cochilo", representam blocos de informação que constroem significados parciais e contribuem para o entendimento geral do texto. Dentro desses agrupamentos, há uma ordem específica em que as palavras se sucedem, condicionada pela gramática. Por exemplo, na sequência "o garoto", o artigo "o" sempre precede o substantivo "garoto", enquanto uma ordem inversa ("garoto o") parece incorreta ou causa estranhamento. O estudo da organização e disposição das palavras em uma estrutura ordenada é realizado pela sintaxe, que tem como objetivo identificar as regras que determinam quais combinações de palavras são gramaticalmente aceitáveis.

Os agrupamentos sintáticos funcionam como "peças" na construção de significados. Cada agrupamento contribui para uma unidade maior até formar uma oração completa. Cada agrupamento formado ao longo dessa estrutura desempenha uma função específica dentro da oração, funcionando como uma unidade única. Por exemplo, o agrupamento “o garoto” é constituído pelo artigo “o” (ou determinante) e o substantivo “garoto”. Esses dois elementos trabalham juntos para cumprir uma função e produzir um significado. Os agrupamentos que excedem o nível da palavra são denominados sintagmas (do inglês, *phrase*). Eles podem ser: nominais, verbais, adverbiais, adjetivais e preposicionais —, cada um com suas regras próprias com relação à organização das palavras, de acordo com a função contextual.

As unidades que compõem um sintagma são chamadas de constituintes (do inglês, constituent). Constituintes podem ser palavras isoladas ou agrupamentos que formam unidades maiores. No exemplo “O garoto tirou um cochilo”, temos cinco palavras que se organizam em constituintes intermediários até formar dois principais: [o garoto] e [tirou um cochilo].

No nível da palavra, exemplos de constituintes incluem substantivos, verbos e adjetivos. O núcleo é o elemento central que define o tipo de sintagma; ele é responsável por determinar a função do sintagma na oração. Por exemplo, o sintagma nominal “um cochilo” atua como objeto no sintagma maior “tirou um cochilo”. Como objetos são representados por sintagmas nominais, o núcleo deste é o substantivo. Ademais, os constituintes não se limitam a um único nível estrutural; eles integram todas as camadas da oração, desde os menores agrupamentos até as unidades mais amplas. No quadro 8, podemos observar os três constituintes da oração:

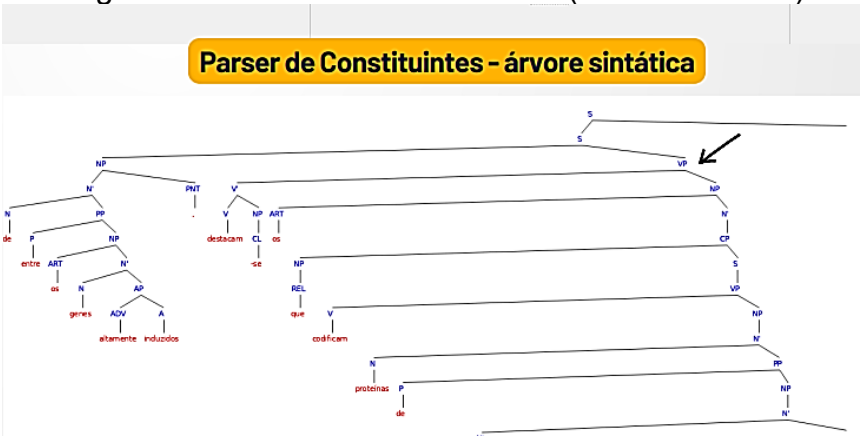
Quadro 8 - Constituintes da oração

Sintagma	**Núcleo**	**Classe de palavra do núcleo**	**Tipo de sintagma**
o garoto	garoto	Substantivo	sintagma nominal
tirou um cochilo	tirou	Verbo	sintagma verbal
Um cochilo	cochilo	Substantivo	sintagma nominal

Fonte: autoria própria

Dessa forma, palavras e sintagmas podem ser constituintes, sempre funcionando como parte de uma estrutura maior, que culmina na oração, considerada a unidade mais abrangente na escala de ordens. A análise das orações com base em seus constituintes e a hierarquia entre eles é conhecida como sintaxe de constituência, uma abordagem analítica muito utilizada no processamento de linguagem natural (PLN), ao lado da sintaxe de dependência. As árvores (em inglês, *trees*) são representações que permitem uma boa visualização da hierarquia dos constituintes.

Figura 26 - *Parser* de constituintes (árvore sintática)



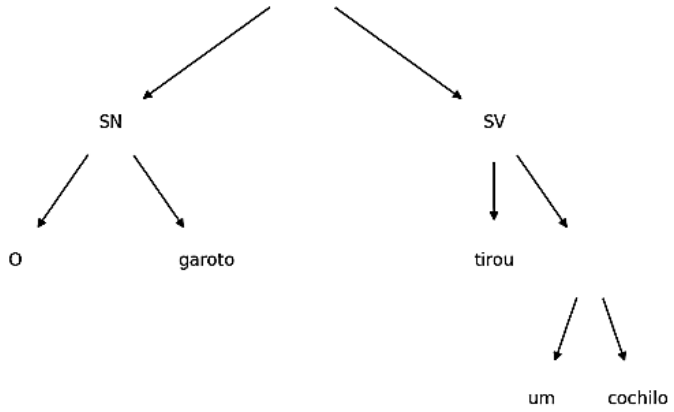
Fonte: autoria própria

Legenda: a imagem representa uma visualização da hierarquia dos constituintes

Nas representações com estruturas arbóreas, são utilizados diagramas conhecidos como árvores, para ilustrar graficamente a estrutura de uma sentença em sua hierarquia. Os nós (ou *nodes*) são os elementos discretos de um grafo, ou representação simbólica, enquanto as arestas (ou *edges*) são as linhas que conectam os nós, indicando as relações entre eles; tal forma de representação pode ser aplicada tanto na sintaxe de constituência quanto na sintaxe de dependência.

Já a representação por setas, vista na figura 27, também conhecida como "*arced arrows*", utiliza setas unidirecionais para indicar a relação de dependência entre um nó pai (governador) e um nó filho (dependente). A seta parte do elemento que rege a relação (pai) e aponta para o elemento regido (filho).

Figura 27 - *Arced arrows*



Fonte: autoria própria
Legenda: a imagem representa uma representação por setas

A representação por parênteses, por sua vez, utiliza pares de parênteses para delimitar as relações de dependência entre palavras. O parêntese mais externo engloba a relação entre o núcleo da sentença (elemento central) e seus dependentes. Cada par de parênteses interno representa a relação entre um governador e seu dependente. Essa notação, embora menos visual que a representação por setas, é computacionalmente mais tratável, facilitando o processamento automático das relações de dependência.

A indentação (figura 28), por fim, explora o recurso de deslocamento horizontal para representar a hierarquia entre as palavras. Os dependentes são indentados em relação aos seus respectivos governadores, revelando a estrutura hierárquica da sentença.

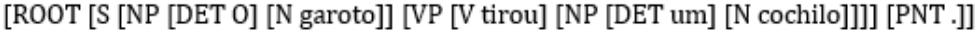
Figura 28 - Processo de indentação

1 (, tirou)
2 (tirou, garoto)
3 (garoto, o)
4 (tirou, cochilo)
5 (cochilo, um)
6 (tirou, .)

Fonte: autoria própria
Legenda: a imagem acima representa o processo de indentação.

Em continuidade, a sintaxe de constituência, também denominada "sintaxe de constituintes", adota uma perspectiva hierárquica na análise da estrutura sentencial. Há o agrupamento de palavras e sintagmas em constituintes; esses, podem ser englobados em constituintes ainda maiores. Esse processo de agrupamento resulta em uma hierarquia de unidades, que pode ser representada por meio de colchetes ou árvores de constituintes. Na representação por colchetes, cada constituinte é delimitado por colchetes, com os constituintes mais internos representando as unidades menores e os mais externos, as unidades maiores. A anotação de cada constituinte acompanha um conjunto de etiquetas morfossintáticas *(tagset),* que indicam a função gramatical do constituinte.

Figura 29 - Hierarquia de constituintes

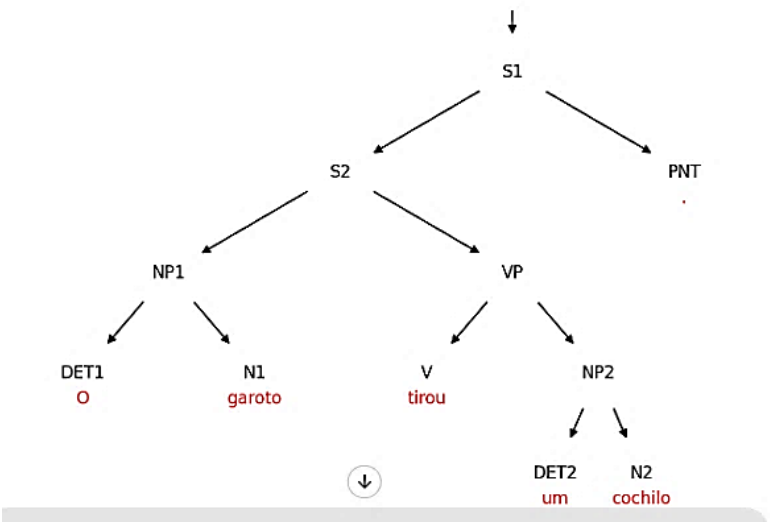


Fonte: autoria própria
Legenda: a imagem representa uma hierarquia de unidades.

A sentença na representação (na sintaxe de constituência) em árvore aparece como um nó raiz, que se ramifica em outros nós, nos constituintes da sentença final; as palavras individuais correspondem as folhas das árvores, enquanto os nós intermediários definem os sintagmas e outras unidades gramaticais. A sintaxe de constituência , portanto, oferece uma visão detalhada da estrutura hierárquica da sentença, revelando as relações entre seus constituintes e suas funções gramaticais.

Por outro lado, a sintaxe de dependência, está centrada nas relações de dependência entre as palavras e não na estrutura hierárquica dos constituintes. Esse modelo de sintaxe é baseado na gramática de dependência de Tesnière (1959), que descreve as relações entre as palavras como unidirecionais, nas quais uma palavra (o núcleo) rege ou governa outra palavra (o dependente). As relações de dependência podem ser de dois tipos: macrorrelações, que se estabelecem entre os núcleos de diferentes sintagmas, e microrrelações, que conectam elementos mais próximos, como um substantivo e seu artigo. A direção da dependência é crucial, pois indica qual palavra exerce influência sobre a outra.

Figura 30 - Sintaxe de dependência



Fonte: autoria própria

3.5 LÉXICO E GRAMÁTICA NO PLN

É sabido que as palavras são o material básico utilizado para realizarmos nossas ações de linguagem. Inicialmente, a gramática pode ser entendida como um conjunto de regras sistematizadas que servem para determinar o que é considerado adequado ou não em uma língua; para exemplificar, conforme a ordem das palavras em contextos não poéticos ou pautados em um alto nível de formalidade, emprega-se a construção “ele apreciou a obra” ao invés de “a apreciou ele obra”. Tais normas remetem à gramática, enquanto o léxico refere-se ao conjunto de palavras que compõem uma língua.

O léxico, além de reunir os vocábulos de uma língua, associa a eles definições morfossintáticas. Normalmente, cada palavra no léxico possui combinações que englobam as suas categorias gramaticais *(features)* e o seu lema. Todavia, pode ocorrer variação na categorização gramatical em conformidade com os critérios de seleção adotados, podendo seguir padrões variados.

Embora na língua portuguesa, definimos categorias como: substantivos, adjetivos, nomes próprios, numerais, pronomes, preposições, conjunções, advérbios

e verbos; a depender do critério adotado, outras categorias podem ser incluídas, como por exemplo: determinantes ou artigos; pode-se ainda realizar subdivisões, como entre conjunções coordenativas e subordinativas, ou entre verbos auxiliares e plenos.

Nesse sentido, a seleção das categorias gramaticais e das características morfológicas que serão consideradas, assim como de seus valores possíveis, é uma etapa crucial na construção de um léxico. Geralmente, a definição dessas categorias e características é acompanhada pela criação de etiquetas (ou *tags*) que representem as informações associadas a cada palavra no léxico. Por exemplo, para a palavra “delas”, o léxico poderia associar a categoria de pronome, o lema “ele” e as características de pronome possessivo, na terceira pessoa do plural e no gênero feminino. Nesse caso, a palavra possui uma única combinação gramatical possível, de lema e de *features*. Por outro lado, existem palavras que apresentam múltiplas combinações possíveis. Por exemplo, a palavra “amarras” pode ter diferentes interpretações:

Quadro 9 - Interpretações da mesma palavra

Forma	Lema	Flexão	Exemplo
Verbo	amarrar	Presente do Indicativo, 2ª pessoa do singular	Tu amarras o sapato.
Substantivo	amarra	Gênero feminino, plural	As amarras do navio estavam soltas.

Fonte: autoria própria

Como destaca Antunes (2012, p. 27), “a língua organiza e disponibiliza termos para expressar essas categorias, atuando como um intermediário na forma como percebemos e estruturamos a realidade.” O dinamismo das línguas nos mostra que mesmo nos léxicos mais abrangentes nem todas as palavras existentes podem ser englobadas, deixando lacunas vocabulares. A explicação para isso é o fato de que a linguagem não estabelece uma relação direta entre as palavras e os objetos do mundo, mas entre as palavras e as categorias cognitivas que construímos ao longo de nossa experiência.

3.5.1 Tipos de léxico: comum e especializado

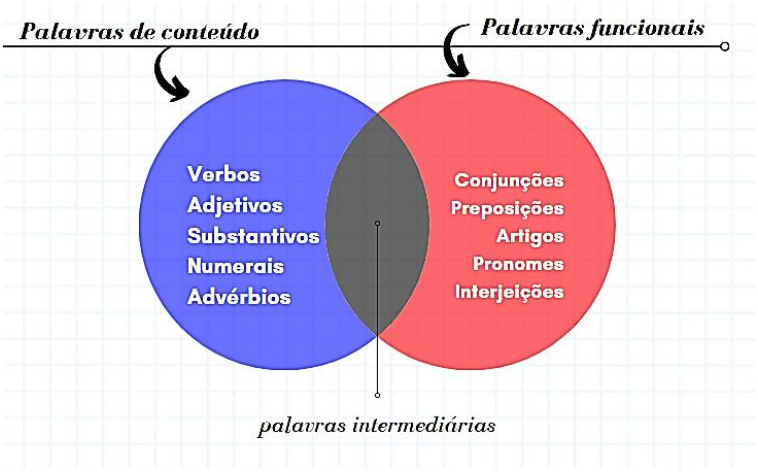
O léxico comum engloba as palavras de uma língua que não possuem um significado técnico-científico específico, ou seja, não estão associadas a conceitos construídos historicamente em uma área do conhecimento.

Em contraste, o léxico especializado refere-se a palavras que assumem um significado particular dentro de um sistema conceitual específico, geralmente vinculado a uma área científica. Nesse contexto, especializado, a palavra pode ser identificada como uma terminologia técnica (ou termo) ou permanecer como uma palavra comum. Por exemplo, o termo “cardiopatia” é visivelmente associado ao léxico especializado, enquanto “livro” é típico do léxico comum.

Contudo, essa categorização nem sempre é rígida, e é possível identificar casos que parecem estar em uma zona intermediária entre os dois tipos de léxico. Além disso, termos técnicos podem se tornar palavras comuns com o passar do tempo, assim como palavras do léxico comum podem assumir um status técnico em contextos especializados.

As palavras podem ser divididas em dois grandes grupos: funcionais ou de conteúdo. Essa distinção está relacionada à forma como elas se organizam em classes linguísticas. As palavras funcionais pertencem a uma categoria fechada, composta por elementos que não se expandem ao longo do tempo. Já as palavras lexicais fazem parte de uma categoria aberta, caracterizada pela possibilidade de criação constante de novos termos. Advérbios e pronomes, por sua vez, podem assumir características de ambas as categorias, dependendo de seu uso específico ou de subclassificação.

Figura 31 - Intersecção de categorias de palavras



Fonte: autoria própria

Por exemplo, no português, as preposições são praticamente fixas e não costumam sofrer alterações significativas ao longo dos anos. Em contraste, adjetivos e substantivos estão em constante transformação e atualização, frequentemente originados de fenômenos sociais e culturais. Podemos realizar a análise dessas categorias com base na ideia de "conjuntos", como se vê acima. Apesar disso, a classificação nem sempre é absoluta. Palavras como "não", dependendo do critério utilizado, podem ser interpretadas como lexicais, por apresentarem um sentido próprio. De modo geral, classes como substantivos, adjetivos e verbos estão relacionadas às categorias abertas, enquanto determinantes (como os artigos), preposições e conjunções fazem parte das classes.

3.6 CAMINHOS DA CORREÇÃO AUTOMÁTICA DE REDAÇÕES

Esta pesquisa, por utilizar uma ferramenta de análise textual fundamentada no PLN, insere-se em um campo mais amplo de aplicações computacionais voltadas à escrita, entre as quais se destaca a Correção Automática de Redação (CAR); trata-se de uma aplicação da área do PLN que visa avaliar e atribuir notas a textos escritos por meio de programas computacionais. É sabido que a correção manual de redações é uma prática canônica e antiga, a CAR, por sua vez, surgiu na década de 1960, mais especificamente para a língua inglesa e, mais recentemente, para o português.

Existem duas áreas distintas, porém, complementares, nesse campo: a Avaliação Automática de Redação (AAR), focada na atribuição de notas, e a Avaliação Automática de Textos (AAT), que busca fornecer feedback aos alunos para auxiliar no aprendizado da escrita.

Quadro 10 - Etapas básicas da correção automática de redações

Número	Aspecto	Descrição
1	Detecção de desvios no texto	Identificação de erros ortográficos, gramaticais e de outros tipos.
2	Atribuição de nota	Avaliação do texto, seja de forma global ou por critérios específicos.
3	Feedback ao aluno	Fornecimento de informações e sugestões para aprimorar a escrita.

Fonte: autoria própria

Cada uma dessas etapas, apresentadas no quadro 10, pode ser considerada uma aplicação independente no campo do PLN. Existem ferramentas, por exemplo, que podem auxiliar a escrita e que se concentram na detecção de desvios, enquanto outras se dedicam à geração de *feedback* para o aluno. A correção de redação, entretanto, deve ser vista como um processo de completude que traz em si todas as 3 etapas.

3.6.1 A redação escolar e a correção automática de textos

Desde o início deste texto, considerações foram feitas acerca da redação escolar; no entanto, cabe aqui, conceituá-la novamente, tanto como uma prática pedagógica quanto como um gênero textual crucial na avaliação de competências, sejam elas relacionadas à escrita a argumentação e a criatividade. Além disso, a redação desempenha um papel essencial no desenvolvimento da criticidade e na capacidade de expressão por meio da linguagem. Os temas abordados em redações são variados, podendo ir de tópicos simples e corriqueiros, há questões mais complexas e abstratas. Para compreender**mos** a avaliação da redação escolar, é fundamental diferenciar os conceitos de tipo textual e gênero textual. Cada redação deve atender a ambos, conforme indicado pela proposta apresentada. Por exemplo,

no Exame Nacional do Ensino Médio (Enem), o texto solicitado sempre pertence ao tipo argumentativo e ao gênero dissertativo-argumentativo.

Quadro 11 - Definições de alguns gêneros textuais

Gênero Textual	Descrição
Dissertação	O autor apresenta e disserta sobre um determinado tema, apresentando informações e argumentos relacionados ao assunto.
Narração	O autor conta uma história, relatando fatos e acontecimentos em alguma ordem que pode ser cronológica ou não.
Carta	O autor se dirige a um destinatário específico, fazendo requisições, solicitações ou expressando suas opiniões e sentimentos.
Artigo de opinião	O autor defende um ponto de vista sobre um tema específico, utilizando argumentos e evidências para sustentá-lo.
Resenha	O autor faz uma análise crítica de um texto, obra ou produto.

Fonte: autoria própria

Esses exemplos que aparecem no quadro 11, são apenas alguns dos gêneros textuais mais comuns na redação escolar. Para além dessas tipologias textuais e desses gêneros textuais, temos a descrição, a exposição, a crônica, o relatório, o conto, a fábula, entre outros. Múltiplos aspectos são considerados na correção de redações, eles podem variar de acordo com os modelos adotados pelas instituições ou pelos exames. Vejamos o quadro 12.

Quadro 12 - Aspectos considerados nas correções de redações

Critério	Descrição

Norma padrão da língua portuguesa	Analisa o uso correto da língua, avaliando a gramática, a ortografia, a escolha do vocabulário e a construção das frases.
Adequação ao tema	Verifica se o texto trata do tema proposto de forma completa e relevante, sem tangenciar ou fugir do assunto.
Estrutura do gênero textual	Considera se o texto segue o tipo e o gênero exigidos na proposta, como uma dissertação argumentativa ou um relato descritivo.
Coerência	Avalia a clareza das ideias, a consistência entre os argumentos e a articulação lógica dos elementos textuais.
Coesão	Examina o uso apropriado de conectivos, pronomes e outros recursos linguísticos que garantem a ligação entre as partes do texto.

Fonte: autoria própria

Outros aspectos específicos também podem ser analisados em alguns modelos de correção, como o do Enem, que analisa a interpretação dos textos motivadores e a qualidade da proposta de intervenção apresentada. Os critérios de avaliação que são utilizados no modelo de correção adotado pelo Enem são minuciosamente organizados em cinco competências, como já mencionado. Cada competência é avaliada em uma escala de 0 a 200 pontos. A soma das notas das cinco competências resulta em uma pontuação final que varia de 0 a mil.

Contudo, os critérios de avaliação e a pontuação que recebem, podem variar entre os modelos adotados pelas instituições. Vejam a descrição abaixo, na tabela 2:

Tabela 2 - Variação de pontuação atribuída aos critérios avaliativos

Modelo	Tipo/Gênero	Critérios Avaliativos	Nota Mín.	Nota Máx.

Enem	Dissertação-argumentativa	Domínio da norma culta da Língua Portuguesa	0	200
Enem	Dissertação-argumentativa	Desenvolvimento do tema	0	200
Enem	Dissertação-argumentativa	Estrutura do texto e conexão entre ideias	0	200
Enem	Dissertação-argumentativa	Proposta de intervenção	0	200
Enem	Dissertação-argumentativa	Pontuação final (soma das competências)	0-mil	mil
Fuvest	Dissertação-argumentativa	Tema/Gênero	1	5
Fuvest	Dissertação-argumentativa	Coerência/Coesão	1	5
Fuvest	Dissertação-argumentativa	Língua Portuguesa	1	5
Unesp	Dissertação-argumentativa	Tema	0	28
Unesp	Dissertação-argumentativa	Gênero/Coerência	0	28
Unesp	Dissertação-argumentativa	Língua Portuguesa/Coesão	0	28
Unicamp	Variável (depende do ano)	Tema	0	2
Unicamp	Variável (depende do ano)	Gênero	0	3
Unicamp	Variável (depende do ano)	Leitura (interpretação crítica dos textos de apoio)	0	3
Unicamp	Variável (depende do ano)	Língua Portuguesa/Coerência/Coesão	0	4

Fonte: autoria própria

Embora haja diferenças entre os modelos, como se vê na tabela 2, as falhas graves são pontualmente penalizadas, como: fuga ao tema, inserção de elementos não aceitos, uso de caligrafia ilegível ou de língua estrangeira. Quanto à presença ou a ausência de título, também pode variar; enquanto na Unesp, não exigem a titulação, a Fuvest e a Unicamp trazem especificações que podem ou não solicitar. Todas essas variações representam uma dificuldade para a Correção Automática de Redação (CAR), o que torna ainda mais complexa a criação de sistemas que abranjam a todos os critérios ao mesmo tempo.

De qualquer forma, ferramentas específicas como sistemas para detecção de desvios podem ser utilizadas em diferentes contextos, a exigência é que haja diretrizes semelhantes entre os modelos. Em alguns casos, os desvios identificados são usados apenas para calcular a nota sem que os alunos tenham uma devolutiva detalhada. Esse *feedback* pode ser realizado de duas formas: com a utilização da abordagem baseada em regras, também conhecida como abordagem simbólica; ou por meio de modelos estatísticos. A primeira é particularmente eficiente para identificar erros gramaticais que são comuns entre os falantes nativos. Já os modelos estatísticos são mais pertinentes para a captura de desvios de uso como aqueles que são frequentemente cometidos por não nativos.

Não obstante os sistemas baseados em regras sejam vistos como menos modernos a depender da complexidade das tarefas, eles ainda são muito utilizados para detecção de desvios em redações escolares. Os desvios em textos podem ser classificados em diferentes categorias, dependendo do tipo de erro que apresentam. Entre os desvios mais comuns estão:

Quadro 13 - Desvios mais comuns e suas categorias

Tipo de Desvio	Descrição
Desvios ortográficos	Englobam erros na grafia das palavras, uso inadequado de acentuação, problemas de capitalização e outros relacionados ao padrão ortográfico vigente.
Desvios gramaticais	Referem-se a problemas de concordância, regência, uso de pronomes, preposições e pontuação.
Desvios lexicais	Envolvem o uso inadequado de vocabulário ou registro, como escolha de palavras incompatíveis com o gênero textual ou contexto.
Desvios coesivos	São problemas relacionados ao uso de conectivos e à articulação entre as partes do texto.

Fonte: autoria própria

A identificação e tratamento de cada um desses desvios exigem abordagens específicas, implicando necessidade de recursos especializados, dentre eles, o que implica a necessidade de recursos especializados, como dicionários, *parsers* sintáticos e regras gramaticais.

Quanto aos desvios de vocabulário, registro e gênero, os diferentes gêneros textuais demandam escolhas específicas de léxico, vocabulário e, até mesmo, de registro linguístico. Alguns gêneros exigem a utilização exclusiva da norma culta, enquanto outros podem admitir variações coloquiais. Se o texto apresentar escolhas linguísticas incompatíveis com aquilo que se espera no gênero solicitado, ocorrem desvios de registro ou de gênero e de vocabulário, que podem ser identificados a partir de regras formais que associam palavras a determinados contextos. Por exemplo, em uma proposta que requer uma dissertação argumentativa, o uso frequente de pronomes e verbos na primeira pessoa do singular, como “eu”, “acho”, “penso” ou “creio”, pode indicar um desvio.

Expressões opinativas como “no meu ponto de vista” também não são condizentes com esse gênero textual, já que a dissertação argumentativa exige uma abordagem impessoal. Existem, ainda, os desvios de recursos coesivos, que desempenham um papel crucial na articulação e fluidez de um texto, todavia, ao serem utilizados inadequadamente, podem comprometer a coesão e levar à penalização em correções de redações. A título de ilustração, um exemplo comum de desvio de conectivo é a utilização do termo “porém” com sentido conclusivo, quando seu significado original é adversativo. Outro exemplo recorrente é o uso inadequado do pronome “onde” para se referir a conceitos não relacionados a localizações, como épocas, pessoas ou instituições. Em tais casos, o aluno poderia empregar “em que” ou “no qual” como alternativa correta. Esses problemas podem ser corrigidos por meio de regras específicas que analisem o contexto em que o conectivo ou pronome foi utilizado.

Do mesmo modo, regras que preconizem o uso correto e rebuscado de conectivos, contabilizando a sua utilização de forma positiva, também podem ser estabelecidas para que os pontos fortes de uma redação sejam destacados. Além de identificar desvios, as regras podem oferecer *feedback* detalhado, explicando por que determinada escolha foi inadequada e sugerindo substituições que melhorem o texto. A abordagem em questão não somente corrige, mas colabora com a aprendizagem

ativa do estudante. Regras formais podem ser usadas para detectar e tratar desvios no uso de recursos coesivos, mas muitas outras categorias de desvios podem ser identificadas e exploradas de forma semelhante.

3.6.2 Atribuição de notas aos textos

A extração de atributos (*features)* do texto está presente na abordagem tradicional para a atribuição de notas. Essa extração é utilizada para descrever características específicas das redações em um conjunto de treinamento. Os atributos passam por um processo de seleção e de transformação, o que podemos chamar de engenharia de atributos, antes de serem utilizados como entradas, algoritmos de aprendizado de máquina. Com base nesses atributos, são gerados modelos capazes de atribuir notas a novas redações. Na primeira versão do PEG (Ajay; Tillet; Page, 1973), os atributos eram baseados em contagens de elementos textuais, divididos em três categorias:

Quadro 14 - Categorias presentes nas contagens de elementos textuais

Tipo de Característica	**Descrição**	
Simples	Características diretas, como o número de adjetivos, que indicam uma relação linear entre sua presença e a nota atribuída (mais adjetivos, notas mais altas).	
Enganosamente Simples	Atributos como o número de palavras, que seguem uma relação não linear, penalizando redações muito curtas, mas perdendo relevância em textos muito extensos.	
Sofisticados	Características que indicam maior complexidade, como o uso de conectivos, que refletem a habilidade de articular ideias no texto.	

Fonte: autoria própria

Page e Petersen (1995) introduziram os conceitos de *proxes* e *trins*. Os *proxes* são variáveis observáveis, ou seja, características mensuráveis diretamente no texto, enquanto os *trins* são variáveis latentes, como a coerência, que não podem ser diretamente observadas, mas são inferidas a partir dos *proxes*. Por exemplo, a coerência textual pode ser representada por métricas como conectividade entre frases ou frequência de conectivos. Diversas ferramentas foram desenvolvidas para auxiliar na extração de atributos, vejamos:

Quadro 15 - Ferramentas utilizadas na extração de atributos

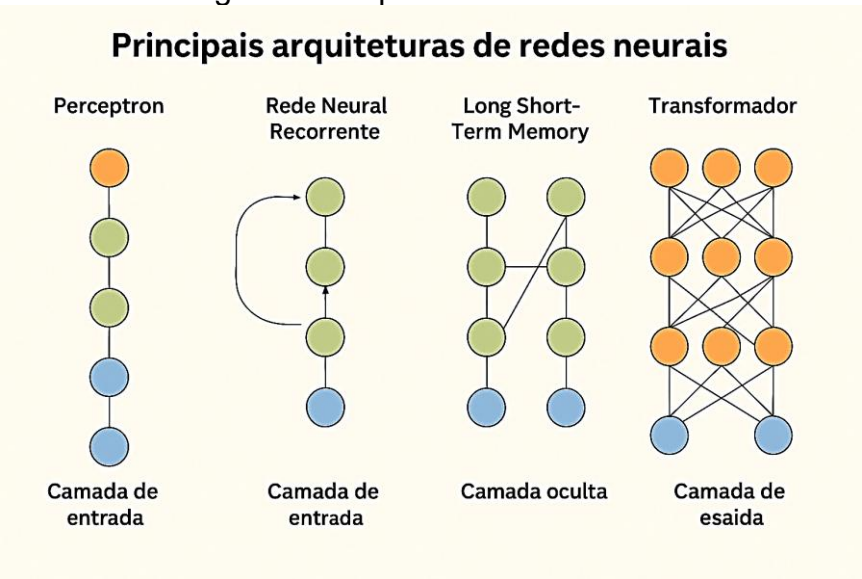
Ferramenta	Descrição
Coh-Metrix	Calcula métricas relacionadas à legibilidade, coesão e outros aspectos linguísticos.
Linguistic Inquiry Word Count (LIWC)	Analisa textos com base em métricas pré-definidas, como categorias lexicais e emocionais.

Fonte: autoria própria

Enquanto atributos independentes de conteúdo (como coesão e legibilidade) são úteis para avaliar critérios gerais, critérios como “tema” exigem atributos dependentes de conteúdo.

Adicionalmente, estratégias como mineração de argumentos têm sido exploradas para representar a qualidade da argumentação, combinando aspectos linguísticos, temáticos e argumentativos para criar modelos mais robustos. É importante fazermos menção dos avanços recentes, como por exemplo, as redes neurais e a representação de texto.

Figura 32 - Tipos de redes neurais



Fonte: autoria própria

É válido considerar que ao longo da evolução do processamento de linguagem natural, diferentes arquiteturas de redes neurais foram desenvolvidas para lidar com tarefas linguísticas mais eficientes. Arquiteturas consideradas mais simples, por possuírem apenas uma camada de processamento, deram espaço para as Redes Neurais Recorrentes (RNNs), por exemplo, capazes de processar sequências por meio de conexões que retroalimentam a rede, possibilitando certa memória contextual. Em seguida, foram criadas as (LSTM) *Long Short-term Memory* para introduzir mecanismos internos de controle para reter ou relegar determinadas informações. Por último, os *Transformers,* que revolucionaram o campo ao conseguirem eliminar a recorrência e adotar mecanismos de processamento paralelo de sequências inteiras, isso com alta eficiência e precisão; essa última arquitetura é a base de modelos avançados para tarefas de geração, de tradução e de análise textual em larga escala.

Embora, como vimos, tenham acontecido muitos avanços significativos, a atribuição automática de notas em português ainda enfrenta obstáculos únicos, como a pouca quantidade de *corpora* rotulados e de modelos linguísticos especializados; nem sempre as ferramentas e métodos desenvolvidos para o inglês são totalmente aplicáveis à língua portuguesa, são necessárias diversas adaptações. Todavia, abordagens combinadas, que integram atributos linguísticos, temáticos e argumentativos, obtêm melhores resultados para a atribuição de notas em critérios

específicos. Ainda há espaço para avanços no uso de redes neurais profundas e na criação de *corpora* diversificados, que contemplem diferentes contextos e modelos de correção.

3.6.3 Atribuição de nota para Redações do Enem

A maior parte dos trabalhos relacionados à atribuição de notas em redações de língua portuguesa utiliza *corpora* compostos por textos do modelo de correção do Enem como base de treinamento. Isso acontece por conta da relevância do exame que é uma aplicação em larga escala no Brasil, com critérios de avaliação (já discutidos) bem definidos. A maioria dos trabalhos que busca automatizar a atribuição de notas para redações, ainda foca a predição de uma nota global, contudo, existem também iniciativas que estão concentradas na predição de notas específicas para cada competência.

3.7 ESTATÍSTICAS BÁSICAS DO TEXTO NA CORREÇÃO AUTOMÁTICA DE REDAÇÃO

Existem inúmeras plataformas e empresas privadas que oferecem os serviços de Correção Automática de Redação (CAR), disponibilizam ao aluno estatísticas básicas do texto, como quantidade de palavras, caracteres, sentenças e parágrafos, além da distribuição de palavras por classes gramaticais (verbos, substantivos, adjetivos, preposições, conjunções, entre outras). Vale destacar, porém, que a maior parte das plataformas que operam no nosso país ou que processam textos em língua portuguesa, ainda se pautam em devolutivas baseadas nas características e nas propriedades linguísticas textuais. Por esse motivo, esta subseção será dedicada a essa abordagem.

Além dessas métricas fundamentais, algumas plataformas oferecem informações adicionais a partir de estatísticas simples, como a proporção de palavras únicas (*types*) em relação ao total de palavras (*tokens*), medidas de similaridade entre sentenças, desvio-padrão dos parágrafos, entre outras.

Entre os recursos disponíveis para análise textual, considerando estatísticas básicas, destaca-se a ferramenta de PLN utilizada nessa pesquisa, a NILC- Metrix (Leal et al., 2021), uma versão brasileira do Coh-Metrix. O NILC- Metrix (essa

ferramenta será detalhada na seção 4.2) é a versão mais recente do Coh-Metrix- Port (Scarton & Aluísio, 2010), contendo 200 métricas organizadas em 14 categorias (Quadro 24.2), que avaliam diferentes aspectos do texto, como coerência, coesão, inteligibilidade e complexidade, atribuindo valores a cada métrica. Na seção de "material e método" considerações importantes serão delineadas sobre a ferramenta, todavia, de antemão, vejamos os grupos de referência e os quantitativos de métricas que a NILC-Metrix nos traz.

Quadro 16 - Grupos de referência e quantitativos de métricas existentes no NILC-Metrix.

Categoria	**Qtde.**	**Descrição das métricas**
Medidas Descritivas	10	Servem para mostrar as características do *corpus*: número de palavras, quantidade de parágrafos, número de sentenças, características estatísticas; não são dados para contraste, mas para descrição.
Simplicidade Textual	8	Indicam o que seria uma sentença curta, média, longa e muito longa. Permite produzir estatísticas acerca da quantidade (quebrada) de tamanhos que o *corpus* tem. Nelas está contido o Dicionário de Palavras Simples (Biderman) para mostrar a proporção de palavras de conteúdo (as que carregam as informações do texto verbos, adjetivos, advérbios e substantivos)
Coesão Referencial	9	Capturam os elementos necessários para a correferência textual, que ocorre quando temos um nome, um pronome, um sintagma nominal encadeados; diz respeito à recuperação de uma entidade que já apareceu no texto antes. Compreendemos um texto quando recuperamos os seus referentes.
Coesão Semântica	11	Diz respeito aos tópicos que estão sendo discutidos no texto, à repetição de palavras, ao uso de sinônimos; essas métricas ajudam a computar as marcas textuais feitas pela

		similaridade das palavras (de palavra a palavra, de palavra a parágrafo e de sentenças com todo o texto).
Medidas Psicolinguísticas	24	São métricas subjetivas capturadas com experimentos para a Psicolinguística. A idade de aquisição das palavras, os valores de concretude, se são fáceis de serem imaginadas; quanto mais baixa a idade de aquisição, mais fácil é a compreensão do texto; o vocabulário básico é facilmente resgatado
Diversidade Lexical	15	Mede a riqueza lexical com índices de diversidade, razão entre palavras do tipo PoS e a proporção de palavras lexicais e funcionais.
Conectivos	12	Essas métricas vem do Coh-Metrix e são divididas 4 grupos de conectivos: aditivos; causais; lógico e temporais; dentro de cada um deles há uma categorização entre positivo e negativo. Por exemplo: se eu uso “também” , a próxima sentença irá agregar a anterior (positivo); se eu uso, “entretanto, ” a segunda sentença trará um contraste (negativo).
Léxico Temporal	12	São métricas que se referem ao conjunto de palavras e expressões relacionadas ao tempo que aparecem em um texto. Essas palavras podem indicar tempo cronológico (dias, meses, anos), sequenciamento temporal (primeiro, depois, finalmente), durações (minutos, horas, séculos) e frequência (sempre, nunca, frequentemente).
Complexidade Sintática	27	Verificam diferentes formas de alinhamento sintático e complexidade estrutural, como frases regulares e irregulares.
Densidade de Padrões Sintáticos	4	Mede a ocorrência de padrões sintáticos frequentes na língua, como sintagmas nominais, adjuntos e estruturas de dependência.

Informações Semânticas de Palavras	27	Analisa o uso de substantivos abstratos, palavras polissêmicas, palavras afetivas e polaridade (positiva e negativa) das palavras.
Informações Morfossintáticas de Palavras	42	São métricas relacionadas com as partes mais baixas da gramática, com as categorias de palavras. Os textos são contabilizados no que diz respeito aos adjetivos, substantivos, verbos, advérbios, preposições, pronomes, conjunções...
Frequência de Palavras	10	Determina a frequência das palavras no *corpus*, baseando-se na curva de Zipf e distribuição estatística.
Índices de Leiturabilidade	5	Inclui índices de legibilidade reconhecidos na PLN, como Fórmula de Chall, Índice de Brunet, Flesch, Gunning Fog e Estatística de Honoré.

Fonte: http://fw.nilc.icmc.usp.br:23380/nilcmetrix

Os cálculos dessas métricas geralmente resultam em valores numéricos, que, isoladamente, podem não ser suficientemente informativos para o aluno. No entanto, há diferentes estratégias para converter essas métricas em *feedbacks* interpretáveis. Por exemplo, considerando quatro métricas de Simplicidade Textual relacionadas ao tamanho das sentenças — *long_sentence_ratio , medium_long_sentence_ratio , medium_short_sentence_ratio e short_sentence_ratio* —é possível oferecer ao aluno uma análise qualitativa. Caso o sistema detecte um excesso de sentenças longas, pode-se fornecer um *feedback* como: "Seu texto apresenta um grande número de períodos longos , o que pode comprometer a clareza e a compreensão das ideias. Tente dividir algumas sentenças para tornar o texto mais fluido." Tanto as estatísticas básicas quanto as métricas avançadas do NILC-Metrix podem ser utilizadas não apenas para fornecer *feedbacks* detalhados aos alunos, mas também como parâmetros para a atribuição de notas, auxiliando na avaliação geral da redação ou de aspectos específicos do texto.

3.7.1 Correção manual *versus* correção automática

Sabe-se que a Correção Automática de Redações (CAR) divide opiniões entre diferentes grupos, como estudantes, professores, avaliadores, pesquisadores, especialistas em Linguística Computacional e afins, cientistas de dados e desenvolvedores de sistemas. Embora ainda exista certa resistência à adoção dessa tecnologia, já se tornou um consenso reconhecer as vantagens de corretores ortográficos e gramaticais quando integrados a ferramentas amplamente utilizadas, como *Microsoft Office*, *Google Drive,* redes sociais e teclados de dispositivos móveis. A principal discussão em torno da correção automática envolve seus benefícios e limitações, além de questões éticas sobre a substituição ou complementação da correção humana. Além disso, há preocupações sobre a possível subversão dos valores pedagógicos e educacionais ao se substituir a avaliação manual por um modelo automatizado.

A partir de agora procurarei elencar as diferenças práticas entre a correção manual e a correção automática, lançando luz às principais vantagens e limitações de cada abordagem. Até meados de 1980 e 1990, o modelo de avaliação de redações no Brasil era holístico, o avaliador atribuía uma nota global ao texto (por exemplo, de 0 a 100), sem a exigência de seguir critérios previamente estabelecidos. A partir dos anos 2000, esse modelo foi substituído por uma abordagem analítica, em que era necessário explicitar cada critério de avaliação.

De modo paralelo, o sistema de avaliação cega, em duplas, foi adotado; nele, duas correções independentes eram realizadas a fim de conferir maior confiabilidade e pertinência na atribuição de notas. Essa transição de modelos (do holístico para o analítico) resultou na padronização das grades de correção que vão abrindo caminhos para o desenvolvimento de sistemas automatizado de avaliação. É notório que tarefas que exigem mais estruturação e repetição tendem a ser executadas com mais eficácia por máquinas do que por seres humanos. Mas apesar da tentativa de padronização, ainda existem alguns obstáculos na objetividade dos critérios que são adotados por algumas metodologias de correção. O aumento da discrepância entre corretores pode ocorrer quando as diretrizes avaliativas são muito amplas ou pouco detalhadas, comprometendo a confiabilidade das notas.

Por outro lado, sistemas como o do Enem, que realizam treinamento rigoroso dos avaliadores, conseguem reduzir as inconsistências, ainda que não mitiguem totalmente as variações nas correções. Um indicativo disso são os índices de redações que precisam de terceira correção (*vide* seção 2.1.1), nos casos em que há

discrepâncias superiores a 80 pontos em uma competência ou 100 pontos na nota final. Por conta dessas limitações, a correção automática começou a ser considerada uma alternativa viável, principalmente, pela capacidade de minimizar as avaliações subjetivas e os vieses associados à correção manual. Entendemos que a correção manual está sujeita à multiplicidade de variáveis humanas que podem, de algum modo, comprometer a sua eficácia: o cansaço, a pressão por produtividade, o desinteresse na tarefa e as interpretações subjetivas. Ademais, o tempo e o custo envolvidos na avaliação manual são fatores relevantes.

A automação da correção de redações permite reduzir esses custos operacionais e eliminar alguns dos desafios associados ao trabalho humano. A confiabilidade dos sistemas automáticos é outra vantagem significativa. Enquanto a avaliação manual pode gerar discrepâncias entre diferentes corretores ou até mesmo em variações na nota atribuída pelo mesmo corretor em momentos distintos, a correção automática apresenta-se de modo mais consistente. Isto é, sempre que uma redação for processada pelo mesmo sistema receberá a mesma avaliação e a mesma pontuação.

Em contrapartida, a correção manual pode apresentar flutuações naturais, avaliadores distintos podem enfatizar aspectos textuais diferentes. Mais ainda, o mesmo corretor humano pode atribuir notas diferentes a uma redação dependendo do momento da avaliação, o que pode gerar questionamentos e solicitações revisionais. Assim sendo, vê-se que a correção automática aparece como uma alternativa de minimização de subjetividades, de redução de custos e de aumento da confiabilidade no processo avaliativo. Sobretudo, ainda que com limitações, a correção manual ainda é essencial para o reconhecimento de capacidades de interpretações mais complexas que estão envolvidas na construção textual algo que o sistema automatizado ainda não consegue fazer eficientemente.

Indubitavelmente, a produção textual é um processo sociocognitivo extremamente complexo, que envolve fatores linguísticos, culturais, psicológicos e emocionais. Por sua vez, a correção automática lida com padrões pré-estabelecidos, mas não consegue contemplar todo o significado do texto, menos ainda a interpretar o raciocínio do autor. Isto é, mesmo com modelos computacionais avançados que busquem tornar os sistemas mais "interpretáveis", ainda é impossível capturar todas as nuances sociais, emocionais e cognitivas envolvidas na produção e avaliação textual.

Logo, também surgem preocupações genuínas em relação ao impacto pedagógico da correção automática. No escopo dessas preocupações, pensamos acerca do risco de o foco excessivo nos aspectos formais da escrita, levar os alunos a priorizarem regras rígidas em detrimento da construção de sentido e da autonomia na produção textual. Isso levanta um questionamento fundamental: "a dependência excessiva da correção automática pode limitar o desenvolvimento da escrita dos alunos, tornando-os apenas treinados para atender aos critérios do algoritmo, em vez de incentivá-los a aprimorar sua capacidade crítica e expressiva?" Refletir sobre isso reforça a necessidade de um modelo híbrido de correção, que combine a eficiência e a objetividade da automação com a profundidade interpretativa da avaliação humana. Dessa forma, seria possível aproveitar as vantagens de cada abordagem, minimizando suas limitações e garantindo uma avaliação mais justa, acurada e eficaz.

4 MATERIAL E MÉTODO

Como afirma Gil (2002), pesquisa é um procedimento sistemático e racional que tem como objetivo principal a propiciação de respostas a determinados problemas. Requer-se uma pesquisa quando as informações dispostas ainda não são suficientes para a resposta das questões levantadas ou quando essas informações estão desordenadas, não podendo ser adequadamente aplicadas.

Para que uma pesquisa seja desenvolvida, servimo-nos dos conhecimentos disponíveis e da utilização criteriosa de métodos, técnicas, bem como, de outros procedimentos científicos. A seguir, explicarei acerca da abordagem e natureza desta pesquisa; na sequência, descreverei os materiais e os procedimentos metodológicos adotados para responder à hipótese levantada.

4.1 A ABORDAGEM QUANTIQUALITATIVA DA PESQUISA

Esta pesquisa adota uma abordagem quantiqualitativa, de natureza documental, descritiva e explicativa, cujo objetivo principal é a análise descritiva das métricas extraídas de redações nota mil do ENEM. A abordagem quantitativa desse estudo se justifica pela mensuração e pela interpretação de dados numéricos gerados a partir do processamento textual na NILC-Metrix e dos desdobramentos estatísticos; a abordagem qualitativa, por sua vez, complementa a análise, fornecendo uma compreensão mais profunda dos padrões linguísticos e recorrentes observados nas métricas linguísticas.

Acerca da dimensão quantitativa, o estudo está fundamentado no pensamento lógico-positivista, com ênfase nos aspectos mensuráveis da linguagem escrita. As métricas linguísticas foram obtidas de maneira sistemática e analisadas por meio de métodos estatísticos, a fim de identificar relações entre as variáveis linguísticas e compreender a totalidade do fenômeno investigado com base em padrões quantitativos observáveis. Para isso, a coleta de dados foi realizada com o auxílio de uma ferramenta computacional para garantir controle e precisão e para assegurar a objetividade em todas as etapas do estudo.

A dimensão qualitativa do estudo, em acréscimo, buscou interpretar os dados quantitativos à luz de conceitos teóricos da Linguística de *Corpus*, da Lexicologia e do

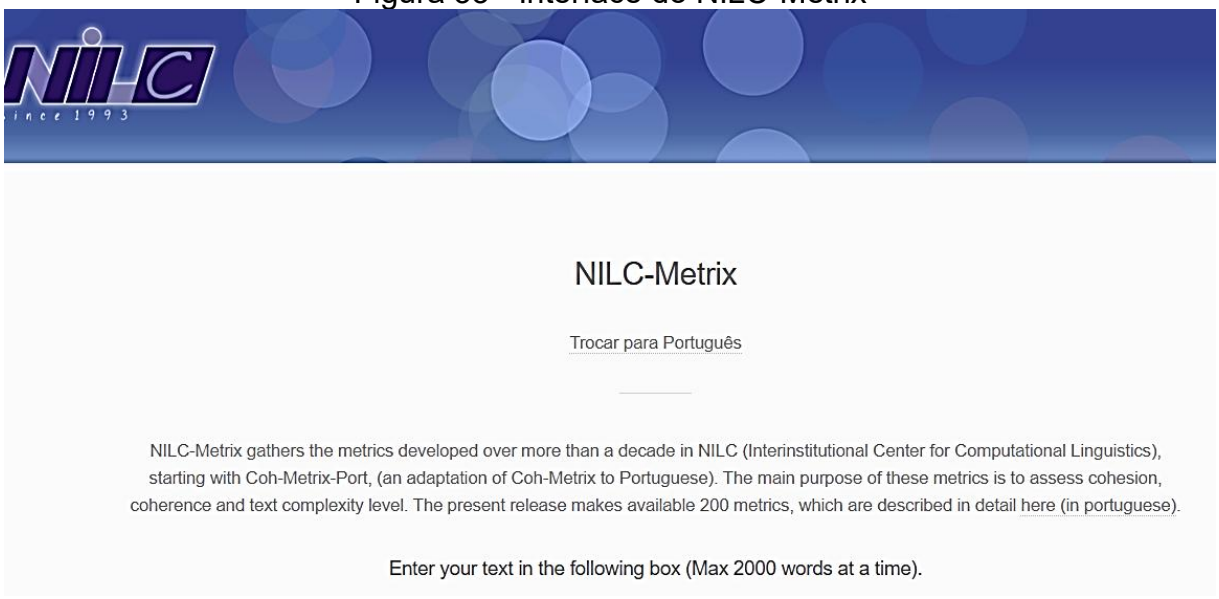
PLN, possibilitando a identificação de tendências e de padrões que não seriam plenamente compreendidos apenas com base nos números.

4.2 A FERRAMENTA NILC-METRIX[24]

A NILC-Metrix, ferramenta computacional principal desta pesquisa, é um conjunto de um pouco mais de 200 métricas coletadas ao longo de quase 16 anos, que funciona como um extrator e como uma facilidade computacional. A idealização e criação da ferramenta foram lideradas pelos professores/pesquisadores Sandra Aluísio, Sidney Leal, Carolina Scarton, Magali Duran e por um escopo de pesquisadores. As métricas são agrupadas em 14 conjuntos de referências e foram desenvolvidas durante o Projeto Coh-Metrix, pelo NILC (Núcleo Interinstitucional de Linguística Computacional) da USP (Universidade de São Paulo). As extrações realizadas pela NILC-Metrix são objetivas e aplicadas uniformemente a todos os textos. A ferramenta realiza uma análise multinível com o objetivo de fornecer índices que avaliem a coesão e a coerência dos textos. Originalmente utilizada para a avaliação da complexidade textual, a NILC-Metrix pode ser aplicada em diversas áreas, incluindo traduções. Além de oferecer uma análise descritiva, a ferramenta permite a criação de modelos computacionais (preditores), analisa automaticamente um texto por completo ou partes específicas dele.

[24] A ferramenta Nilc-Metrix pode ser encontrada no domínio: http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en. O Código-fonte também está disponível: https://github.com/nilc-nlp/nilcmetrix.git Under AGPLv3 license.

Figura 33 - Interface do NILC-Metrix



Fonte: http://fw.nilc.icmc.usp.br:23380/nilcmetrix

Quadro 17 - Caracterização dos grupos de referência do NILC-Metrix

Categoria:	Descrição:
Informações Semânticas de Palavras	Esta categoria agrupa métricas que analisam elementos semânticos, incluindo a porcentagem de palavras positivas e negativas, o grau de ambiguidades e a utilização de substantivos abstratos e nomes próprios em um texto.
Diversidade Lexical	Refere-se à variedade de vocabulário em um texto, medida pela relação entre o total de palavras únicas e o total geral. Um vocabulário mais diversificado geralmente reflete maior complexidade textual, enquanto uma baixa diversidade pode indicar maior coesão.
Informações Morfossintáticas de Palavras	Inclui métricas relacionadas à densidade e distribuição de elementos gramaticais, como adjetivos, advérbios, verbos, substantivos, preposições e pronomes, tanto por frase quanto por texto.
Medidas Psicolinguísticas	Analisa aspectos como a idade estimada em que as palavras são aprendidas, seu grau de concretude, familiaridade e capacidade de evocar imagens mentais, fornecendo 24 métricas que avaliam a facilidade de compreensão do texto.
Coesão Semântica	Concentra-se na conexão semântica entre as partes de um texto, considerando a similaridade entre palavras e frases com base em suas relações de significado. Essas métricas frequentemente utilizam modelos computacionais como Análise Semântica Latente (LSA).
Frequência de Palavras	Examina com que frequência as palavras aparecem em um texto, com base em dados de *corpora* como o BrWaC e o *Corpus*

	Brasileiro. As métricas são ajustadas por escala logarítmica para maior precisão.
Índices de Leiturabilidade	Reúne fórmulas clássicas, como índices de Flesch e Dale-Chall, que avaliam a dificuldade de leitura com base em fatores como o comprimento de palavras e sentenças e a presença de palavras incomuns.
Simplicidade Textual	Agrupa métricas que avaliam a facilidade de leitura de um texto. Inclui análises de proporção de sentenças de diferentes comprimentos, conexões entre frases e uso de vocabulário simples.
Medidas Descritivas	Analisam características gerais do texto, como o número de palavras, frases e parágrafos, além de calcular a média e o desvio-padrão do comprimento das sentenças e palavras. Essas métricas são úteis para identificar padrões de organização textual.
Complexidade Sintática	Foca nas estruturas gramaticais que impactam a complexidade do texto, como o uso de sentenças subordinadas, passivas ou com ordem sintática não convencional, além da quantidade de palavras antes do verbo principal.
Coesão Referencial	Avalia a utilização de elementos que conectam diferentes partes do texto, como pronomes e termos que se referem a ideias já mencionadas, ajudando a criar continuidade e clareza.
Conectivos	Foca na presença de palavras ou expressões que ligam ideias no texto, como 'além disso', 'entretanto' e 'portanto', com análises específicas para conectivos aditivos, causais, lógicos e temporais.
Léxico Temporal	Inclui métricas que detalham o uso de tempos e modos verbais no texto, como o presente, passado e subjuntivo, além da presença de conectivos temporais.
Densidade de Padrões Sintáticos	Examina estruturas gramaticais complexas, como a quantidade de palavras em sintagmas nominais, o uso de gerúndios e variações na extensão de frases nominais, que podem indicar maior dificuldade de processamento textual.

Fonte: http://fw.nilc.icmc.usp.br:23380/metrixdoc

É coerente indagarmos: “como as métricas oriundas da NILC-Metrix ajudam o pesquisador?” Elas ajudam na descoberta de como as características do texto se correlacionam com a compreensão da leitura, a partir dos seguintes aspectos:

1) Possibilitam que o professor identifique quais habilidades dos alunos precisam ser desenvolvidas naquela série e, assim, selecione textos com características que auxiliem os alunos na aprendizagem.

2) Permitem avaliar se o texto tem partes muito complexas e se ele deve ser simplificado para que o conteúdo do texto seja compreendido.

É importante salientar que a ferramenta possui algumas limitações. Por exemplo, ela não serve para avaliar a fluência de leitura ou quantas palavras a criança

consegue ler por minuto; não avalia a evolução nos níveis mais básicos do início da alfabetização; não ajuda a trabalhar com processos cognitivos, categorias psicológicas, que estão presentes, por exemplo, no LIWC[25] (Léxico categorizado).

Considerando a trajetória histórica, a imagem abaixo nos mostra uma linha do tempo com os principais projetos envolvidos na simplificação do português para a inclusão digital. Na sequência, farei menção específica a uma versão do PorSimples, do CohMetrix e trarei considerações sobre o índice Flesch.

Figura 34 - Projetos envolvidos na simplificação do português para a inclusão digital

Fonte: autoria própria

Dando ênfase ao projeto PorSimples[26], o seu principal objetivo era o de desenvolver tecnologias de Processamento de Linguagem Natural (PLN) relacionadas à Adaptação de Texto (AT) para promover inclusão digital e acessibilidade para pessoas com baixos níveis de alfabetização. Diversas ferramentas computacionais foram criadas ao longo da trajetória do PorSimples, ferramentas essas relacionadas à simplificação de textos, como o Facilita (WATANABE et al., 2009) e o Simplifica [27](SCARTON et al., 2010b).

---

[25] O LIWC é um programa de análise de texto que categoriza palavras em categorias derivadas de gramática e psicologia.

[26] Considerando o trabalho sobre leiturabilidade, destaca-se, no NILC, o projeto PorSimples (ALUÍSIO et al., 2008). Mais recentemente, o NILC desenvolveu o trabalho de Leal (2019) sobre medidas de complexidade textual, o que culminou na criação da ferramenta Simpligo (LEAL et al., 2019). O Simpligo encontra-se disponível para uso on-line na plataforma do NILC (LEAL, 2020).

[27] Disponível em: http://nilc.icmc.usp.br/porsimples/simplifica/

Leal (2019) possui um trabalho relevante sobre métodos de predição de complexidade de frases para o PB (Português brasileiro). Para tal, foram criados dois *corpora*, um com sentenças alinhadas pelo PorSimples, o PorSimplesSent (Leal et al., 2018), e outro com métricas de rastreamento ocular e normas de previsibilidade para estudantes de nível superior, o RastrOS[28] (Leal et al., 2022b). O estado-da-arte foi atingido por Leal (2019), que atingiu 97,5% de precisão. Baseando-se no melhor método desenvolvido, o autor, criou a aplicação Simpligo (Leal, 2020), que para cada sentença informada atribuiu um índice de complexidade individual.

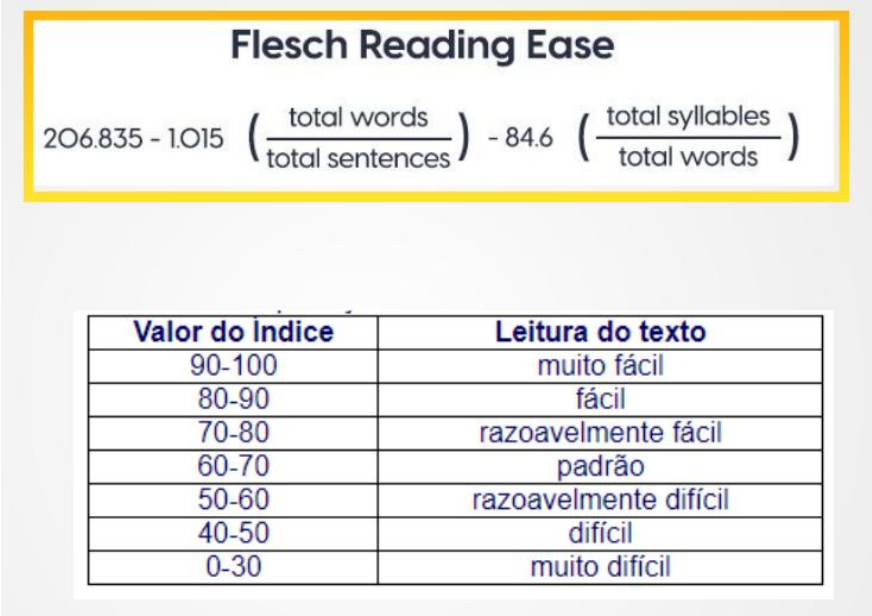
Figura 35 - Interface do projeto PorSimples

Fonte: http://fw.nilc.icmc.usp.br:23080/

Para analisar a inteligibilidade dos textos, as medidas principais são as seguintes fórmulas: os índices *Flesch Reading Ease* e o *Flesch-Kincaid Grade Level*. Apesar de serem superficiais, elas merecem destaque, pois a primeira é a única métrica de inteligibilidade já adaptada para o português (Martins et al., 1996) e incorpora o conceito de séries escolares da segunda. Essas métricas são consideradas superficiais, pois medem características relacionadas à superficialidade do texto, como o número de palavras em sentenças e o número de letras ou sílabas por palavra. Em termos gerais, a fórmula base para medição do Índice *Flesch Reading Ease* é esta:

[28] Disponível em: http://www.nilc.icmc.usp.br/nilc/index.php/rastros

Figura 36 - Fórmula do índice *Flesch Reading Ease*



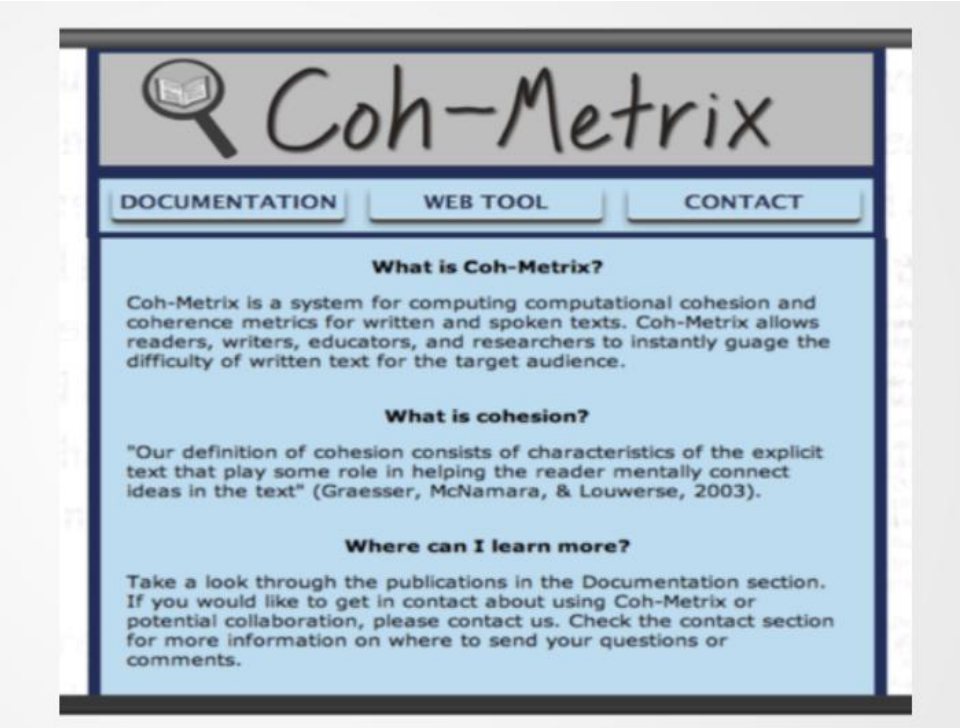
Valor do índice	Leitura do texto
90-100	muito fácil
80-90	fácil
70-80	razoavelmente fácil
60-70	padrão
50-60	razoavelmente difícil
40-50	difícil
0-30	muito difícil

Fonte: autoria própria

O índice *Flesch* adaptado para o português vale-se das mesmas variáveis (métricas), isto é, o comprimento das frases e o número de sílabas por palavra, mas, com algarismos diferentes. Os adaptadores da fórmula para o português brasileiro usaram diferentes procedimentos estatísticos para a adaptação da fórmula, considerando as diferentes palavras, tipos e tamanhos de frases.

É importante lembrar que, considerar um texto fácil de ser lido, ou não, requer uma visão global, não unilateral; há que se compreender o todo para o conhecimento do que é específico. Todavia, quando nos referimos às análises lexicais e/ou sintáticas de um texto podemos contar com previsões e ou estimativas de complexidade; a esses indícios chamamos de métricas de complexidade. Isto é, são estruturas microtextuais que precisam ser enxergadas num conjunto que compõe o texto, o que nos leva a entender que o que determina o potencial de complexidade de um texto é o conjunto de métricas, e não uma métrica isolada.

Figura 37 - CohMetrix



Fonte: http://143.107.183.175:22680/

Já o *Coh-Metrix-Port*, é uma adaptação da ferramenta Coh-Metrix (cujo foco é a Língua Inglesa) para o português brasileiro. A proposta da ferramenta é calcular índices para avaliar coesão, coerência e dificuldade de compreensão de um texto, servindo-se de vários níveis de análise linguística: lexical, sintático, discursivo e conceitual. Para a implementação dessas métricas, diferentes recursos e ferramentas de processamento de linguagem natural foram utilizados. A versão 2.0 do Coh-Metrix-Port apresenta 48 métricas.

É útil mencionar um importante trabalho de Scarton e Aluísio (2010), que exploram a inteligibilidade de textos por meio de ferramentas de Processamento de Línguagem Natural e da adaptação das métricas do Coh-Metrix para o Português; o estudo apresenta um projeto de adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil, ferramenta essa que, posteriormente, torna-se o NILC-Metrix.

4.3 CONSTITUIÇÃO DO *CORPUS* DA PESQUISA

A condução desta pesquisa foi realizada em etapas sequenciais, começando pelo cálculo amostral (*vide* figura 39) representativo da população (setecentas e vinte) de redações nota mil do ENEM no período de 2014 a 2023. Esse cálculo assegurou

que o *corpus* coletado refletisse, de forma estatisticamente válida, as características textuais de excelência avaliadas ao longo dos anos. Com base nesse cálculo, foram selecionadas 279 redações[29] (28 redações além do necessário), que compõem a amostragem desta pesquisa.

Diante da preocupação com o conceito de representatividade, considerei os aspectos que caracterizam as redações nota mil do ENEM, incluindo a sequência cronológica, os critérios de excelência e os temas abordados. Para que o *corpus* representasse fielmente os resultados produzidos pelos participantes que atingiram a nota máxima, houve a necessidade de compilar um conjunto de redações que fosse o mais representativo possível desse contexto. Assim, o *corpus* foi constituído por textos selecionados a partir de fontes públicas (digitais), com características de excelência textual exigidas nas avaliações do ENEM, haja vista que todas as notas dos textos são 1000 (mil).

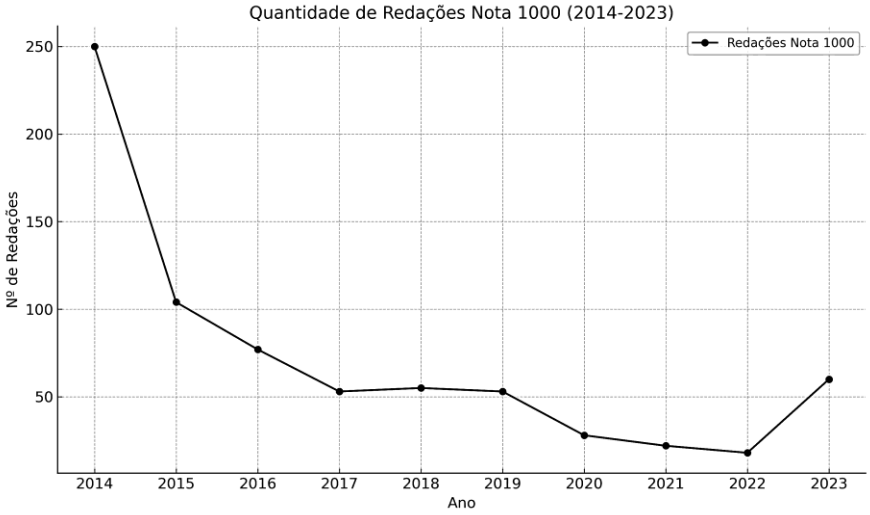
O *corpus*, portanto, compreende a modalidade escrita, formada exclusivamente por redações nota mil do ENEM, que representam um gênero textual específico dentro do contexto brasileiro de avaliação e de ingresso acadêmicos. Escolhi esse conjunto de textos com suas respectivas notas, em virtude da possível ausência de interferências externas no processo de escrita, garantindo que os textos reflitam as competências exigidas pela banca avaliadora do exame, a partir de uma análise rigorosa das características linguísticas e textuais que fundamentam as notas máximas.

Embora existam diversas iniciativas voltadas para a compilação de *corpora* de redações ou de textos acadêmicos disponíveis ao público, a obtenção de informações específicas, como as métricas linguísticas analisadas nesta pesquisa, enfrentou desafios consideráveis, pois, muitos *corpora* apresentam limitações de acesso, o que dificulta a realização de consultas mais avançadas. Para superar essas limitações e garantir a viabilidade das análises propostas, o *corpus* desta pesquisa foi compilado de forma direcionada, assegurando sua representatividade em relação às redações nota mil do ENEM e permitindo a extração precisa de métricas textuais essenciais para o estudo.

---

[29] Disponibilizo o corpus desta pesquisa, neste link: https://docs.google.com/document/d/1OicgKcD83jJpO7gNIFh4U0nMzE-VmGscicF1fkdI_4g/edit?usp=sharing

Figura 38 - Ilustração gráfica da quantidade de redações avaliadas com nota mil ao longo dos anos



Fonte: Sinopses Estatísticas do INEP

Frente ao exposto, fez-se necessário desenvolver um *corpus* específico, concebido especialmente para servir como material de análise e exploração das métricas linguísticas que caracterizam textos de excelência.

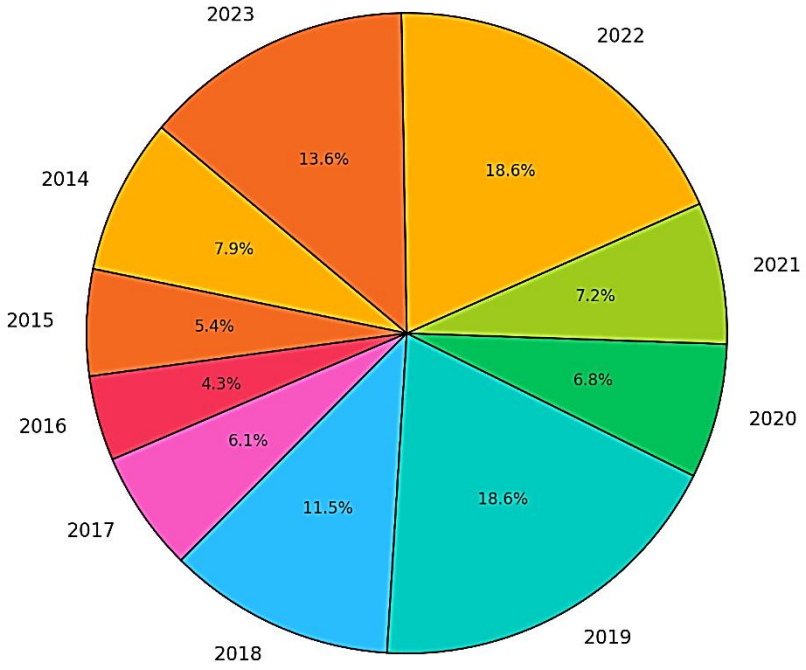
Tabela 3 - Quantitativo de redações avaliadas com nota mil considerando a população da pesquisa

Ano	Redações Nota Mil
2014	250
2015	104
2016	77
2017	53
2018	55
2019	53
2020	28
2021	22
2022	18
2023	60
	Total: 720

Fonte: autoria própria

O propósito inicial do *corpus* foi permitir sua exploração com base em diferentes princípios analíticos, utilizando métricas extraídas pela NILC-Metrix, evidenciando a coerência e a coesão textual, os cálculos estatísticos, identificação de padrões linguísticos recorrentes e outros elementos que contribuíssem para uma compreensão aprofundada das práticas textuais associadas à excelência esperada de um texto nota mil.

Gráfico 5 - Distribuição das redações coletadas, por ano, em % (Corpus da pesquisa)



Fonte: autoria própria

Tabela 4 - Distribuição de redações coletadas por ano (*corpus* da pesquisa)

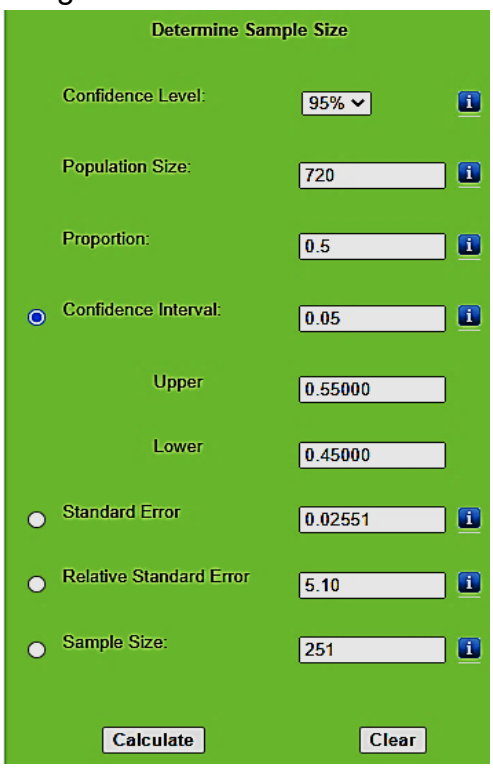
Ano	**Quantidade de Redações**
2014	22
2015	15
2016	12
2017	17
2018	32
2019	52
2020	19

2021	20
2022	52
2023	38
	Total: 279

Fonte: autoria própria

No que tange ao cálculo amostral, podemos entendê-lo como uma técnica utilizada para determinar o tamanho de uma amostra representativa de uma população, especialmente em situações em que analisar toda a população é inviável devido às limitações de tempo, custo ou recursos.

Figura 39 - Cálculo amostral



Fonte: Australian Bureau of Statistics (2024)

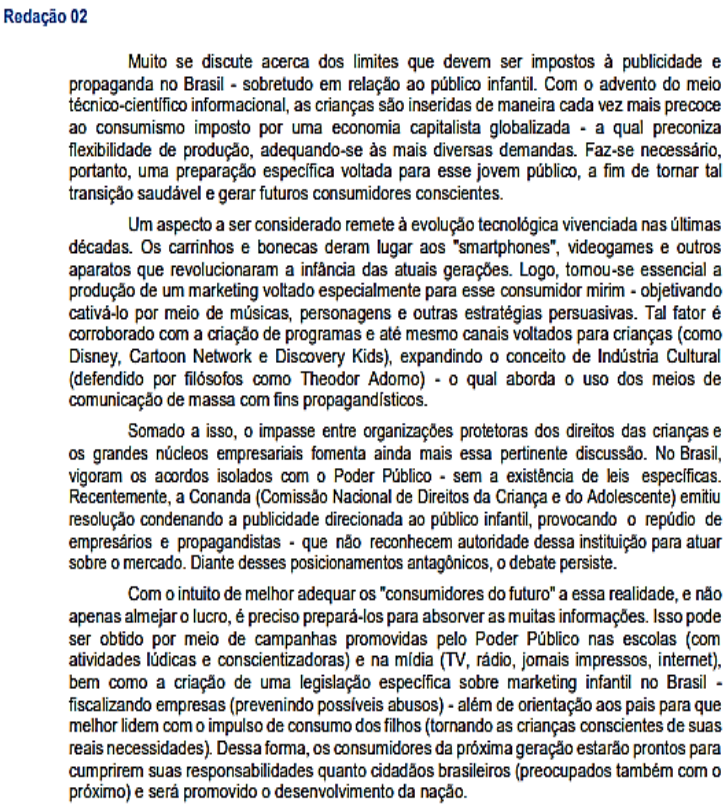
Para este estudo, no cálculo amostral, como aparece na figura 39, adotou-se um nível de confiança de 95%, que corresponde a um valor crítico de 1,96 na distribuição normal, uma margem de erro de 5%, e uma proporção esperada de 50%; essa última, maximiza a variabilidade e exige o maior tamanho amostral, proporcionando resultados abrangentes. O cálculo amostral para uma população finita é obtido por meio da seguinte fórmula:

$$n_{ajustado} = \frac{n}{1 + \frac{n-1}{N}} \quad (1)$$

Na etapa inicial do processamento do *corpus*, foi realizada a coleta das redações, priorizando fontes públicas que disponibilizam textos integrais de redações nota mil na Web. A maior parte do material encontrava-se no formato PDF, o que demandou um processo criterioso de organização e limpeza textual para tornar os textos adequados à análise automatizada. É importante lembrar de que para o uso adequado da ferramenta NILC-Metrix, é essencial que os textos estejam limpos; qualquer elemento que possa comprometer a análise, como símbolos matemáticos, legendas de figuras e tabelas mal posicionadas, deve ser removido. Além disso, as transcrições precisam ser devidamente pontuadas e segmentadas para garantirem a consistência dos dados. O princípio fundamental é: dados de má qualidade geram resultados de má qualidade.

Para viabilizar a manipulação dos dados pela ferramenta NILC-Metrix e garantir a legibilidade computacional, todos os textos foram convertidos para o formato de "texto simples" com extensão txt. A seguir, veremos duas imagens com a representação dos textos tanto em PDF quanto em txt.

Figura 40 - Redação 2/279 em formato PDF



Redação 02

Muito se discute acerca dos limites que devem ser impostos à publicidade e propaganda no Brasil - sobretudo em relação ao público infantil. Com o advento do meio técnico-científico informacional, as crianças são inseridas de maneira cada vez mais precoce ao consumismo imposto por uma economia capitalista globalizada - a qual preconiza flexibilidade de produção, adequando-se às mais diversas demandas. Faz-se necessário, portanto, uma preparação específica voltada para esse jovem público, a fim de tornar tal transição saudável e gerar futuros consumidores conscientes.

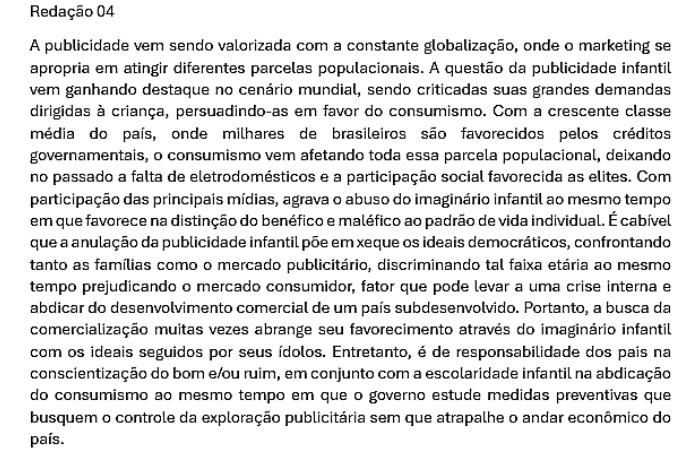
Um aspecto a ser considerado remete à evolução tecnológica vivenciada nas últimas décadas. Os carrinhos e bonecas deram lugar aos "smartphones", videogames e outros aparatos que revolucionaram a infância das atuais gerações. Logo, tomou-se essencial a produção de um marketing voltado especialmente para esse consumidor mirim - objetivando cativá-lo por meio de músicas, personagens e outras estratégias persuasivas. Tal fator é corroborado com a criação de programas e até mesmo canais voltados para crianças (como Disney, Cartoon Network e Discovery Kids), expandindo o conceito de Indústria Cultural (defendido por filósofos como Theodor Adorno) - o qual aborda o uso dos meios de comunicação de massa com fins propagandísticos.

Somado a isso, o impasse entre organizações protetoras dos direitos das crianças e os grandes núcleos empresariais fomenta ainda mais essa pertinente discussão. No Brasil, vigoram os acordos isolados com o Poder Público - sem a existência de leis específicas. Recentemente, a Conanda (Comissão Nacional de Direitos da Criança e do Adolescente) emitiu resolução condenando a publicidade direcionada ao público infantil, provocando o repúdio de empresários e propagandistas - que não reconhecem autoridade dessa instituição para atuar sobre o mercado. Diante desses posicionamentos antagônicos, o debate persiste.

Com o intuito de melhor adequar os "consumidores do futuro" a essa realidade, e não apenas almejar o lucro, é preciso prepará-los para absorver as muitas informações. Isso pode ser obtido por meio de campanhas promovidas pelo Poder Público nas escolas (com atividades lúdicas e conscientizadoras) e na mídia (TV, rádio, jornais impressos, internet), bem como a criação de uma legislação específica sobre marketing infantil no Brasil - fiscalizando empresas (prevenindo possíveis abusos) - além de orientação aos pais para que melhor lidem com o impulso de consumo dos filhos (tornando as crianças conscientes de suas reais necessidades). Dessa forma, os consumidores da próxima geração estarão prontos para cumprirem suas responsabilidades quanto cidadãos brasileiros (preocupados também com o próximo) e será promovido o desenvolvimento da nação.

Fonte: organizado pela autora.

Figura 41 - Texto limpo para o processamento; redação nota mil – 2014



Redação 04

A publicidade vem sendo valorizada com a constante globalização, onde o marketing se apropria em atingir diferentes parcelas populacionais. A questão da publicidade infantil vem ganhando destaque no cenário mundial, sendo criticadas suas grandes demandas dirigidas à criança, persuadindo-as em favor do consumismo. Com a crescente classe média do país, onde milhares de brasileiros são favorecidos pelos créditos governamentais, o consumismo vem afetando toda essa parcela populacional, deixando no passado a falta de eletrodomésticos e a participação social favorecida as elites. Com participação das principais mídias, agrava o abuso do imaginário infantil ao mesmo tempo em que favorece na distinção do benéfico e maléfico ao padrão de vida individual. É cabível que a anulação da publicidade infantil põe em xeque os ideais democráticos, confrontando tanto as famílias como o mercado publicitário, discriminando tal faixa etária ao mesmo tempo prejudicando o mercado consumidor, fator que pode levar a uma crise interna e abdicar do desenvolvimento comercial de um país subdesenvolvido. Portanto, a busca da comercialização muitas vezes abrange seu favorecimento através do imaginário infantil com os ideais seguidos por seus ídolos. Entretanto, é de responsabilidade dos pais na conscientização do bom e/ou ruim, em conjunto com a escolaridade infantil na abdicação do consumismo ao mesmo tempo em que o governo estude medidas preventivas que busquem o controle da exploração publicitária sem que atrapalhe o andar econômico do país.

Fonte: organizado pela autora

Após a coleta, as redações foram submetidas ao processamento na NILC-Metrix, a submissão dos textos não foi unitária foi total. Esse processamento total (todas as redações de uma só vez) foi facilitado devido à colaboração do pesquisador Dr. Sidney Leal, (um dos idealizadores da ferramenta), com a minha pesquisa.

4.4 SOFTWARE ESTATÍSTICO

Para que a análise estatística dos dados acontecesse de forma acurada, realizei a normalização dos dados e um teste de homogeneidade (*vide* subseção 5.3); utilizei a linguagem de programação *Python* 3 associada à interface online *Google Colab,* [30]que torna possível a execução de códigos em nuvem, sem a necessidade de instalação de *software* ou de configuração de biblioteca local. Para tal, utilizei as seguintes bibliotecas:

Figura 42 - Bibliotecas *Python* aplicadas à pesquisa

Fonte: autoria própria

4.5 O CAMINHO METODOLÓGICO PERCORRIDO

O método adotado nesta pesquisa consiste em um conjunto articulado de procedimentos quantitativos aplicados aos dados extraídos do *corpus.* A finalidade central foi identificar padrões linguístico-textuais consistentes e recorrentes entre os

[30] https://colab.research.google.com/

textos. Foram aplicadas técnicas de normalização, de estatísticas descritivas e teste de homogeneidade de variância. Em seguida, foi realizada uma interpretação linguística das métricas textuais, com base nos valores de referência dispostos na ferramenta NILC-Metrix e na matriz de correção e de competências do Enem.

Estas foram as etapas da análise estatística:

Figura 43 - Etapas da Análise Estatística

Fonte: autoria própria..

4.5.1 Normalização dos dados com Z-score

As métricas extraídas do NILC-Metrix apresentam naturezas e escalas distintas. Algumas métricas, por exemplo, expressam proporções que variam de 0 a 1, ao passo que outras são constituídas por valores absolutos. Ressalta-se que essa heterogeneidade pode ser um dificultador da análise comparativa direta, uma vez que diferenças em escalas podem induzir a interpretações estatísticas equivocadas, privilegiando artificialmente métricas com maior magnitude ou com maior dispersão natural.

A fim de superar esse obstáculo e de garantir homogeneidade na escala de análise, empreguei uma técnica estatística de normalização por escore-padrão, também conhecida como *z-score,* que transforma os valores de cada métrica em uma nova escala com média zero e com desvio-padrão 1, de tal modo que todas as métricas sejam avaliadas sob os mesmos parâmetros de dispersão relativa. Esta é a fórmula utilizada:

$$Z = \frac{X - \mu}{\sigma}$$

Na qual:

- **X** representa o valor original da métrica para uma determinada redação;
- **μ (mi)** é a **média** dos valores dessa métrica no corpus;
- **σ (sigma)** é o **desvio padrão** correspondente.

As principais vantagens que obtive com essa padronização englobam a eliminação dos efeitos das diferentes escalas de medição, a preservação da distribuição original dos dados e a possibiidade de interpretações estatísticas robustas; ademais, facilitou a comparação direta entre as métricas com propriedades distintas, tais como proporções, contagens e médias.

4.5.2 Estatísticas descritivas e seleção de métricas

Após a normalização dos dados, foram calculadas as principais estatísticas descritivas (média, desvio-padrão, coeficiente de variação (CV), valores mínimos e máximos, e intervalo interquartil), ilustradas na figura 43. O principal critério para a seleção das métricas mais estáveis foi o CV (coeficiente de variação) cuja fórmula é: CV = (Desvio padrão / Média) × 100.

É importante destacar que, embora a ferramenta NILC-Metrix disponibilize um conjunto superior a 200 métricas linguísticas, optei por trabalhar com um recorte de 90 métricas selecionadas com base em critérios de consistência estatística. Tomei essa decisão metodológica a fim de garantir a viabilidade analítica da investigação, visto que a análise de um volume tão elevado de variáveis (200), sem redução

fundamentada, comprometeria a clareza interpretativa e a objetividade dos resultados do estudo. As 90 métricas com menor CV foram consideradas as mais consistentes entre os textos analisados e, portanto, as mais adequadas e representativas para os padrões linguísticos recorrentes das redações nota mil do Enem. Essa seleção foi feita considerando a hipótese da pesquisa (*vide* 1.4.2), que afirma que as redações nota mil apresentam padrões linguísticos consistentes. Logo, as métricas com menor variação foram as que revelaram traços recorrentes e, portanto, configuram pistas confiáveis como marcas de alto desempenho.

Figura 44 - Estatísticas descritivas (recorte)

	Média	Desvio Padrão	Tam. (n)	Mínimo	Máximo	Amplitude	Coef. de variação
adverbs	0,046623779	0,000565016	279	0	0	0	0%
content_words	0,024932462	126,2935685	279	5,07673	5,74074	0,66401	2%
flesch	0,010491755	84,03305032	279	4,01176	4,85831	0,84655	3%
function_words	0,004278946	2,284943782	279	0,74971	1	0,25029	4%
sentences_per_paragraph	0,101177472	0,469593573	279	0,20697	0,95715	0,75018	7%
syllables_per_content_word	0,102378006	0,467256043	279	0,20362	0,94993	0,74631	7%
words_per_sentence	0,021114652	0,466537638	279	0,19527	0,93133	0,73606	7%
noun_ratio	0,296231363	2,155027469	279	0,72031	5,05136	4,33105	7%
paragraphs	31,36683374	1,680655728	279	0,60612	4,12	3,51388	7%
sentences	3,921240039	1,088829015	279	0,5216	4,34752	3,82592	8%
words	0,008546678	1,744485353	279	0,64516			
pronoun_ratio	18,78222028	3,466743558	279	1,27729			
verbs	0,066666956	1,556837174	279	0,60412			
logic_operators	2,155027469	0,344275934	279	0,06977			
and_ratio	0,089212532	0,143431394	279	0,00468			
if_ratio	0,024862145	0,874749871	279	0,4			
or_ratio	0,057734225	0,281580039	279	0,04016			
negation_ratio	0,059837998	0,63534187	279	0,29268			

Fonte: autoria própria

É útil compreendermos o que cada estatística descritiva nos revela. A média, por exemplo, expressa o valor médio de cada métrica no conjunto de textos. Já o desvio padrão indica o grau de dispersão dos valores em torno da média, ou seja,

quanto maior esse valor, maior a variação da métrica no corpus. O coeficiente de variação (CV), por sua vez, a estatística-chave para esta pesquisa, nos fornece a variação relativa da métrica em relação à sua média, ele é expresso em porcentagem e nos permite identificar as métricas mais estáveis, isto é, quanto menor o CV, maior a regularidade da métrica entre os textos.

A título de ilustração, se a média de uma métrica é **0,05** e o desvio padrão é **0,002**, o CV é de **4%,** o que indica baixa variabilidade e alta estabilidade. Também foram considerados os valores mínimo e máximo, que indicam os limites inferior e superior registrados para cada métrica, bem como a amplitude, que representa a diferença entre esses dois extremos. Embora não esteja ilustrado diretamente na imagem, o intervalo interquartil (IQR) — diferença entre os quartis Q3 e Q1 — foi utilizado para detectar a concentração de valores no centro da distribuição e para identificar possíveis dispersões atípicas.

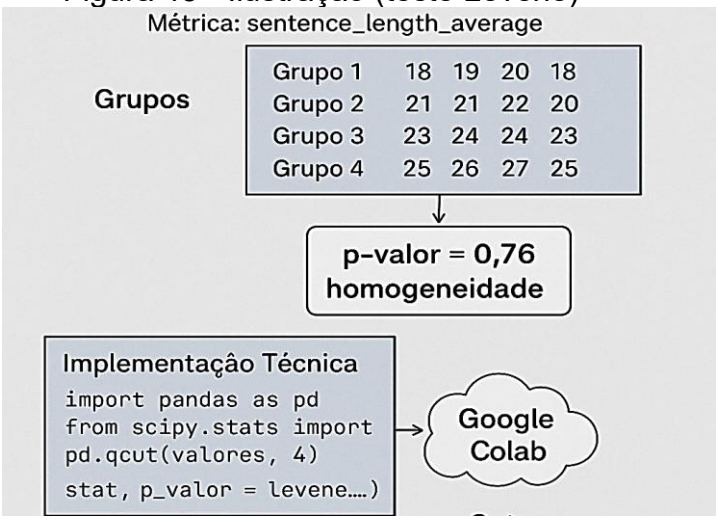
4.5.3 Teste de Levene: homogeneidade de variância

Para confirmar estatisticamente a estabilidade das métricas selecionadas após a análise descritiva, apliquei o teste de Levene, [31]que tem por finalidade verificar a homogeneidade das variâncias entre diferentes grupos de dados. No contexto desta pesquisa, os 279 textos foram segmentados em quartis, ou seja, divididos em quatro grupos de igual tamanho com base nos valores de cada métrica. O teste de Levene foi, então, aplicado a cada métrica para investigar se a sua variabilidade interna permanecia constante entre os grupos. A interpretação estatística do teste considera que, quando o p-valor é superior a 0,05, não há diferença significativa entre as variâncias, indicando que a métrica se comporta de maneira homogênea ao longo do corpus e pode ser considerada estatisticamente estável. À guisa de ilustração, pode-se citar a métrica *sentence_length_average*: se os valores observados em quatro grupos forem, por exemplo, [18, 19, 20, 18], [21, 21, 22, 20], [23, 24, 24, 23] e [25, 26,

[31] O teste de Levene foi realizado nesta pesquisa por meio da linguagem Python, utilizando a função `levene()` da biblioteca `scipy.stats`. Esse teste também pode ser executado em softwares estatísticos como R (com a função `leveneTest()` da biblioteca `car`), SPSS (na análise de variância, ativando a opção 'Test of Homogeneity of Variances') e Jamovi, uma plataforma estatística gratuita com interface intuitiva. O Excel não possui essa funcionalidade nativamente, mas é possível simular o teste por meio de fórmulas complementares.

27, 25], o teste retornaria um p-valor elevado **(como 0,76),** sinalizando homogeneidade entre os grupos.

Figura 45 - Ilustração (teste Levene)



Fonte: autoria própria

Abaixo, apresento o código utilizado nesta pesquisa para aplicar o teste de Levene nas métricas extraídas do corpus. O código, como dito, foi implementado na linguagem *Python,* utilizando as bibliotecas pandas e scipy.stats, executado no ambiente *Google Colab:*

Quadro 18 - Código de aplicação do teste de Levene

```
from scipy.stats import levene
import pandas as pd

df = pd.read_csv("jeane_279redacoes_metricas.csv")
coluna = "noun_ratio"
valores = df[coluna]
quartis = pd.qcut(valores, 4, labels=False, duplicates='drop')
grupos = [valores[quartis == i].values for i in range(4)]

stat, p_valor = levene(*grupos)
print(f"Estatística de Levene: {stat:.4f}")
print(f"p-valor: {p_valor:.4f}")
```

Fonte: autoria própria

5 ANÁLISE DOS DADOS

A partir das métricas geradas pelo NILC-Metrix e dos cálculos estatísticos realizados (o percurso metodológico que foi descrito na seção anterior), foram identificados os elementos de maior variação entre as redações. Isso me permitiu selecionar os aspectos mais relevantes para análise, direcionando a pesquisa para os elementos textuais que apresentam maior impacto na diferenciação das redações nota mil. Com as métricas selecionadas, iniciou-se a interpretação qualitativa dos dados, buscando compreender como as variações observadas se relacionam com as características estruturais das redações de excelência.

Os dados resultantes foram organizados e visualizados por meio de tabelas e gráficos, elaborados no MS Excel. Com base nos resultados obtidos, foram discutidos os achados e elaboradas conclusões que destacam as contribuições descritiva desta pesquisa para o ensino de redação e para o uso de ferramentas de processamento de linguagem natural.

Figura 46 - Relação de métricas processadas pelo NILC-Metrix (recorte)[32]

	B	C	D	E	F	G	H
1	adjective_rati	adverbs	content_word	flesch	function_word	sentences_per_paragrap	syllables_per_content_wo
2	0,02907	0,56977	6,04979	0,43023	2,6	3,5	26,46154
3	0,0223	0,56134	17,51529	0,43866	1,6	3,22517	33,625
4	0,09296	0,04271	0,61055	9,34085	0,38945	3	3,30453
5	0,125	0,02041	0,57143	18,25198	0,42857	3,6	3,33036
6	0,12227	0,04367	0,59825	7,72648	0,40175	1,6	3,36496
7	0,09974	0,0315	0,60105	15,66187	0,39895	3	3,25328
8	0,09863	0,03288	0,59726	18,0047	0,40274	3,2	3,26147
9	0,10526	0,04986	0,59834	24,74196	0,40166	2,5	3,12963
10	0,08571	0,05055	0,58901	17,51322	0,41099	3,4	3,22015
11	0,11932	0,0303	0,57008	3,88757	0,42992	8,5	3,45847
12	0,09982	0,02773	0,57301	0,06721	0,42699	7	3,4
13	0,09833	0,04603	0,56485	-17,03664	0,43515	5,5	3,65556
14	0,07741	0,01883	0,5523	4,28238	0,4477	7,5	3,55682
15	0,09884	0,04264	0,58527	6,69072	0,41473	8,5	3,40066
16	0,09874	0,04622	0,57773	14,39481	0,42227	9,5	3,30545
17	0,09957	0,01515	0,57143	15,42751	0,42857	8,5	3,32197
18	0,10465	0,01395	0,56047	-6,51211	0,43953	6	3,62241
19	0,08805	0,0283	0,59119	21,93587	0,40881	3	3,20213
20	0,08925	0,04189	0,57923	9,22832	0,42077	9	3,34591
21	0,10458	0,02832	0,57081	1,22122	0,42919	6,5	3,4313
22	0,08984	0,02734	0,56445	2,52443	0,43555	7,5	3,51557
23	0,06954	0,02398	0,56115	7,59642	0,43885	7,5	3,4188
24	0,10219	0,03893	0,58151	7,0229	0,41849	7	3,37657
25	0,08758	0,02648	0,58248	10,59571	0,41752	9	3,33566

Fonte: organizado pela autora

Ademais, a título de acréscimo, foi delimitada a medida de riqueza lexical presente no *corpus*. A type-token ratio (TTR), que como é explicado em seções

[32] Os valores de todas as métricas geradas a partir do corpus estão neste link: 200 métricas_jeane_279_redacoes_metricas (1).xlsx

anteriores, é uma das medidas de riqueza lexical mais utilizada por pesquisadores nas áreas de Linguística de *Corpus* e Processamento de Linguagem Natural.

O cálculo dessa medida é relativamente simples: divide-se o número de palavras únicas de um texto (types ou tipos) pelo número total de palavras nesse mesmo texto (tokens). Por exemplo, considere um texto com 100 palavras no total (100 tokens), no qual se verifica que 65 palavras são diferentes (65 types). A TTR é calculada dividindo-se o número de types pelo de tokens, ou seja, 65/100. Para facilitar a interpretação, o resultado é multiplicado por 100, resultando em 65%. Esse valor, expresso em forma percentual, demonstra que o texto possui uma significativa variação vocabular, uma vez que 65% das palavras não se repetem. Veja a fórmula:

$$TTR = \frac{\text{types}}{\text{tokens}} \times 100 \tag{2}$$

Na análise do *corpus* utilizado neste estudo foram processados 121.518 *tokens* e identificados 14.430 *types*, resultando em um *type-token* ratio (TTR) de 11,87%. Esses dados foram sistematizados na Tabela 4, que apresenta a relação das métricas processadas. A inclusão do TTR como indicador quantitativo é relevante, pois evidencia a riqueza lexical presente nas redações analisadas, para a compreensão da variedade e repetição lexical no conjunto de textos. Essa métrica, ao lado de outras ferramentas linguísticas, aprofunda a análise detalhada dos padrões de escrita no contexto do Enem, especialmente, o repertório lexical escolhido pelos estudantes na construção textual.

Tabela 5 - Medida da Riqueza lexical do *corpus*

Corpus	***Tokens***	***Types***	***Type-token ratio* (TTR).**
279 redações	121518	14430	11, 87%

Fonte: autoria própria

5.1 PROCESSAMENTO DE DADOS NO NILC-METRIX E ORGANIZAÇÃO DAS MÉTRICAS OBTIDAS

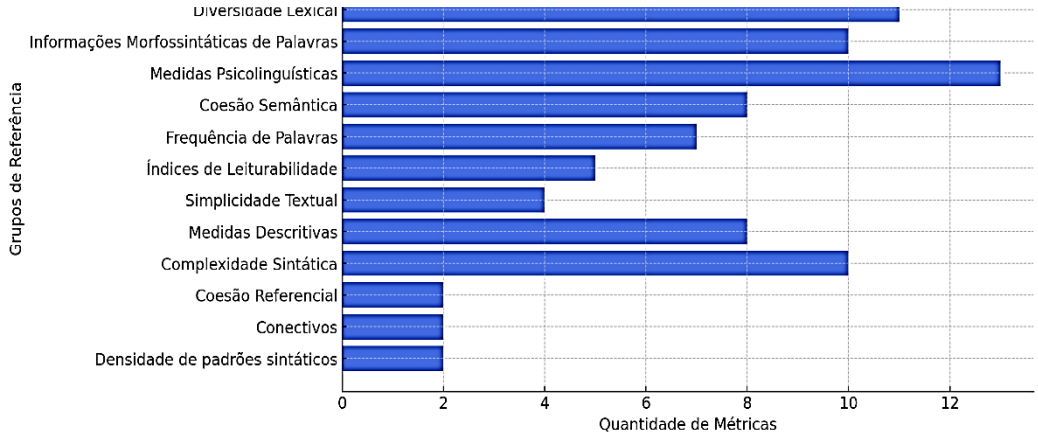
As métricas indicadas pela análise dos dados do presente estudo estão organizadas em 13 grandes grupos, conforme a natureza das análises e as técnicas de processamento de linguagem natural (PLN) empregadas, de acordo com Leal *et al.* (2022).

Quadro 19 - Métricas de variação mais significativa *(no corpus)* e seus respectivos grupos de referência

Grupos de referência (NILC-Metrix)	**Quantidade de métricas**
Informações Semânticas de Palavras	10
Diversidade Lexical	9
Informações Morfossintáticas de Palavras	9
Medidas Psicolinguísticas	10
Coesão Semântica	9
Frequência de Palavras	8
Índices de Leiturabilidade	4
Medidas Descritivas	7
Complexidade Sintática	3
Coesão Referencial	3
Conectivos	5
Densidade de padrões sintáticos	3
	Total: 90

Fonte: autoria própria

Gráfico 6 - Distribuição de métricas por grupo de referência, no NILC-Metrix



Fonte: autoria própria

O gráfico 6 nos mostra a organização das métricas utilizadas, categorizadas em diferentes grupos de referência, conforme os princípios do NILC-Metrix. Os dados nos apontam uma distribuição de métricas em 14 (quatorze) grupos distintos. Destaca-se o grupo "Medidas Psicolinguísticas", que concentra o maior número de métricas (12), seguido pelo da "Diversidade Lexical" com 10 (dez) métricas, o de "Informações Morfossintáticas de Palavras" com 10 (dez) métricas e o de “Complexidade Sintática”, também com 10 (dez). Outros grupos relevantes incluem o de “Informações Semânticas das Palavras”, com 9 (nove); os de "Coesão Semântica" e o de “Medidas Descritivas”, com 8 (oito) cada; o de “Frequência de Palavras”, com 7 (sete); o de “Índices de Leiturabilidade”, com 5 (cinco); o de “Simplicidade Textual”, com 4 (quatro); e o de “Coesão Referencial”, com 3 (três). Os demais grupos, a saber: “Conectivos””, e “Densidade de Padrões Sintáticos””, trazem duas métricas cada.

O gráfico 6 reflete, ainda, a ênfase dada a diferentes dimensões da análise textual, como semântica, diversidade lexical, e aspectos psicolinguísticos e morfossintáticos, evidenciando a abrangência e a diversidade do *corpus* analisado. A análise multinível mencionada na subseção 4.2 foi corroborada. Dos 14 grupos de referência (categorias) presentes na ferramenta NILC-Metrix, vê-se que somente um não possui um valor de métrica significativo: o “léxico temporal”.

No quadro 17, elenquei as principais características de cada grupo de referência, assim como nos mostra a ferramenta, a fim de elucidar e descrever quais métricas do NILC-Metrix mostraram maior consistência ou variação nas redações nota mil do

Enem. Por conseguinte, apresentarei um esboço geral das métricas que foram significativamente variadas na análise do *corpus*.

Quadro 20 - Visualização/caracterização das 90 métricas mais relevantes no *corpus* da pesquisa

Métrica	Descrição:
abstract_nouns_ratio	Proporção de substantivos abstratos em relação ao total de substantivos.
add_pos_conn_ratio	Razão de conectores aditivos positivos em relação ao total de palavras.
adjective_diversity_ratio	Razão entre o número de adjetivos únicos e o total de adjetivos.
adjective_ratio	Proporção de adjetivos em relação ao total de palavras no texto.
adjectives_ambiguity	Avalia o grau de ambiguidade associado aos adjetivos.
Adverbs	Contagem total de advérbios no texto.
adverbs_ambiguity	Grau de ambiguidade dos advérbios, avaliando múltiplos significados.
adverbs_diversity_ratio	Razão entre o número de advérbios únicos e o total de advérbios.
arg_ovl	Sobreposição de argumentos no texto, medindo a repetição de elementos.
cau_pos_conn_ratio	Razão de conectores causais no texto.
concretude_25_4_ratio	Proporção de palavras com concretude entre 2,5 e 4.
concretude_mean	Média de concretude das palavras no texto.
concretude_std	Desvio padrão da concretude das palavras no texto.
content_word_diversity	Diversidade de palavras de conteúdo no texto.
content_word_max	Frequência máxima de palavras de conteúdo no texto.
content_words	Número total de palavras de conteúdo no texto.
content_words_ambiguity	Grau de ambiguidade associado às palavras de conteúdo.
cw_freq_bra	Frequência média das palavras de conteúdo no *corpus* brasileiro.
cw_freq_brwac	Frequência média das palavras de conteúdo no *corpus* BrWaC.
dalechall_adapted	Índice de legibilidade adaptado de Dale-Chall.
dep_distance	Distância de dependência sintática média no texto.
easy_conjunctions_ratio	Razão de conjunções subordinativas no texto.
familiaridade_25_4_ratio	Proporção de palavras com familiaridade entre 2,5 e 4.
familiaridade_4_55_ratio	Proporção de palavras com familiaridade entre 4 e 5,5.
familiaridade_55_7_ratio	Proporção de palavras com familiaridade entre 5,5 e 7.
familiaridade_mean	Média da familiaridade das palavras no texto.
familiaridade_std	Desvio padrão da familiaridade das palavras.
Flesch	Índice de legibilidade Flesch, que mede a facilidade de leitura.
freq_bra	Frequência média das palavras no *corpus* brasileiro.
freq_brwac	Frequência média das palavras no *corpus* BrWaC.
function_words	Número total de palavras funcionais no texto.
gunning_fox	Índice de legibilidade Gunning-Fog, que mede a complexidade sintática.
hard_conjunctions_ratio	Razão de conjunções coordenativas no texto.
Honore	Medida de riqueza lexical conhecida como Estatística de Honoré.
idade_aquisicao_25_4_ratio	Proporção de palavras adquiridas entre 2,5 e 4 anos.
idade_aquisicao_4_55_ratio	Proporção de palavras adquiridas entre 4 e 5,5 anos.
idade_aquisicao_55_7_ratio	Proporção de palavras adquiridas entre 5,5 e 7 anos.

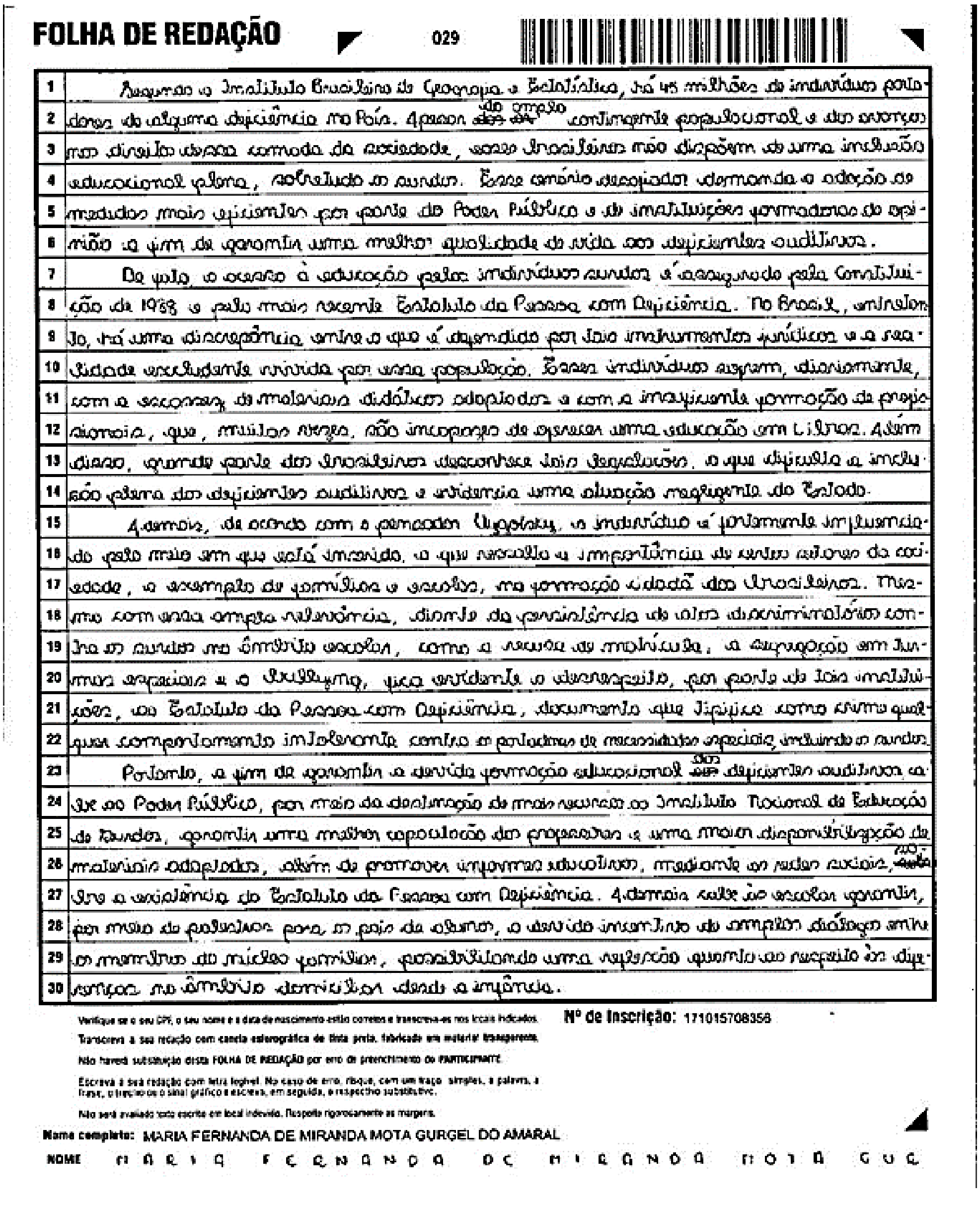
idade_aquisicao_mean	Média da idade de aquisição das palavras no texto.
idade_aquisicao_std	Desvio padrão da idade de aquisição das palavras.
imageabilidade_25_4_ratio	Proporção de palavras com imagemabilidade entre 2,5 e 4.
imageabilidade_mean	Média de imagemabilidade das palavras no texto.
logic_operators	Proporção de operadores lógicos no texto.
long_sentence_ratio	Razão de sentenças longas em relação ao total de sentenças.
lsa_adj_mean	Similaridade semântica média entre adjetivos.
lsa_adj_std	Desvio padrão da similaridade semântica em adjetivos.
lsa_all_mean	Similaridade semântica média geral (LSA).
lsa_givenness_mean	Similaridade semântica média relacionada à 'dadação'.
lsa_givenness_std	Desvio padrão da similaridade semântica relacionada à 'dadação'.
lsa_span_mean	Similaridade semântica média entre spans do texto.
max_noun_phrase	Comprimento máximo de um sintagma nominal.
min_cw_freq	Frequência mínima das palavras de conteúdo no texto.
min_cw_freq_bra	Frequência mínima das palavras de conteúdo no *corpus* brasileiro.
min_cw_freq_brwac	Frequência mínima das palavras de conteúdo no *corpus* BrWaC.
min_freq_bra	Frequência mínima de palavras no *corpus* brasileiro.
min_freq_brwac	Frequência mínima no *corpus* BrWaC.
min_noun_phrase	Comprimento mínimo de um sintagma nominal.
negative_words	Contagem de palavras com carga negativa.
noun_diversity	Diversidade de substantivos no texto.
noun_ratio	Proporção de substantivos em relação ao total de palavras.
nouns_ambiguity	Grau de ambiguidade associado aos substantivos.
Paragraphs	Número total de parágrafos no texto.
preposition_diversity	Diversidade de preposições utilizadas no texto.
pronoun_diversity	Diversidade de pronomes utilizados no texto.
pronoun_ratio	Proporção de pronomes em relação ao total de palavras.
punctuation_diversity	Diversidade de sinais de pontuação usados no texto.
punctuation_ratio	Razão de pontuação em relação ao total de palavras.
ratio_function_to_content_words	Razão entre palavras funcionais e palavras de conteúdo.
sentence_length_max	Comprimento máximo de uma sentença no texto.
sentence_length_standard_deviation	Desvio padrão do comprimento das sentenças.
Sentences	Número total de sentenças no texto.
sentences_per_paragraph	Média de sentenças por parágrafo.
simple_word_ratio	Proporção de palavras simples em relação ao total de palavras.
std_noun_phrase	Desvio padrão do comprimento dos sintagmas nominais.
syllables_per_content_word	Média de sílabas por palavra de conteúdo.
Ttr	Índice de Type-Token Ratio, que mede a diversidade lexical.
verb_diversity	Diversidade de verbos utilizados no texto.
Verbs	Número total de verbos no texto.
verbs_ambiguity	Grau de ambiguidade associado aos verbos.
Words	Número total de palavras no texto.
words_per_sentence	Média de palavras por sentença.
Yngve	Métrica que avalia a profundidade sintática do texto segundo a hierarquia de dependência.
lsa_paragraph_mean	Similaridade semântica média entre os parágrafos do texto.
indefinite_pronouns_diversity	Diversidade no uso de pronomes indefinidos.
and_ratio	Proporção do conectivo “e” em relação ao total de palavras.
mean_noun_phrase	Média de expressões nominais (sintagmas nominais) por sentença.
pronouns_standard_deviation	Grau de variação no uso de pronomes ao longo do texto.

Fonte: http://fw.nilc.icmc.usp.br:23380/metrixdoc#flesch

5.1.1 Exemplo de processamento na NILC-Metrix no *Corpus* analisado

A seguir, apresento o processamento minucioso de uma das redações do *corpus* desta pesquisa , a fim de ilustrar como algumas métricas aparecem nos textos.

Figura 47 - Redação 50 do *corpus* (PDF)



FOLHA DE REDAÇÃO 029

Segundo o Instituto Brasileiro de Geografia e Estatística, há 45 milhões de indivíduos portadores de alguma deficiência no País. Apesar do amplo contingente populacional e dos avanços nos direitos dessa camada da sociedade, esses brasileiros não dispõem de uma instrução educacional plena, sobretudo os surdos. Esse cenário desafiador demanda a adoção de medidas mais eficientes por parte do Poder Público e de instituições formadoras de opinião a fim de garantir uma melhor qualidade de vida aos deficientes auditivos.

De fato, o acesso à educação pelos indivíduos surdos é assegurado pela Constituição de 1988 e pelo mais recente Estatuto da Pessoa com Deficiência. No Brasil, entretanto, há uma discrepância entre o que é defendido por tais instrumentos jurídicos e a realidade excludente vivida por essa população. Esses indivíduos sofrem, diariamente, com a escassez de materiais didáticos adaptados e com a insuficiente formação de profissionais, que, muitas vezes, são incapazes de oferecer uma educação em Libras. Além disso, grande parte dos brasileiros desconhece tais legislações, o que dificulta a inclusão plena dos deficientes auditivos e evidencia uma atuação negligente do Estado.

Ademais, de acordo com o pensador Vygotsky, o indivíduo é fortemente influenciado pelo meio em que está inserido, o que ressalta a importância de certos setores da sociedade, a exemplo de famílias e escolas, na formação cidadã dos brasileiros. Mesmo com essa ampla relevância, diante da persistência de atos discriminatórios contra os surdos no âmbito escolar, como a recusa de matrícula, a segregação em turmas especiais e o bullying, fica evidente o desrespeito, por parte de tais instituições, ao Estatuto da Pessoa com Deficiência, documento que tipifica como crime qualquer comportamento intolerante contra os portadores de necessidades especiais, incluindo os surdos.

Portanto, a fim de garantir a devida formação educacional dos deficientes auditivos, cabe ao Poder Público, por meio da destinação de mais recursos ao Instituto Nacional de Educação de Surdos, garantir uma melhor capacitação dos professores e uma maior disponibilização de materiais adaptados, além de promover informes educativos, mediante as redes sociais, sobre a existência do Estatuto da Pessoa com Deficiência. Ademais, cabe às escolas garantir, por meio de palestras para os pais de alunos, o devido incentivo de amplos diálogos entre os membros do núcleo familiar, possibilitando uma reflexão quanto ao respeito às diferenças no âmbito domiciliar desde a infância.

Verifique se o seu CPF, o seu nome e a data de nascimento estão corretos e transcreva-os nos locais indicados.
Transcreva a sua redação com caneta esferográfica de tinta preta, fabricada em material transparente.
Não haverá substituição desta FOLHA DE REDAÇÃO por erro de preenchimento do PARTICIPANTE.
Escreva a sua redação com letra legível. No caso de erro, risque, com um traço simples, a palavra, a frase, o trecho ou o sinal gráfico e escreva, em seguida, o respectivo substitutivo.
Não será avaliado texto escrito em local indevido. Respeite rigorosamente as margens.

Nº de Inscrição: 171015708356

Nome completo: MARIA FERNANDA DE MIRANDA MOTA GURGEL DO AMARAL

NOME M A R I A F E R N A N D A D E M I R A N D A M O T A G U G

Fonte: *corpus* da pesquisa

Analisaremos a seguir a transcrição integral da redação selecionada, cuja temática era 'Desafios para a formação educacional de surdos no Brasil', produzida em 2017.

Redação 50

Segundo o Instituto Brasileiro de Geografia e Estatística, há 45 milhões de indivíduos portadores de alguma deficiência no País. Apesar do amplo contingente populacional e dos avanços nos direitos dessa camada da sociedade, esses brasileiros não dispõem de uma inclusão educacional plena, sobretudo os surdos. Esse cenário desafiador demanda a adoção de medidas mais eficientes por parte do Poder Público e de instituições formadoras de opinião a fim de garantir uma melhor qualidade de vida aos deficientes auditivos. De fato, o acesso à educação pelos indivíduos surdos é assegurado pela Constituição de 1988 e pelo mais recente Estatuto da Pessoa com Deficiência. No Brasil, entretanto, há uma discrepância entre o que é defendido por tais instrumentos jurídicos e a realidade excludente vivida por essa população. Esses indivíduos sofrem, diariamente, com a escassez de materiais didáticos adaptados e com a insuficiente formação de profissionais, que, muitas vezes, são incapazes de oferecer uma educação em Libras. Além disso, grande parte dos brasileiros desconhece tais legislações, o que dificulta a inclusão plena dos deficientes auditivos e evidencia uma atuação negligente do Estado. Ademais, de acordo com o pensador Vygotsky, o indivíduo é fortemente influenciado pelo meio em que está inserido, o que ressalta a importância de certos setores da sociedade, a exemplo de famílias e escolas, na formação cidadã dos brasileiros. Mesmo com essa ampla relevância, diante da persistência de atos discriminatórios contra os surdos no âmbito escolar, como a recusa de matrícula, a segregação em turmas especiais e o bullying, fica evidente o desrespeito que justifica como crime qualquer comportamento intolerante contra os portadores de necessidades especiais, incluindo os surdos. Portanto, a fim de garantir a devida formação educacional dos deficientes auditivos, cabe ao Poder Público, por meio da destinação de mais recursos ao Instituto Nacional de Educação de Surdos, garantir uma melhor capacitação dos professores e uma maior disponibilização de materiais adaptados, além de promover informes educativos, mediante as redes sociais, sobre a existência do Estatuto da Pessoa com Deficiência. Ademais, cabe às escolas garantirem, por meio de palestras para os pais de alunos, o devido incentivo de amplos diálogos entre os membros do núcleo familiar, possibilitando uma reflexão quanto ao respeito às diferenças no âmbito domiciliar desde a infância.

A redação do exemplo acima (redação 50) foi processada no NILC-Metrix. Para melhor compreensão, vejamos alguns detalhes desse processamento.

Figura 48 - Métricas geradas no processamento – NILC-Metrix

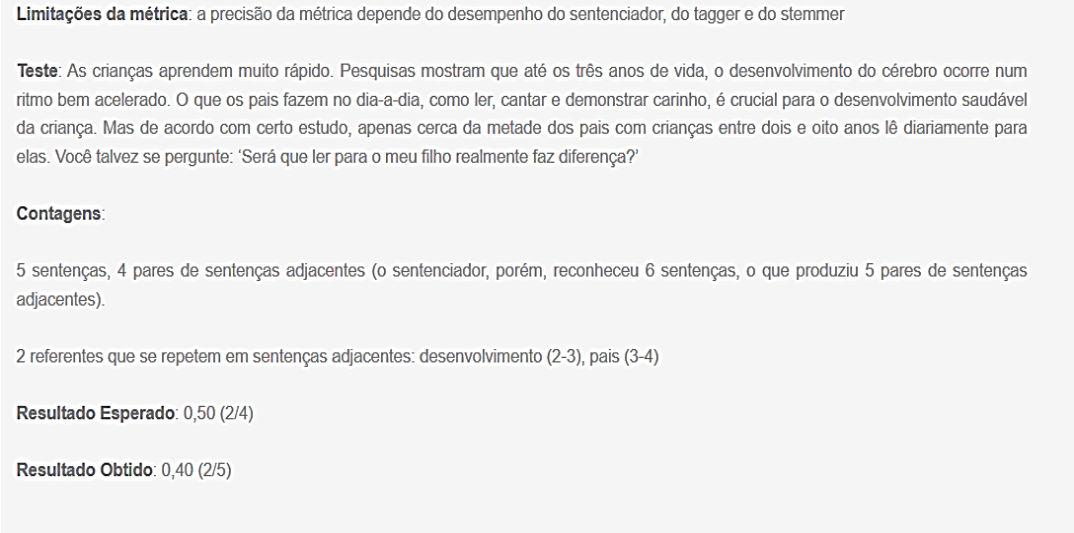
Resultados

	Group	Metric	Description	Value
1	Coesão Referencial	adj_arg_ovl	Quantidade média de referentes que se repetem nos pares de sentenças adjacentes do texto	0.90909
2	Coesão Referencial	adj_cw_ovl	Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto	0.72727
3	Coesão Referencial	adj_stem_ovl	Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes do texto	0.81818
4	Coesão Referencial	adjacent_refs	Média das proporções de candidatos a referentes na sentença anterior em relação aos pronomes pessoais do caso reto nas sentenças	0.0
5	Coesão Referencial	anaphoric_refs	Média das proporções de candidatos a referentes nas 5 sentenças anteriores em relação aos pronomes anafóricos das sentenças	0.0

Fonte: http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en

À guisa de exemplificação da análise apresentada, vejamos algumas métricas reveladas no texto. A interpretação delas se deu pelos valores de referência dispostos na própria ferramenta. Para cada métrica são estabelecidos valores esperados. Vejam como esse valor aparece:

Figura 49 - Contagem e resultado da métrica de exemplo (adj_arg_ovl)



Limitações da métrica: a precisão da métrica depende do desempenho do sentenciador, do tagger e do stemmer

Teste: As crianças aprendem muito rápido. Pesquisas mostram que até os três anos de vida, o desenvolvimento do cérebro ocorre num ritmo bem acelerado. O que os pais fazem no dia-a-dia, como ler, cantar e demonstrar carinho, é crucial para o desenvolvimento saudável da criança. Mas de acordo com certo estudo, apenas cerca da metade dos pais com crianças entre dois e oito anos lê diariamente para elas. Você talvez se pergunte: 'Será que ler para o meu filho realmente faz diferença?'

Contagens:

5 sentenças, 4 pares de sentenças adjacentes (o sentenciador, porém, reconheceu 6 sentenças, o que produziu 5 pares de sentenças adjacentes).

2 referentes que se repetem em sentenças adjacentes: desenvolvimento (2-3), pais (3-4)

Resultado Esperado: 0,50 (2/4)

Resultado Obtido: 0,40 (2/5)

Fonte: http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en

As métricas que aparecem no exemplo de processamento da figura 46 são estas:

Tabela 6 - Exemplificação dos valores de algumas métricas

1.adj_arg_ovl.	48. and_ratio	72.pronoun_diversity	198. Flesch
1.0	**0.02439**	**0.54545**	**3.86953**

Fonte: organizado pela autora

A métrica 1 *(1.adj_arg_ovl.),* representada na tabela 6, que mensura a média de referentes repetidos entre pares de sentenças consecutivas, obteve o valor **1.0** no *corpus*. Tal valor indica que, em média, um referente é retomado a cada par de sentenças adjacentes, ou seja, podemos inferir que há uma coesão referencial consistente ao longo do texto.

Já a métrica 48 *(48. and_ratio),* apresentou um valor de **0,02439**. No teste fornecido pela ferramenta, um trecho de 38 palavras com apenas um operador lógico ("e") resultou em um valor de 0,026. Já no texto analisado, do *corpus*, a métrica

apresentou valor muito próximo. O valor sugere que, a cada aproximadamente 40 palavras, um operador lógico foi utilizado, o que configura uma frequência baixa.

A métrica *pronouns_uniq_ratio* mede a razão entre o número de pronomes únicos (sem repetições) e o total de pronomes presentes no texto. No teste da ferramenta, temos um trecho com 10 pronomes totais e 9 únicos, resultando em um valor de 0,90, considerado alto; indica grande variedade e refinamento no uso de pronomes, com poucas repetições. No texto do *corpus,* o valor registrado foi **0,54545,** o que significa que aproximadamente 54,5% dos pronomes usados são diferentes entre si, o que indica que há repetição significativa de referentes pronominais ao longo do texto.

A última métrica que coloquei como exemplo *(198. Flesch),* está relacionada aos índices de leiturabilidade. O valor indicado pela métrica para o texto de exemplo, **3.86953,** traz um valor extremamente baixo dentro da escala de legibilidade; isso indica que o texto é muito difícil de ser lido (*vide* figura 36), talvez pelo fato de possuir frases longas, palavras complexas ou baixa segmentação conteudinal.

Essas foram métricas exemplificadas para entendermos como a NILC-Metrix nos apresenta os valores e como eles podem ser interpretados sobre um texto. A análise do *corpus* desta pesquisa parte de todos os valores numéricos que foram gerados pela ferramenta nos níveis sintáticos, lexicais, morfológicos, psicolinguísticos, de simplicidade textual, de conectivos etc.

Para a convergência entre a análise estatística e a interpretação das métricas, a padronização por *z-score*, como mencionada na seção “Material e Método” permitiu comparar métricas como, por exemplo, *sentence_length_average,* com média de 20,5 palavras por sentença e desvio de 3,2, a outras como *adjectives_ambiguity,* que variava de 0 a 1. Após a normalização de todos os dados, foram calculadas estatísticas descritivas *(vide* figura 44) para todas as métricas, especialmente o coeficiente de variação (CV), que foi o critério principal de seleção. Por exemplo, *a métrica cw_freq* (frequência de palavras de conteúdo) apresentou um CV de **36,50%,** sugerindo alta variação entre os textos quanto à densidade de palavras informativas. Já a métrica *freq_brwac,* relacionada à frequência média de palavras no corpus BrWaC, apresentou apenas **1,93%** de variação, revelando um padrão bastante homogêneo. A métrica *adverbs_max,* por sua vez, registrou um CV de **43,88%,** indicando que o número máximo de advérbios em uma sentença variou amplamente entre os textos.

Figura 50 - Exemplos de valores de variação no *corpus*

Métrica	Variação (%)
lsa_all_mean	8,81%
lsa_paragraph_mean	50,08%
cross_entropy	17,22%
lsa_adj_std	37,25%
lsa_span_mean	6,78%
lsa_givenness_mean	6,844%
lsa_span_std	19,77%
idade aquisicao std	11,59%
concretude_25_4_ratio	9,01%
imageabilidade_4_55_ratio	14,77%
concretude_4_55_ratio	17,42%
familiaridade_4_55_ratio	10,05%
idade aquisicao_mean	7,85%
concretude_std	11,76%
familiaridade_mean	7,39%
content_word_diversity	4,37%
type_token_ratio	8,18%
pronoun_diversity	18,41%
content_word_min	10,73%
verb_diversity	9,22%
content_word_max	8,57%
adiective_diversity_ratio	11,38%
preposition_diversity	13,16%

Fonte: organizado pela autora

Quando o teste de Levene foi aplicado, foi possível verificar a homogeneidade da variância entre os quartis dos dados. Tomando como exemplo a métrica *noun_ratio*, os textos foram divididos em quatro grupos (Q1 a Q4) com base em seus valores. O teste de Levene foi então utilizado para comparar a dispersão entre esses quartis. A título ilustrativo, para a métrica sentence_length_average, o valor do teste foi: Levene = **0,5879; p = 0,6241**, indicando variância homogênea, considerando o p-valor superior a 0,05.

Tabela 7 - Exemplo de resultados do teste de Levene

Métrica	**Levene**	**p-valor**	**Interpretação**

sentence_length_average	0,5879	0,6241	Variância homogênea (p > 0,05)
noun_ratio	1,9876	0,0432	Variância heterogênea (p < 0,05)

Fonte: dados da pesquisa

Como visto na subseção anterior, a 5.1.1, a interpretação individual das métricas se efetuou por medidas e parâmetros que estão dispostos na própria NILC-Metrix. Embora na subseção “Material e Método” tenha sido descrito o percurso metodológico da pesquisa, considerei pertinente detalhá-lo um pouco mais para que a articulação dos procedimentos seja facilmente compreendida nas subseções sequentes, referentes à interpretação das variações.

5.2 INTERPRETAÇÃO DAS VARIAÇÕES DAS MÉTRICAS EXTRAÍDAS DO CORPUS

Nos termos elucidados, baseando-me nas verificações estatísticas até aqui feitas, respondo às duas questões de pesquisa, das três delimitadas para este estudo (*vide* subseção 1.4.1).

A primeira questão respondida foi: 1. Nas redações nota mil do Enem, existe variação entre os valores mínimos e máximos nas métricas fornecidas pelo NILC-Metrix? Os cálculos de amplitude de variação mostraram que mesmo em redações cuja nota foi igual, existe variação entre os valores das métricas.

Por conseguinte, a segunda questão foi: 2. Quais métricas do NILC-Metrix mostraram maior consistência ou variação nas redações nota mil do Enem? A amplitude de variação obtida pela diferença entre o valor máximo e o valor mínimo de cada métrica, indicou resultados variados, alguns, inclusive, com valores na classe dos milhões, que extrapolaram o limite da pesquisa.

A partir dos valores obtidos foi possível comprovar que algumas métricas possuem variações mais significativas e mais consistentes do que outras. A interpretação dessas variações será detalhada a partir de agora, para que de modo subsequente, a terceira questão desta pesquisa seja respondida, no que diz respeito

ao alinhamento (ou não) das métricas do NILC-Metrix (nas redações nota mil) com as competências avaliativas do Enem. Faz-se útil reiterar que, as métricas a serem descritas e interpretadas já estão delimitadas no quadro 20 desta pesquisa.

As interpretações que serão feitas daqui para frente estão fundamentadas nos parâmetros postos pela própria ferramenta, nos valores esperados para cada métrica dentro do NILC-Metrix; a partir deles, podemos entender como a variação se reflete no *corpus*, e o que isso significa em termos linguístico-textuais.

Conforme abordado anteriormente, a análise se pautará em 90 métricas; todavia, ressalto que a exclusão das demais não implica desconsiderar sua relevância teórica ou descritiva. A opção metodológica pela delimitação deve-se ao elevado número total de métricas extraídas, bem como à ocorrência de valores nulos ou extremos em parte delas, o que poderia comprometer a consistência estatística da análise. Assim, optei por um recorte que privilegia as métricas com maior estabilidade e potencial interpretativo no contexto do presente estudo. A ocorrência desses valores em determinadas métricas pode ser atribuída a limitações inerentes à própria ferramenta.

Conforme descrito na documentação da própria ferramenta, tais desvios são reflexos da arquitetura sobre a qual ela foi construída, a qual se apoia em diferentes camadas de análise — como segmentadores, etiquetadores morfossintáticos (*PoS taggers*) e parsers sintáticos — que, embora apresentem níveis elevados de acurácia, não são infalíveis. Por exemplo, parsers de constituintes, amplamente utilizados em ferramentas de PLN para o português, possuem taxas de acerto em torno de 90%, o que implica uma margem considerável de erro residual. Quando essas imprecisões ocorrem na base da cadeia de processamento — como a segmentação incorreta de palavras ou a etiquetagem falha de categorias gramaticais — elas podem ser propagadas para as camadas superiores, afetando diretamente o cálculo de determinadas métricas. Como consequência, algumas dessas métricas podem ser zeradas ou assumir valores inconsistentes muito baixos ou altos (*outliers*), distorcendo as medidas de tendência central e dispersão.

Reitero, neste momento, a relevância da continuidade desta pesquisa - tanto por mim quanto por outros investigadores - no que tange à análise abrangente de todas as métricas para fins específicos.

Por conseguinte, cada métrica a ser descrita e interpretada pertence a uma categoria diferente, que na ferramenta aparece como grupo de referência. Essa

categorização pode ser relembrada no quadro 16, desta pesquisa. Ainda que no quadro 20, vê-se uma visualização/caracterização das 90 métricas que foram mais relevantes no *corpus* da pesquisa, farei questão de reexplicá-las concomitantemente à demonstração dos valores obtidos.

5.2.1 Medidas descritivas

Partindo do grupo de “Medidas Descritivas”, obtive algumas métricas de variação considerável. A categoria em questão quantifica sílabas por palavra, palavras por sentença, sentenças por parágrafo, além de contagens absolutas de palavras, sentenças e parágrafos. As métricas foram:

- *Words:* essa métrica faz a contagem de palavras no texto; a variação nos indicou que há um coeficiente de variação de **19,44%** entre os textos mais curtos e os mais longos do *corpus*. Podemos entender isso como uma significativa regularidade textual, que nos termos de extensão (número de palavras/tokens) indica uma homogeneidade textual.
- A métrica *sentences_per_paragraph* mede o número médio de frases por parágrafo. Nessa métrica, o valor de variação entre os textos foi de **47, 91%**. O resultado de variação obtido demonstra que há estratégias distintas na construção dos parágrafos: alguns textos agrupam ideias em parágrafos mais longos e densos; outros, optam por parágrafos mais enxutos e concisos. Essa diversidade parece nos indicar que o alto desempenho textual pode se manifestar tanto na concisão quanto na exploração mais ampla de ideias por parágrafo, desde que exista coesão e desenvolvimento temático, aspectos esses, vinculados à competência 3
- *Sentence_length_max:* essa métrica está relacionada ao comprimento máximo de uma sentença. A variação de **22, 70%** nas palavras na sentença mais longa do texto, indicou que algumas redações possuem sentenças máximas significativamente mais extensas do que outras. Percebe-se que a variação foi razoável, o que evidencia uma possível liberdade estilística entre os candidatos. Ainda assim, o alto desempenho foi mantido (haja vista, a nota mil). A variação nessa métrica parece não comprometer a qualidade do texto, mas aponta diferentes estratégias sintáticas adotadas pelos candidatos.

- *Sentence_length_standard_deviation: o* desvio-padrão do comprimento das sentenças é medido por essa métrica; a variação foi de **26, 02%**, sugerindo que há uma diversidade regular nas formas de distribuição do tamanho das sentenças nas redações. Constata-se que os autores das redações adotaram, de algum modo, algumas estratégias distintas de controle do ritmo textual. A alternância bem conduzida de estruturas sintáticas complexas e simples pode ser um indicativo de domínio linguístico, aspecto diretamente vinculado à competência 1.
- A métrica *words_per_sentence* mensura a média de palavras por sentença. O valor foi de **19, 48%,** o que nos sugere diferenças moderadas na extensão média das sentenças. Textos com maior média tendem a adotar períodos mais complexos, enquanto os de menor média indicam maior concisão.
- *long_sentence_ratio:* a métrica avalia a proporção de sentenças consideradas longas (geralmente definidas por um dado número de palavras) em relação ao total de sentenças no texto. A variação foi de **10,84,** sugerindo uma distinção considerável entre as redações, na frequência de sentenças longas.
- *simple_word_ratio: essa* métrica calcula a proporção de palavras consideradas simples (palavras curtas ou comuns) em relação ao total de palavras no texto. A variação obtida foi de **9, 18%,** alguns textos parecem ter priorizado a simplicidade lexical, ao passo que outros preferiram um estilo mais elaborado.

5.2. 2 Coesão Referencial

No grupo de “Coesão Referencial”, que avalia quantidades de palavras de conteúdo, radicais de conteúdo, medidas de coesão referencial e a proporção de repetições lexicais, essas foram as métricas:

- *content_word_diversity*: essa métrica representa o grau de diversidade lexical entre as palavras de conteúdo (substantivos, verbos, adjetivos e advérbios) utilizadas no texto; ela apresentou um coeficiente de variação de **70%,** indicando alta variabilidade dessa métrica ao longo do corpus. O CV nos revelou que, mesmo entre textos nota mil, o uso de diversidade lexical não é uniforme. Ou seja, há produções com maior variedade de palavras de conteúdo, enquanto outras se mantêm mais restritas lexicalmente, sem que isso tenha comprometido, necessariamente, a nota máxima atribuída.

- Ainda no mesmo grupo, as métricas *content_word_min* e *content_word_max*, que indicam respectivamente o menor e o maior número de palavras de conteúdo identificadas em unidades textuais (como sentenças ou parágrafos), apresentaram um coeficiente de variação de **71%.** O valor indicou uma alta amplitude no comportamento dessas métricas ao longo do corpus analisado. Em outras palavras, há redações nota mil que parecem apresentar passagens com uso mínimo de palavras de conteúdo bastante reduzido, enquanto outras alcançam picos máximos de densidade lexical.

5.2.3 Coesão Semântica

O próximo grupo é o da "Coesão Semântica". Essa categoria calcula aspectos como: similaridade, entropia[33], desvio-padrão, taxa de gêneros linguísticos, além de modelos baseados em LSA[34] para medir coesão. Neste grupo, algumas métricas indicaram uma variação mais significativa entre os textos, são elas:

- *lsa_all_mean:* a métrica calcula a similaridade semântica média entre todas as palavras do texto, entendida como coesão global; essa métrica se baseia na Análise Semântica Latente (LSA). A variação obtida foi de **8,81%,** sugerindo que parte dos textos mantém uma consistência temática mais coesa, já outros, apresentam maior dispersão semântica. Essa estabilidade na progressão semântica está diretamente ligada à competência quatro, que avalia a coesão e a continuidade argumentativa.
- *lsa_paragraph_mean:* essa métrica teve uma variação de **50, 08%.** É uma métrica que mede a similaridade semântica média entre os parágrafos do texto. A variação obtida indica diferenças expressivas na coesão semântica global entre os textos. Ou seja, parte das redações apresenta forte conexão temática entre os parágrafos, enquanto outras evidenciam parágrafos mais independentes semanticamente.
- *cross_entropy:* essa métrica mede a imprevisibilidade lexical do texto em relação a um modelo de linguagem. A variação não tão expressiva de **17,22%** sugere

[33] Na linguística, entropia é a medida de informação discursiva ponderada em função da quantidade de lexemas.
[34] Análise Semântica Latente (LSA, em inglês - http://lsa.colorado.edu/) foi adotada no Coh-Metrix e também no NILC Metrix como uma medida de coesão e de coerência semânticas.

que alguns textos apresentam construções mais previsíveis, enquanto outros parecem inovar mais lexicalmente.

- *lsa_adj_std:* essa métrica avalia a consistência na utilização de adjetivos em termos semânticos. A variação foi de **37, 25%,** indicando uma diferença moderada na forma como os adjetivos contribuíram para a coesão semântica nos textos do *corpus*. Ou seja, alguns textos exploraram intensamente a nuance semântica dos adjetivos e outros mantiveram o uso mais repetitivo ou restrito, o que pode estar atrelado ao grau de precisão e de refinamento na caracterização de ideias.
- *lsa_span_mean:* Essa métrica mede a coesão semântica entre os segmentos consecutivos do texto. A variação foi de **6, 78%,** é relativamente baixa, o que nos mostra que a coesão entre segmentos é bastante consistente no *corpus*, evidenciando poucas distinções entre os textos, nesse aspecto.
- *lsa_givenness_mean:* a métrica nos mostra a similaridade semântica média relacionada à 'dadação' (introdução de informações novas *versus* informações já conhecidas no texto). A variação nessa métrica foi de **6, 84%**, indicando que os candidatos mantiveram estratégias consistentes de progressão informativa nos textos, o que pode favorecer a coesão textual.
- *lsa_span_std:* essa métrica indica variabilidade na coesão semântica entre segmentos do texto. A variação de **19, 77%,** aponta para diferenças moderadas no que condiz à consistência da coesão entre segmentos nos textos do *corpus*. Ou seja, alguns textos apresentam coesão uniforme, enquanto outros apresentam flutuações.
- *lsa_givenness_std:* essa métrica avalia o desvio-padrão relacionado à consistência na introdução de informações novas *versus* conhecidas. A variação dessa métrica foi de **39, 33%,** pode ser considerada alta, indicando que há textos com forte continuidade temática e outros com saltos discursivos mais evidentes, o que se relaciona diretamente com a competência quatro, que avalia a progressão textual e a articulação entre as partes.
- *lsa_adj_mean:* essa métriva mede a coesão semântica específica entre os adjetivos. Nos textos, a variação foi de **7, 20%**, sugerindo-nos uma diferença não tão significativa na forma como os adjetivos contribuíram para a coesão semântica nos textos; podemos inferir, assim, que alguns poucos textos utilizaram adjetivos de maneira mais coesa do que outros.

5.2.4 Medidas Psicolinguísticas

O outro grupo de referência é o de “Medidas Psicolinguísticas”, que verifica quatro critérios principais no texto: concretude, imagemabilidade, familiaridade e idade de aquisição de palavras. Para esse grupo, sobressaíram-se tais métricas:

- idade_aquisicao_std: essa métrica mede o desvio-padrão da idade de aquisição das palavras no texto. Isto é, ela indica a variabilidade na idade em que as palavras do texto são geralmente adquiridas pelos falantes. A variação foi de **11,59%,** o que nos sugere que há diferenças moderadas entre as redações em termos da consistência na complexidade lexical. Em outras palavras, podemos considerar que alguns textos podem utilizar um vocabulário com idades de aquisição mais homogêneas, enquanto alguns outros apresentam maior heterogeneidade.
- A métrica *concretude_25_4_ratio*, que mede a proporção de palavras concretas aprendidas entre 2,5 e 4 anos, teve um valor de **9, 01%,** sugerindo que, de modo geral, as redações podem fazer uso de um vocabulário mais abstrato e formal.
- A métrica *concretude_mean* teve um valor de variação **7, 42%.** É uma métrica que mensura a média de concretude das palavras. A concretude das palavras mede o quanto os termos utilizados remetem a conceitos tangíveis (como "mesa" e "casa") ou abstratos (como "justiça" e "ética"). O valor sugere pouca variação no balanceamento entre palavras concretas e abstratas.
- *imageabilidade_4_55_ratio:* é uma métrica que mensura a proporção de palavras altamente imagéticas (fáceis de visualizar mentalmente) aprendidas até os 4,55 anos. A variação foi relativamente baixa, de **14,77%.** Textos com maior uso dessas palavras tendem a ser mais acessíveis e concretos, facilitando a compreensão.
- *concretude_4_55_ratio:* essa métrica mede a proporção de palavras concretas adquiridas até os 4,5 anos. A variação foi de **17,42%,** sugerindo algumas diferenças entre textos mais ancorados em termos concretos (facilitando a compreensão) e outros mais abstratos (mais conceituais).
- A métrica *familiaridade_4_55_ratio* obteve um valor de variação de **10, 05%.** Essa métrica quantifica a proporção de palavras aprendidas entre 4 e 5,5 anos. Mede a quantidade de palavras que fazem parte do vocabulário básico da língua. O valor obtido sugere que também há um balanceamento entre palavras simples e mais formais, com pouca variação.

- *idade_aquisicao_mean:* essa métrica reflete a média das idades em que as palavras do texto são adquiridas. A variação foi de **7, 85%,** podendo ser considerada como moderada; ela aponta para um padrão estável de seleção lexical entre os candidatos.
- *concretude_std:* essa métrica está relacionada ao desvio-padrão da concretude das palavras. A variação foi de variação **11, 76%,** sugerindo uma diferença de baixa a moderada entre as redações, no que diz respeito, à consistência do uso de palavras concretas *versus* abstratas.
- *familiaridade_mean:* essa métrica indica o grau médio de familiaridade das palavras para os falantes nativos. A variação foi de **7, 39%** e aponta para diferenças sutis entre as redações na escolha de palavras mais ou menos comuns, influenciando a facilidade de leitura dos textos.
- *imageabilidade_mean:* por fim, neste grupo de medidas psicolinguísticas, essa métrica reflete a média da capacidade das palavras de evocarem imagens mentais. A variação foi de **7, 62%,** sugerindo que existe alguma diferença (ainda que baixa) entre as redações, no que concerne à utilização de palavras que facilitam a formação de imagens mentais.

5.2.5 Diversidade Lexical

A próxima categoria recebe o nome de “diversidade lexical”, ela contém métricas que medem a riqueza lexical com índices de diversidade, de razão entre palavras e de proporção de palavras lexicais e funcionais presentes no texto. Nessa categoria, obtive várias métricas consideráveis. Farei uma análise panorâmica, a partir de então.

De modo geral, a análise da diversidade lexical e estrutural das redações do *corpus* revelou variações consideráveis em métricas distintas, apontando para uma heterogeneidade textual no que diz respeito às escolhas lexicais, ao uso de palavras funcionais e de conteúdo, assim como, à estrutura sintática.

A métrica *content_word_diversity,* que mede a variedade de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios), trouxe-nos uma variação de **4, 37%**, sugerindo pouca variação entre as redações que utilizaram um vocabulário mais rico e diversificado, enquanto, outras, repetiram frequentemente os mesmos termos. Essa premissa foi reforçada pela variação da métrica *Type-Token Ratio (TTR),* que

atingiu **8, 18%**, revelando diferenças sutis na relação entre o número de palavras únicas e o número total de palavras nos textos.

A métrica *pronoun_diversity* que avalia a diversidade de pronomes utilizados no texto variou **18, 41%,** uma variação não tão expressiva, mas que mostra nos diferentes graus de sofisticação coesiva nos textos, uma maior diversidade e domínio dos recursos referenciais, favorecendo a coesão textual.

A métrica *content_word_min* representa o menor número de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios) registrado em um texto. A variação foi de **10,73%,** indicando que poucos textos utilizaram esse tipo de vocabulário com mais economia, possivelmente com foco em estruturas sintáticas mais funcionais. Textos com maior número de palavras de conteúdo tendem a apresentar maior densidade semântica.

A métrica *verb_diversity* avalia a diversidade de verbos, calculada pela razão entre o número de verbos distintos e o total de de verbos. Um valor de **9, 22%,** como o que foi gerado, pode indicar uma baixa variação na diversidade verbal no texto.

Ademais, a métrica *content_word_max,* que mede a frequência da palavra de conteúdo mais recorrente no texto, variou **8, 57%**, isso nos leva a entender que poucas redações repetiram certos termos com mais frequência, ao passo que as demais, distribuíram melhor o uso de palavras de conteúdo.

Outro aspecto importante revelado pelas métricas, refere-se à análise da diversidade de classes gramaticais específicas. Nesses aspectos, a diversidade de substantivos *(noun_diversity)* variou **8, 11%**, enquanto a razão de diversidade de adjetivos *(adjective_diversity_ratio*) apresentou uma variação de **11, 38%**. A análise que podemos fazer é a de que a maior parte dos textos são mais descritivos e detalhados do que outros.

Foi verificada ainda, a métrica *preposition_diversity, que* avalia a diversidade de preposições utilizadas em um texto. No *corpus*, a variação foi de **13, 16%,** o que indica que a maior parte das redações utilizou uma variedade maior de preposições, refletindo, possivelmente, em uma construção sintática mais complexa, com mais relações temporais e lógicas.

5.2.6 Conectivos

Por conseguinte, o grupo de “Conectivos”, apresentou, nesta análise de variação, a métrica, a *cau_pos_conn_ratio.* A categoria avalia o uso de conectivos e de operadores lógicos (positivos e negativos) no encadeamento das ideias do texto. A métrica, por sua vez, mediu os conectores causais, como "porque", "portanto" e "assim", que são fundamentais para estabelecer as relações de causa e efeito, decisivos na coerência e a clareza argumentativa do texto. A variação observada no *corpus*, foi de **22, 70%**. Tal variação indicou que há diferenças moderadas entre as redações quanto ao uso de conectores causais positivos. Em suma, alguns textos parecem ter utilizado esses conectores de maneira mais frequente, reforçando a explicitação de relações causais, outros textos apresentaram as ideias de argumentação mais implícitas nas relações de causa e efeito.

Outra métrica foi a *Add_pos_conn_ratio, que* é uma métrica que mensura a razão de conectores positivos (ex.: "além disso", "por outro lado"). O valor de variação foi de **27, 83%,** indicando que parte moderada das redações fez uso desses conectores.

*And_ratio***:** nessa métrica, a variação foi de **44,37%.** A métrica *and_ratio* indica a frequência relativa do uso da conjunção "e". A alta variação que obtivemos, sugere que há diferenças significativas nas escolhas coesivas dos autores: enquanto alguns utilizaram amplamente essa conjunção aditiva (e), outros recorreram a estratégias mais variadas de coesão (como advérbios, conectivos lógicos ou substituições pronominais).

*Verbs***:** essa métrica, que mede o número total de verbos no texto, teve variação de **18,51%.** A variação sugere diferentes níveis de dinamismo textual: textos com mais verbos tendem a ser mais processuais e dinâmicos; textos com menos verbos podem se apoiar mais em construções nominais e abstratas.

*Logic_operators***:** essa métrica mede a presença de operadores lógicos (como "se", "então", "mas", "porém"), fundamentais para a estrutura argumentativa. A variação observada foi de **35,13%**, indicando que acentuada parte dos textos parece ter empregado com mais frequência esses marcadores, enquanto outros se valeram de estruturas mais implícitas.

5.2.7 Complexidade Sintática

As métricas a serem analisadas na sequência, dizem respeito à categoria da “complexidade sintática”. Importa mencionar que a categoria em questão, verifica diferentes formas de alinhamento sintático e de complexidade estrutural, como frases regulares e irregulares.

A métrica *yngve*[35]*,* mede a complexidade sintática a partir do índice de Yngve; avalia o quão complexa é a estrutura das sentenças. Um valor de **15, 05%** como o que foi obtido, indica uma variação moderada quanto ao nível de complexidade da sentença, sugerindo que a maior parte das frases dos textos são bem estruturadas, sem excessiva subordinação.

A métrica *std_noun_phrase* teve um valor de variação de **26, 66%.** Essa métrica está relacionada ao desvio-padrão da quantidade de sintagmas nominais. O valor obtido nos sugere que nas redações há certa oscilação na estrutura frasal, com alguns textos possuindo partes mais carregadas de substantivos e outros menos.

Já a métrica *dep_distance* mede a distância de dependência sintática; ou seja, mede a complexidade estrutural das sentenças com base na distância entre os elementos dependentes de uma frase. Um valor de **27, 26%**, que foi a variação obtida no *corpus*, indica que uma parte moderada das redações apresenta frases mais longas e complexas; podemos inferir, assim, que houve um determinado nível de elaboração sintática, característica típica de textos bem construídos, no aspecto argumentativo.

5.2.8 Frequência de Palavras

No grupo de “Frequência de Palavras”, a análise revelou variações consideráveis em diferentes métricas, principalmente, naquelas relacionadas à distribuição de palavras entre os textos.

A métrica *cw_freq_brwac*, que mede a frequência média das palavras de conteúdo no *corpus* BrWaC, revelou um valor de variação de **3, 22%,** indicando que a utilização de palavras de conteúdo nos textos é equilibrada, ou seja, os textos se fundamentaram mais em palavras de conteúdo do que em palavras simples.

---

[35] A complexidade de Yngve baseia-se na premissa de que as árvores sintáticas das sentenças da língua inglesa tendem a se ramificar para a direita, e que desvios em relação a esse padrão correspondem a uma maior complexidade na linguagem.

A métrica *cw_freq* teve uma variação de **36,50%**. Essa métrica calcula a frequência média das palavras de conteúdonos textos. A alta variação sugere que os textos utilizaram essas palavras em densidades distintas, refletindo diferenças na carga informativa. É útil relembrar que, textos com maior frequência de palavras de conteúdo tendem a apresentar maior densidade argumentativa, aspecto valorizado pela competência três.

A métrica *freq_brwac,* que mede a frequência média das palavras no *corpus* BrWaC, apresentou uma variação de **1, 93%** gerando indícios de que a maioria dos textos continha palavras mais comuns no português brasileiro, enquanto, pouquíssimos, empregaram termos não tão frequentes. Uma variação baixa.

Já a *freq_bra,* que compara a frequência média das palavras utilizadas nas redações com a de um *corpus* brasileiro de referência[36], apresentou uma variação de **7, 68%,** indicando diferenças sutis na variação das escolhas lexicais entre os textos do *corpus* e os padrões do português do Brasil.

A análise das frequências mínimas também revelou variações relevantes. A métrica *min_cw_freq_bra,* que avalia a menor frequência de palavras de conteúdo no *corpus* brasileiro, variou **20, 99%**; algumas redações utilizaram palavras de conteúdo que são raras nesse *corpus*, enquanto outras empregam termos mais comuns.

Da mesma forma, a *min_freq_brwac*[37]*,* que mede a menor frequência de palavras no *corpus* BrWaC, teve uma variação de **23, 22%**, o que indica moderada parte das redações continha palavras que são extremamente raras ou quase inexistentes no BrWaC, possivelmente relacionadas ao uso de neologismos, de termos técnicos ou de regionalismos.

Já a *min_freq_bra,* que mede a menor frequência de palavras no *corpus* brasileiro, variou **24, 50%**, reforçando a ideia de que algumas redações utilizaram termos pouco comuns, outras, permaneceram dentro de um vocabulário mais recorrente.

---

[36] O Corpus Brasileiro (http://corpusbrasileiro.pucsp.br/cb/Inicial.html e https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS) é uma coletânea de aproximadamente um bilhão de palavras de português brasileiro, resultado de projeto coordenado por Tony Berber Sardinha, (GELC, LAEL, Cepril, PUCSP), com financiamento da Fapesp.

[37] O Corpus BrWac (https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC) foi disponibilizado em Janeiro de 2017 e é composto por 3.53 milhões de documentos da Web, 2.68 bilhões de tokens e 5.79 milhão de types (TTR 0.0021). Os textos do BrWac foram etiquetados pelo tagger nlpnet, que gera uma etiquetação de PoS (part-of-speech).

Ademais, a métrica *min_cw_freq_brwac*[38]*,* que foca a menor frequência de palavras de conteúdo específicas no *corpus* BrWaC, apresentou uma variação de **23, 07%**, o que significa que um quantitativo moderado de textos contém palavras de conteúdo que são muito raras nesse *corpus* de referência, indicando o uso de um vocabulário mais especializado ou menos convencional em certos textos.

5.2.9 Informações Semânticas de Palavras

A categoria "Informações Semânticas de Palavras", analisa características como sinonímia, antonímia, hiperonímia, hiponímia, polissemia, homonímia, paronímia, metáforas e metonímias, a fim de fornecer uma compreensão mais profunda do significado das palavras e de suas inter-relações dentro do texto. Nessa categoria, as métricas foram:

- A métrica *adjectives_ambiguity* teve um valor de variação de **28, 87%.** Essa métrica mede a ambiguidade dos adjetivos, o grau de ambiguidade dos adjetivos utilizados. Um valor como o obtido, sugere que boa parte dos textos utilizou adjetivos que podem ter múltiplos significados a depender do contexto.
- A métrica *nouns_ambiguity* teve um valor de **13, 27%.** Trata-se de uma métrica que mensura a ambiguidade dos substantivos, o grau de polissemia dos substantivos utilizados. O valor obtido sugere que há um certo nível de ambiguidade nos substantivos e de palavras com múltiplos significados, em uma parte sutil dos textos.
- adjectives_ambiguity: a variação nessa métrica foi de **28,87**%. É uma métrica que mede a ambiguidade semântica dos adjetivos. Textos com valores mais altos utilizam adjetivos mais genéricos ou polissêmicos; já os de menor ambiguidade tendem a ser mais precisos. Nos textos, a variação aparece como moderada no que condiz com a utilização dos dois grupos de adjetivos.
- *verbs_ambiguity:* essa métrica avalia o grau de ambiguidade presente nos verbos utilizados no texto. A variação foi de **17, 65%**, o que indica que algumas redações, não muitas, apresentaram verbos com múltiplos significados ou interpretações, podendo afetar a clareza da mensagem a ser transmitida.

38

- *adverbs_ambiguity:* de modo similar à métrica acima, essa métrica avalia a ambiguidade nos advérbios. A variação foi de **40, 85%,** sugerindo que um quantitativo significativo dos textos, utilizou advérbios que podem ser interpretados de maneiras diferentes, introduzindo de modo potencial, ambiguidades na modificação de verbos, adjetivos ou outros advérbios.
- *adjectives_max:* essa métrica quantifica o número máximo de adjetivos em uma sentença. A variação foi de **28, 01%** e aponta diferentes estratégias de valorização ou de qualificação de conceitos. O valor moderado de variação sugere que o *corpus* possui textos mais densos, que podem soar mais opinativos ou subjetivos, o que, se bem fundamentado, fortalece a argumentação.
- *abstract_nouns_ratio: essa* métrica avalia a proporção de substantivos abstratos (conceitos intangíveis) em relação ao total de substantivos no texto. A variação foi de **38, 40%**, sugerindo que uma representativa parte das redações fez uso mais frequente de conceitos abstratos.
- *content_words_ambiguity: essa* métrica avalia a ambiguidade geral das palavras de conteúdo (substantivos, verbos, adjetivos e advérbios) no texto. A variação foi de **20, 59%,** indicando moderada variação nas disparidades na clareza e na precisão vocabular entre as redações.
- *negative_words:* essa métrica contabiliza o número de palavras com conotação negativa no texto. A variação foi de **23, 77%,** moderada, indicando que algumas redações contêm mais termos negativos, o que pode influenciar a percepção textual
- *positive_words*: diferentemente da métrica anterior, essa contabiliza palavras com conotação positiva. Uma variação de **20, 02%**, como a verificada, sugere moderadas diferenças na utilização de termos positivos entre as redações, afetando o tom geral dos textos.

5.2.10 Índices de Leiturabilidade

A ferramenta também nos traz a categoria de “Índices de Leiturabilidade”. Ela inclui índices de legibilidade altamente reconhecidos no PLN, a saber: Fórmula de

Chall[39], Índice de Brunet[40], Flesch[41], Gunning Fox[42] e Estatística de Honoré[43]. Para essa categoria, as métricas foram as seguintes:

- *Honore: a* Estatística de Honoré é uma medida de riqueza lexical que considera a proporção de palavras que aparecem apenas uma vez no texto. Valores mais altos indicam um vocabulário mais diversificado. A variação encontrada no *corpus* foi de **18, 89%;** temos, então, diferenças moderadas na riqueza lexical entre as redações.
- *dalechall_adapted***:** o índice de legibilidade adaptado de Dale-Chall, combina frequência de palavras com comprimento médio de frases. A variação observada foi de **9,45%,** sugerindo que pequena parte dos textos oscilou nos níveis de acessibilidade, que configura textos diretos e fáceis de ler ou densos e complexos.
- *Gunning_fox* é uma métrica associada à dificuldade de leitura de um determinado texto. Para essa métrica, o *corpus* revelou uma variação de **19, 28%,** o que pode indicar um alto nível de elaboração textual, com frases longas e com palavras mais elaboradas. Se esse valor variar para cima, temos a indicação de um aumento da complexidade textual, o que pode estar associado ao refinamento na escrita dos candidatos, já que atingiram nota máxima.
- *dalechall_adapted:* essa métrica teve uma variação de **9, 45%** no índice Dale-Chall adaptado, o que indica uma diferença atenuada na legibilidade, entre as redações do *corpus*. Valores mais altos, nesse índice, sugerem textos com maior complexidade; a variação observada nas redações foi pequena, indicando uma uniformidade entre os textos.

5.2.11 Informações Morfossintáticas de Palavras

---

[39] A fórmula de leiturabilidade de Dalechall adaptada combina a quantidade de palavras não familiares com a quantidade média de palavras por sentença, seguindo os respectivos níveis escolares.
[40] Estatística de Brunet é uma forma de type/token ratio menos sensível ao tamanho do texto. Eleva-se o número de types à constante -0,165 e depois eleva-se o número de tokens a esse resultado.
[41] Reveja a figura 34 deste trabalho.;
[42] o índice de leiturabilidade Gunning Fog (também conhecido como Gunning FoX) soma a quantidade média de palavras por sentença ao percentual de palavras difíceis no texto e multiplica tudo por 4. O resultado está diretamente ligado aos 12 níveis do ensino americano. Índices superiores a 12 representam textos extremamente complexos.
[43] A estatística de Honoré é um tipo de type/token ratio que leva em consideração, além da quantidade de types e tokens, a quantidade de hapax legomena. Um hapax legomena é um type que só apresenta um token no texto.

O penúltimo grupo a ser considerado nesta análise é o de “Informações Morfossintáticas de Palavras”. Esse grupo analisa o uso de substantivos abstratos, palavras polissêmicas, palavras afetivas e polaridade (positiva e negativa) das palavras. Vejamos as métricas:

A métrica *function_words,* que contabiliza o total de palavras funcionais (artigos, preposições, conjunções e pronomes), apresentou uma variação de **34, 27%.** Algumas redações pareceram conter esses elementos com muito mais frequência do que outras, refletindo a variedade da construção sintática e interferindo na fluidez textual, já que, as palavras funcionais são essenciais para a organização estrutural das sentenças.

A métrica *pronoun_ratio*, cujo valor de variação foi de **23, 84%,** avalia a proporção de pronomes no texto; ela mede a frequência de pronomes em relação ao total de palavras. O valor nos revelou que o uso de pronomes no lugar de substantivos e de expressões nominais, possui uma variação moderada nos textos.

A métrica *punctuation_ratio* corresponde à razão de sinais de pontuação em relação ao número total de palavras. A atenuada variação de **16,37%** mostra que não houve tantas estratégias dferentes de marcação sintática e de ritmo textual.

Em continuidade, a métrica *nouns_min*, que mensura o menor número de substantivos identificados em textos do corpus, teve uma variação de **21,71%**. A variação sugere que, moderadamente, em alguns textos, predomina a ação verbal ou estruturas menos nominais, enquanto outros concentram mais conceitos por meio de substantivos.

A métrica *pronouns_standard_deviation* teve uma variação de **26,29%** . Essa métrica calcula o desvio padrão da distribuição de pronomes ao longo do texto, o que permite avaliar o grau de regularidade ou oscilação no uso de elementos referenciais. A variação de 26,29% entre os textos indica, moderadamente, estilos distintos de retomada de informações e de coesão referencial. Isso porque textos com menor desvio padrão tendem a utilizar os pronomes de forma mais sistemática, o que pode contribuir para a fluidez e clareza da progressão temática.

A métrica *nouns_standard_deviation* teve uma variação de **17,98%.** Essa métrica mede o desvio padrão da frequência de substantivos ao longo do texto. A variação de 17,98% sugere que os textos nota mil variraam em sua distribuição temática: alguns mantiveram o uso de substantivos de forma mais equilibrada entre

os parágrafos, enquanto outros concentraram mais termos conceituais em trechos específicos.

Em continuidade, a métrica *adjective_ratio*, que mede a proporção de adjetivos em relação ao total de palavras, apresentou uma variação moderada de **21, 54%,** sugerindo que algumas redações fizeram um uso muito mais intenso de descrições qualificativas do que outras. Enquanto alguns textos priorizaram descrições detalhadas, outros foram mais diretos e objetivos.

No que se refere ao equilíbrio entre palavras funcionais e palavras de conteúdo, a métrica *ratio_function_to_content_words* variou **7, 63%,** indicando que a maior parte das redações apresentou uma densidade de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios), sendo mais informativas.

Finalmente, neste grupo, a métrica *adverbs_diversity_ratio*, que avalia a diversidade de advérbios, apresentou uma variação de **18, 32%**, demonstrando que alguns textos utilizaram uma gama mais ampla desses elementos para que a argumentação fosse enriquecida e para que o discurso se tornasse mais expressivo.

5.2.12 Densidade de Padrões Sintáticos

No grupo de “Densidade de Padrões Sintáticos”, que mede a ocorrência de padrões sintáticos frequentes na língua, como sintagmas nominais, adjuntos e estruturas de dependência, a análise obteve as seguintes métricas:

- A métrica *min_noun_phrase* que avalia o número mínimo de sintagmas nominais por sentença. O valor foi de **8, 42%,** mostrando que uma representativa parte dos textos contém pelo menos um sintagma nominal; isto é, a estrutura das frases parece ser gramaticalmente equilibrada, com ao menos um núcleo substantivo presente.
- A métrica *max_noun_phrase* quantifica o comprimento máximo de sintagma nominal. O valor de **32, 46%** é alto, sugere relevante variação na presença de uma construção nominal muito longa; isso, pode afetar a clareza e a legibilidade do texto.
- A métrica *mean_noun_phrase* que se refere à média de frases nominais por sentença obteve uma variação de **20,84%,** que sugere que parte considerável dos textos usou frases nominais mais elaboradas, o que contribui para a densidade conceitual do argumento.

De modo suscinto, em reforço ao que foi exposto, elucido que após as considerações feitas sobre as análises das métricas e de como elas aparecem nos textos, o *corpus* parece revelar algumas características típicas e similares. As variações observadas nas redações indicaram certa heterogeneidade no que tange aos termos de riqueza lexical, de densidade de conteúdo, do comprimento dos textos, do uso de palavras raras e da complexidade de leitura. Outrossim, o *corpus* indica a predominância de redações extensas e desenvolvidas, com palavras de conteúdo bastante frequentes, mas sem um vocabulário excessivamente raro.

Além disso, diferenças notáveis também puderam ser percebidas em termos de complexidade sintática e de escolha lexical. Distinções, por exemplo, no uso de conjunções coordenativas e subordinativas, na frequência de sentenças longas e na proporção de palavras simples, indicaram textos com níveis de complexidade e de formalidade um pouco dicotômicos.

No que concerne às questões de concretude, de familiaridade e de imageabilidade das palavras utilizadas, as discrepâncias pareceram refletir certa diversidade na acessibilidade dos textos. De qualquer modo, a legibilidade geral dos textos pareceu ser consistente. A complexidade argumentativa por sua vez e a variação das métricas relacionadas a ela, nos indicou medidas de preferência por estruturas mais explícitas ou implícitas no delineamento de relações causais dentro do texto. Os níveis de carga semântica e de termos específicos nos textos também trouxeram essa percepção de consistência.

Outro aspecto relevante esteve relacionado à densidade informacional. Os textos analisados, a partir da variação da obtida, apresentaram um alto índice de palavras de conteúdo por sentença, apontando para uma construção densa e substantiva, que foi percebida pela caracterização e pela predominância de substantivos e de expressões nominais em detrimento de estruturas mais verbais.

Por fim, a estrutura organizacional das redações apresenta-se planejada e com segmentação lógica, já que, o número de parágrafos e a média de sentenças por parágrafo indicaram a possibilidade de uma progressão textual bem definida, assim como o esperado. Embora os parágrafos sejam longos, a organização interna parece não ter comprometido a legibilidade e a fluidez argumentativa.

Dessa forma, os achados desta análise, a partir dos cálculos de variação e dos parâmetros já estabelecidos pela ferramenta, corroboraram a premissa de que as redações nota mil do Enem (2014-2023) se destacaram por um conjunto de escolhas

linguístico-textuais consistentes que favoreceram a coesão, a coerência, a proficiência e a escrita formal; aspectos tipicamente exigidos nas competências propostas no exame.

5.3 ALINHAMENTO DAS MÉTRICAS OBTIDAS COM AS COMPETÊNCIAS AVALIATIVAS DO ENEM

O Exame Nacional do Ensino Médio (Enem), conforme abordado anteriormente, adota um quadro de competências que se desdobram em critérios avaliativos dos textos (redações) escritos no Exame; isso, em consonância com o que é estabelecido pela Matriz de Correção (*vide* subseção 2.1.1). Esta pesquisa trouxe um panorama daquilo que as competências contemplam, desde o domínio da norma culta e da adequação ao gênero dissertativo-argumentativo até a organização argumentativa e a elaboração de uma proposta de intervenção (*vide* figura 3).

O intuito desta subseção é responder à terceira pergunta desta pesquisa, que busca saber se as métricas do NILC-Metrix nas redações nota mil se alinham às competências avaliativas do Enem (*vide* subseção 1.4.1). A hipótese norteadora desta pesquisa, como mencionada na subseção 1.4.2), sustenta que as redações analisadas apresentam pistas confiáveis sobre as características linguístico-textuais que distinguem produções de alto desempenho no exame. Logo, ao relacionar as métricas do NILC-Metrix com os critérios de cada competência, pude comprovar que existem padrões linguísticos predominantes nessas produções e identificar elementos que corroboram a atribuição de pontuações elevadas.

Esta pesquisa por estar inserida na Linguística de Corpus, considera o texto como uma produção linguística concreta; enquanto pesquisadora, mantenho o foco na materialidade linguística a fim de demonstrar que os padrões extraídos estatisticamente são observáveis no nível microtextual, o que fortalece o caráter aplicado da investigação. Ao descrever por meio de métricas, o funcionamento interno das redações nota mil, busquei reforçar a contribuição deste estudo para o ensino da escrita com base em evidências empíricas, acessíveis e aplicáveis à sala de aula. Para tanto, trarei exemplos de textos do próprio *corpus* da pesquisa, digitados, digitalizados e analisados, com exemplos explícitos e explicações linguísticas ancoradas nas competências do Enem, que também configuram o objeto de estudo desta pesquisa.

A seguir, à luz das comprovações, reconsidero cada uma das cinco competências do Enem, em conformidade com as métricas computacionais extraídas do NILC-Metrix, para uma compreensão última de como essas características se manifestaram nos textos de nota máxima. Antes, urge ressaltar que as minuciosidades contidas nas competências do Enem foram devidamente abordadas e explicadas na subseção 2.1.1.

5.3.1 Competência I – Domínio da norma culta da Língua Portuguesa

Como já sabemos, a Competência I (*vide* figura 5) avalia a conformidade da redação à norma culta da língua portuguesa; o que engloba aspectos ortográficos, gramaticais e sintáticos. As métricas analisadas revelaram sofisticação sintática e lexical nas redações nota mil. A título de exemplificação, o valor reduzido de desvio-padrão da idade de aquisição das palavras (idade_aquisicao_std) sugeriu que os candidatos utilizaram um vocabulário homogêneo e adequado ao contexto formal. Além disso, a riqueza lexical evidenciada pela alta diversidade de verbos e pronomes *(verbs), (pronouns)* demonstrou que os textos possuem um refinamento da linguagem, que está associado à precisão vocabular presente nos textos.

Ainda no que tange à competência I, outro fator relevante foi a complexidade estrutural das frases, refletida pelo alto número médio de palavras por sentença *(words_per_sentence)* e de sílabas por palavra de conteúdo *(syllables_per_content_word).* A análise dos valores evidenciou a existência de períodos bem estruturados e complexos. Ao mesmo tempo, a frequência elevada de palavras de conteúdo *(content_words)* nos comprovou que os textos apresentam uma densidade informacional consistente, com construções vazias ou redundantes quase inexistentes.

Ademais, a análise evidenciou o uso frequente de estruturas sintáticas formais, período composto, colocação pronominal adequada e vocabulário compatível com a norma padrão. Um exemplo típico é o trecho: *“Ao longo do processo de formação da sociedade, o pensamento cinematográfico consolidou-se em diversas comunidades. No início do século XX, com os regimes totalitários, por exemplo, o cinema era utilizado como meio de dominação à adesão das massas ao governo.”* (Redação 128, 2019). A escolha lexical de alta abstração (como “dominação”, “pensamento

cinematográfico") e o uso de construções impessoais demonstram o domínio da norma culta exigido pela competência I.

O exemplo de redação, da figura 47, foi retirado do *corpus* da pesquisa (redação 8, 2014) e ilustra como algumas das características supramencionadas podem ser entendidas no texto. Trarei ilustrações para todas as competências.

Figura 51 - Ilustração da análise no texto 8 (Competência I)

A publicidade infantil movimenta bilhões de dólares e é responsável por considerável aumento no número de vendas de produtos e serviços direcionados às crianças. No Brasil, o debate sobre a publicidade infantil representa uma questão que envolve interesses diversos. Nesse contexto, o governo deve regulamentar a veiculação e o conteúdo de campanhas publicitárias voltadas às crianças, pois, do contrário, elas podem ser prejudicadas em sua formação, com prejuízos físicos, psicológicos e emocionais. Em primeiro lugar, nota-se que as propagandas voltadas ao público mais jovem podem influir nos hábitos alimentares, podendo alterar, consequentemente, o desenvolvimento físico e a saúde das crianças. Os brindes que acompanham as refeições infantis ofertados pelas grandes redes de lanchonetes, por exemplo, aumentam o consumo de alimentos muito calóricos e prejudiciais à saúde pelas crianças, interessadas nos prêmios. Esse aumento da ingestão de alimentos pouco saudáveis pode acarretar o surgimento precoce de doenças como a obesidade. Em segundo lugar, observa-se que a publicidade infantil é um estímulo ao consumismo desde a mais tenra idade. O consumo de brinquedos e aparelhos eletrônicos modifica os hábitos comportamentais de muitas crianças que, para conseguir acompanhar as novas brincadeiras dos colegas, pedem presentes cada vez mais caros aos pais. Quando esses não podem comprá-los, as crianças podem ser vítimas de piadas maldosas por parte dos outros, podendo também ser excluídas de determinados círculos de amizade, o que prejudica o desenvolvimento emocional e psicológico dela. Em decorrência disso, cabe ao Governo Federal e ao terceiro setor a tarefa de reverter esse quadro. O terceiro setor – composto por associações que buscam se organizar para conseguir melhorias na sociedade – deve conscientizar, por meio de palestras e grupos de discussão, os pais e os familiares das crianças para que discutam com elas a respeito do consumismo e dos males disso. Por fim, o Estado deve regular os conteúdos veiculados nas campanhas publicitárias, para que essas não tentem convencer pessoas que ainda não têm o senso crítico desenvolvido. Além disso, ele deve multar as empresas publicitárias que não respeitarem suas determinações. Com esses atos, a publicidade infantil deixará de ser tão prejudicial e as crianças brasileiras poderão crescer e se desenvolver de forma mais saudável.

Legenda:

Sofisticação Sintática e Lexical	Vocabulário formal	Densidade de verbos e de alguns pronomes
Complexidade estrutural das frases	Densidade informacional consistentes	

Fonte: organizado pela autora

5.3.2 Competência II – Compreensão do tema, adequação ao tipo textual e o uso do repertório cultural

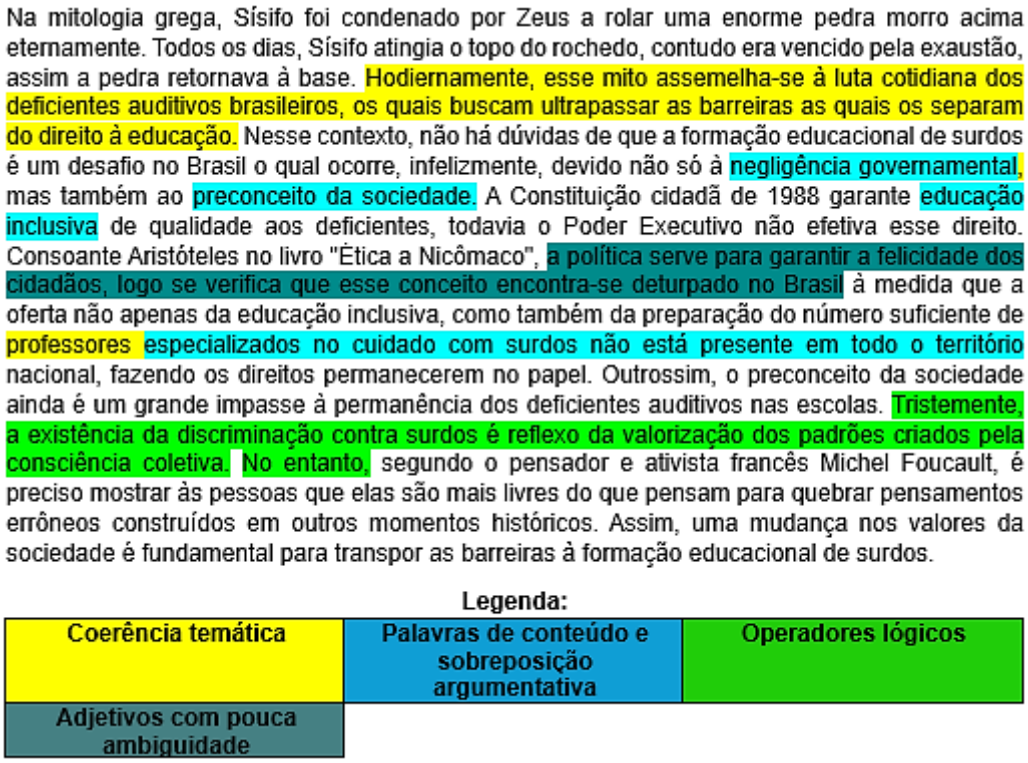
O trecho a seguir, também retirado de uma das redações do *corpus,* mostra como a compreensão do tema é explicitada: *"O Brasil cresceu nas bases parternalistas da sociedade europeia, visto que as mulheres eram excluídas das decisões políticas e sociais, inclusive do voto. Diante desse fato, elas sempre foram*

tratadas como cidadãs inferiores cuja vontade tem menor validade que as demais." (Redação 23, 2015). O candidato contextualiza historicamente o problema e desenvolve sua argumentação de forma crítica e alinhada à proposta temática. A progressão do raciocínio e o uso de vocabulário temático específico são indícios de compreensão e domínio do gênero exigido.

A competência II, como dito, avalia a capacidade do candidato de interpretar corretamente a proposta temática, mobilizando um repertório sociocultural produtivo para sustentar a sua argumentação. A variação das métricas extraídas revelou que as redações analisadas demonstraram coerência temática e uma abordagem argumentativa bem embasada; isso também pôde ser comprovado pela frequência elevada de palavras de conteúdo e pela sobreposição argumentativa *(arg_ovl),* que sugerem uma progressão lógica das ideias ao longo do texto em conformidade com a temática proposta e com os possíveis repertórios escolhidos.

Além disso, o baixo índice de operadores lógicos explícitos *(logic_operators)* sugeriu que os candidatos estruturaram seus argumentos com fluidez, sem recorrência de marcadores discursivos tão excessivos. Outro ponto, atrelado ao que exige a competência II, é a presença moderada de adjetivos ambíguos *(adjectives_ambiguity*) para evitar generalizações e imprecisões.

Figura 52 - Ilustração da análise no texto 48 (Competência II)



Fonte: organizado pela autora

5.3.3 Competência III – Seleção, organização e interpretação de informações para sustentar um ponto de vista

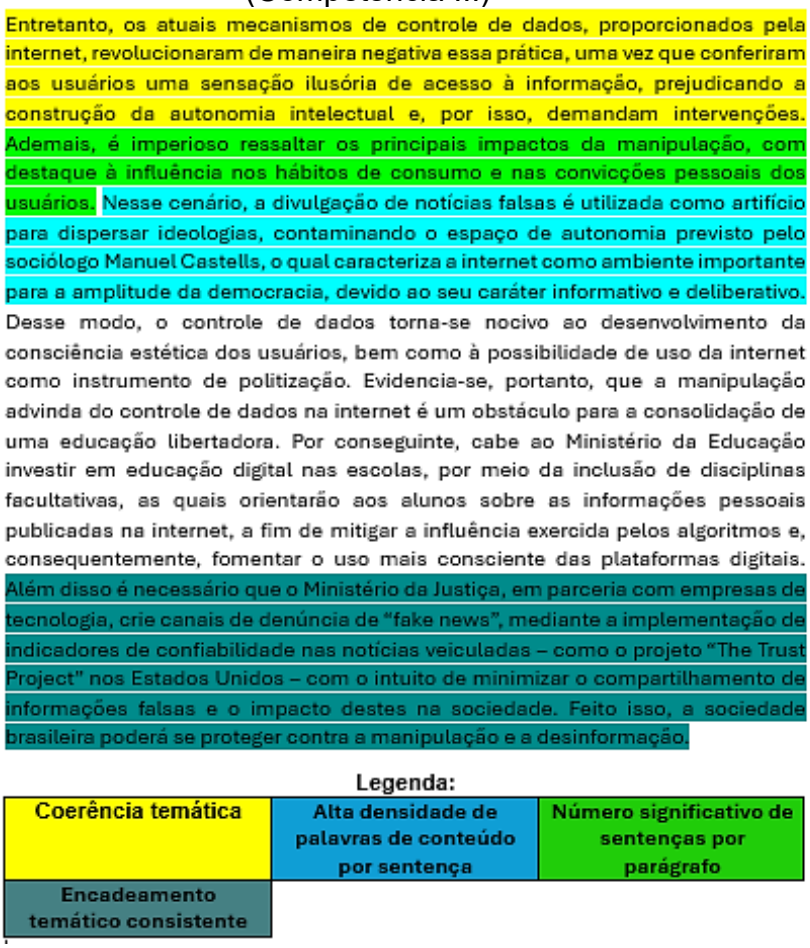
Em continuidade, a progressão argumentativa aparece em construções como: *“Infere-se, portanto, que medidas são necessárias para combater o machismo e facilitar o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil.”* (Redação 259, 2023); é possível perceber o uso de conectores argumentativos e como eles orientam a organização textual e sustentam a coerência interna da redação.

A competência III está relacionada à coerência textual e à capacidade do candidato de organizar sua argumentação de maneira progressiva e lógica. A análise revelou que as redações nota mil apresentaram um alto nível de planejamento textual, evidenciado pelo elevado índice de distância sintática *(dep_distance),* que indica um uso eficiente de orações subordinadas para articular os argumentos.

A estruturação da informação foi reforçada pela alta densidade de palavras de conteúdo por sentença *(content_density)* e pelo número significativo de sentenças

por parágrafo *(sentences_per_paragraph)*, nos indicando que os candidatos evitaram construções fragmentadas. A sobreposição lexical e argumentativa também foi um indicador relevante, sugerindo o uso de um encadeamento temático consistente, que está associado à inteligibilidade do texto e à defesa da argumentação.

Figura 53 - Ilustração da análise no texto (Redação 72, 2018) (Competência III)



Fonte: organizado pela autora

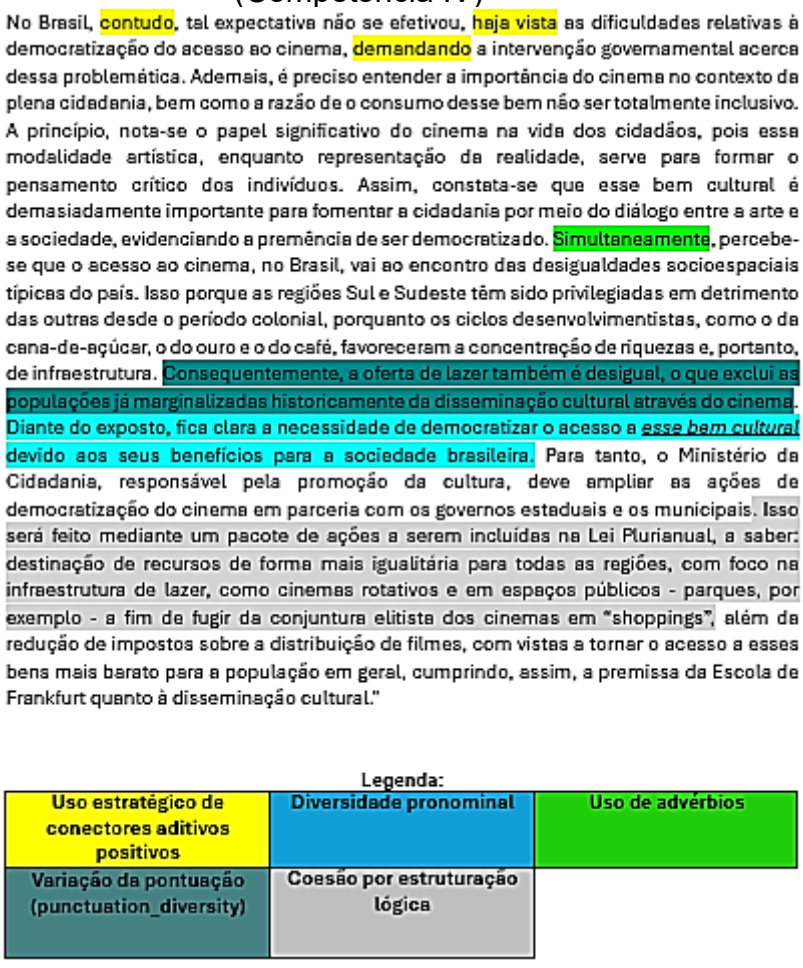
5.3.4 Competência IV – Uso de conectivos e de operadores argumentativos

Na competência IV, a coesão é a base, sendo avaliada pela forma como as ideias são conectadas ao longo do texto. As métricas indicaram um uso estratégico dos recursos coesivos, como demonstrado pelo índice de conectores aditivos positivos *(add_pos_conn_ratio).* A análise constatou que conectores aditivos, adversativos e conclusivos parecem ter sido amplamente utilizados para garantir a fluidez do texto. Observa-se o uso de articuladores como em: *"Portanto, haja vista que a educação é fundamental para o desenvolvimento econômico do referido público e,*

logo, da nação, ela deve ser efetivada aos surdos pelos agentes adequados, a partir da resolução dos entraves vinculados (Redação 43, 2017)". O equilíbrio entre as orações, o uso de pronomes anafóricos e as relações lógico-semânticas contribuem para a construção da coesão sequencial do texto.

A análise também revelou que as redações utilizam diversidade pronominal *(pronoun_diversity)* e uso moderado de advérbios *(adverbs),* aspectos relacionados à fluidez e a clareza textual. A variação da pontuação *(punctuation_diversity)* evidenciou um controle eficaz dos recursos formais, de tal forma que a coesão não dependeu apenas de conectores explícitos, mas também da organização estrutural dos períodos.

Figura 54 - Ilustração da análise no texto (Redação 147, 2019) (Competência IV)



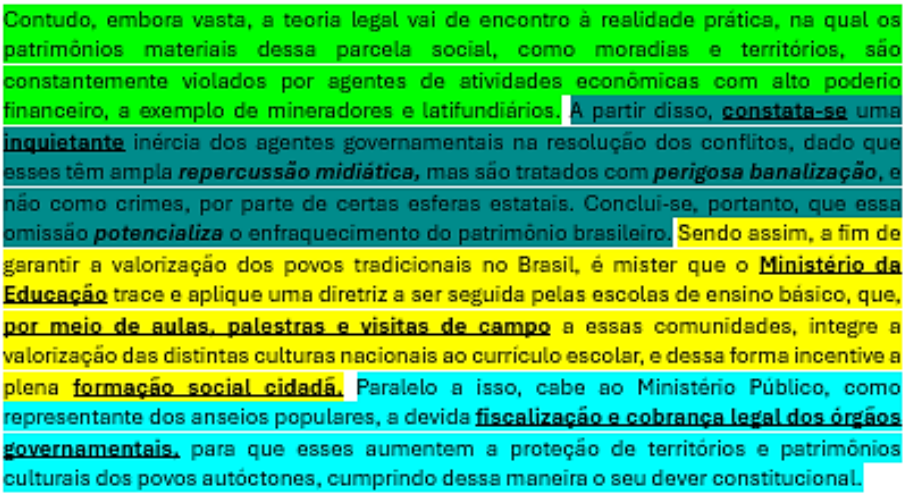
Fonte: organizado pela autora

5.3.5 Competência V – Elaboração da proposta de intervenção

A clareza da proposta de intervenção é ilustrada em: *“Nesse âmbito, cabe ao Ministério da Educação e da Cultura promover um maior acesso ao conhecimento e ao lazer, por meio da instalação de cinemas públicos nas áreas urbanas mais periféricas - que deverão possuir preços acessíveis à população local -, a fim de evitar a situação de alienação e insuficiência intelectual presente nos membros das classes mais baixas.”* (Redação 95, 2019). A intervenção apresentada contempla agente, ação, meio e intuito, aspectos exigidos pela matriz de correção. Além disso, observa-se a adequação aos direitos humanos e a viabilidade da proposta (Democratização do acesso ao cinema no Brasil). A Competência V requer dos candidatos, a apresentação de uma proposta de intervenção detalhada e articulada ao tema. A figura 10, deste trabalho, mostra os quesitos necessários à proposta de intervenção exigida.

Os dados obtidos com as métricas indicaram que as redações nota mil analisadas não fizeram o uso exacerbado de substantivos abstratos *(abstract_nouns_ratio),* de modo que as propostas pareçam ser tangíveis e aplicáveis. Além disso, o nível de complexidade sintática revelou que a formulação das propostas é detalhada e viável, isso, pela presença de palavras de alto nível de familiaridade *(familiaridade_4_55_ratio)* sugerindo que os candidatos empregaram um vocabulário compreensível para a exposição do texto como um todo, incluindo a proposta.

Figura 55 - Ilustração da análise no texto (Redação 230, 2022) (Competência V)



Contudo, embora vasta, a teoria legal vai de encontro à realidade prática, na qual os patrimônios materiais dessa parcela social, como moradias e territórios, são constantemente violados por agentes de atividades econômicas com alto poderio financeiro, a exemplo de mineradores e latifundiários. A partir disso, **constata-se** uma **inquietante** inércia dos agentes governamentais na resolução dos conflitos, dado que esses têm ampla ***repercussão midiática,*** mas são tratados com ***perigosa banalização***, e não como crimes, por parte de certas esferas estatais. Conclui-se, portanto, que essa omissão ***potencializa*** o enfraquecimento do patrimônio brasileiro. Sendo assim, a fim de garantir a valorização dos povos tradicionais no Brasil, é mister que o **Ministério da Educação** trace e aplique uma diretriz a ser seguida pelas escolas de ensino básico, que, **por meio de aulas, palestras e visitas de campo** a essas comunidades, integre a valorização das distintas culturas nacionais ao currículo escolar, e dessa forma incentive a plena **formação social cidadã.** Paralelo a isso, cabe ao Ministério Público, como representante dos anseios populares, a devida **fiscalização e cobrança legal dos órgãos governamentais,** para que esses aumentem a proteção de territórios e patrimônios culturais dos povos autóctones, cumprindo dessa maneira o seu dever constitucional.

Legenda:

Detalhamento da proposta	Poucos substantivos abstratos	Complexidade Sintática (Períodos compostos)
Vocabulário formal, mas compreensível.	Coesão por estruturação lógica	

Fonte: organizado pela autora

As exemplificações trazidas ocorreram em virtude da necessidade de uma compreensão mais concreta acerca de como os valores de variação obtidos das métricas podem ser entendidos, a partir de um olhar sobre o próprio texto. A tentativa foi a de mostrar de forma mais acurada, os desdobramentos palpáveis dos resultados desta pesquisa.

No encadeamento das discussões e das análises realizadas ao longo da pesquisa, bem como, na tessitura da escrita desta tese, partirei para o último capítulo, que retomará as perguntas de pesquisa, articulando-as com as considerações finais deste trabalho. Sequencialmente, apresentarei alguns encaminhamentos destinados aos pesquisadores/professores que almejem aprofundar os seus conhecimentos acerca da Linguística de *Corpus* e do processamento de linguagem natural, especificamente, no contexto do uso de métricas textuais para o redirecionamento do ensino, para a análise e para a correção de redações. Aos que desejarem trabalhar com ferramentas de análise e de descrição textuais, especialmente, com a NILC-

Metrix, acredito que encontrarão nesta pesquisa, muitas contribuições e pontos de partida.

6 CONSIDERAÇÕES FINAIS E ENCAMINHAMENTOS

> O mais importante num *corpus* é saber o que fazer com ele, como usá-lo, e para que tarefas ele é útil.
>
> Diana Santos

A totalidade desta tese não é a mesma coisa da soma de suas partes. Além de saber que existem outras finalidades para o *corpus* compilado para esta tese, acredito ser totalmente possível refletir acerca do ensino de escrita a partir da análise dos textos pela ótica da linguística de *corpus*, com auxílio de recursos de PLN.

Portanto, finalizado o trabalho descritivo e analítico ao qual me propus fazer, com questões e hipóteses que pude responder e verificar, penso que agora caibam alguns encaminhamentos destinados ao professor que ensina redação no seu fazer docente.

Sabemos que teorias são sempre uma tentativa de explicação a partir de dados, e dados são simultaneamente, a motivação e a exemplificação de uma teoria. Nesse sentido, com alteração nos dados, em quantidade e qualidade, é razoável que teorias sejam repensadas.

A partir disso, muito do que escrevo nesta tese pode contribuir para que práticas avaliativas de produção textual e práticas pedagógicas de ensino de escrita, sejam repensadas por meio de evidências e de aspectos totalmente palpáveis. Nesses 4 anos de doutoramento, eu as repensei. As minhas reflexões me levaram a um lugar: o de ensinar com intencionalidade para fins específicos.

A minha compreensão é de que, mesmo que de modo parcial, a reprodução de um processo avaliativo performado por um professor avaliador, com a ajuda de uma máquina que identifica desvios ortográficos, que quantifica palavras e que descobre padrões que não são explicitamente avaliados em algumas grades de avaliação utilizadas, não seja capaz de mitigar a nossa revisão constante da forma como estamos fazendo a avaliação de redações.

Entendo que o modo de ensinar, avaliar, corrigir e analisar os textos tem certa retroatividade nas salas de aula. Como vimos até aqui, por meio de métricas quantificáveis e de modelos estáveis de redação, o atalho necessário para quem deseja ingressar no ensino superior, é o saber escrever para atender a critérios e exigências avaliativas pré-estabelecidos nos grandes exames.

Não obstante pareça simples aprender a reproduzir modelos, advogo, por meio das análises realizadas até o momento e em virtude do meu fazer docente, que as estratégias necessárias para chegar a esses modelos, na medida que não são intencionalmente ensinadas e contextualizadas, são, por si só, complexas de serem aprendidas e, do mesmo modo, complexas de serem ensinadas. Aqui, endosso as palavras de Schlatter e Garcez (2012), “ensinar a escrever significa fomentar a participação das pessoas em eventos variados que exijam leitura e escrita”.

No caso da redação de vestibular/Enem, defendo que os professores das salas de ensino médio e dos cursinhos preparatórios, especialmente, os da área de linguagens, devam refletir, junto com os seus estudantes, sobre o que estão fazendo ao escrever uma redação, ao aprenderem macetes, ao repensarem modelos, ao definirem estratégias de escrita. Existem pistas confiáveis que nos permitem encontrar formas de ensino que tenham um conhecimento-alvo, uma boa prática definida e alunos pensantes.

Esta tese buscou, por meio de uma análise feita por uma ferramenta de PLN, compreender os padrões linguísticos presentes em um *corpus* de redações nota mil do Enem. Almejei, ademais, estabelecer as relações existentes entre esses padrões e as competências avaliativas do exame.

Por meio do percurso investigativo empreendido, pude verificar como as métricas extraídas pela NILC-Metrix podem contribuir com uma análise detalhada e verificável de textos. A partir da investigação das variações e das consistências nas métricas (proposta inicialmente traçada), emergiram descobertas que não somente nos possibilitam identificar elementos característicos das produções de alto desempenho, mas também abrem caminhos para o redirecionamento de um ensino que vá ao encontro dos modelos avaliativos que fazem parte da vida dos estudantes.

Os achados desta pesquisa revelaram que a padronização textual das redações, evidenciada pelo uso consistente de estruturas léxico-sintáticas, conectivas, semânticas e argumentativas, dentro de um espectro específico, estão fortemente alinhadas às competências exigidas pelo Enem, a partir de pistas confiáveis. No entanto, vale admitir que essa padronização não significa, necessariamente, um domínio pleno da escrita, por parte do aluno, em diferentes contextos acadêmicos e sociais; a não ser que o ensino de redação seja compreendido e conduzido para além de um meio de conformidade com um modelo

rigidamente estabelecido, contudo, como uma oportunidade de que a proficiência linguística seja intencionalmente alcançada por estratégias repensadas.

Ao propor um novo estudo sobre a análise de métricas textuais direcionadas ao ensino de redação, este estudo contribui com a literatura ao e*vide*nciar que é possível obter dados quantiqualitativos de um *corpus* e utilizá-los como evidências para o ensino de escrita de textos.

A partir das questões de pesquisa formuladas, foi possível constatar que existe variação entre os valores mínimos e máximos de algumas métricas nas redações analisadas, associadas a diferentes grupos linguístico-textuais, como os elencados na ferramenta NILC-Metrix. Tornou-se evidente que, na análise do *corpus*, algumas métricas demonstraram maior estabilidade do que outras. Em contraste, determinados elementos apresentaram maior variação.

A primeira questão proposta foi: *nas redações nota mil do Enem, existe variação entre os valores mínimos e máximos nas métricas fornecidas pelo NILC-Metrix?* Os cálculos de amplitude de variação demonstraram que, mesmo em textos com pontuação máxima, há significativa flutuação nos valores das métricas analisadas. As amplitudes foram discutidas com base nos cálculos estatísticos de variação já explicados no decorrer deste trabalho.

A segunda questão investigada foi: *quais métricas do NILC-Metrix mostraram maior consistência ou variação nas redações nota mil do Enem?* Por meio da análise da amplitude entre valores mínimos e máximos, observou-se que algumas métricas apresentaram estabilidade notável, enquanto outras variaram significativamente, chegando a alcançar números na ordem dos milhões. Essas variações extremas, inclusive, motivaram o descarte de certas métricas da análise interpretativa final, por não contribuírem com resultados confiáveis.

A terceira questão proposta foi: *as métricas do NILC-Metrix nas redações nota mil se alinham às competências avaliativas do Enem?* A análise das métricas demonstrou um alinhamento significativo com as competências da matriz avaliativa. Houve correspondência com as competências I (domínio da norma padrão), III (organização dos argumentos), IV (coesão textual) e V (clareza na proposta de intervenção). Ao observar os indicadores, pude reforçar a correlação mensurável que pode existir entre os padrões linguísticos computacionalmente extraídos e os critérios subjetivos de correção adotados pelos avaliadores humanos; os dados quantitativos refletem por meio de evidências, as exigências avaliativas do exame e, portanto,

podem ser utilizados como recurso para aprimoramento didático no ensino de redação.

Sob essa perspectiva, a abordagem adotada nesta pesquisa, ao articular métodos estatísticos de análise, técnicas de processamento de linguagem natural e a Linguística de *Corpus* com o ensino de escrita de redações, demonstra a complexidade da questão e amplia o escopo de análise na área de linguística aplicada.

Os resultados encontrados corroboram a hipótese inicial de que as redações nota mil oferecem pistas confiáveis sobre as características linguístico-textuais valorizadas na correção do Enem. Os achados também nos oferecem desdobramentos que podem ser utilizados para a compreensão de que a adesão de padrões de escrita não garante, por si só, a capacidade de transitar com desenvoltura por diferentes escritas acadêmicas. Isto é, ainda que o estudante de ensino médio precise considerar os requisitos de um exame, é essencial que o ensino de redação, para o processo avaliativo e para a vida, perpasse os limites que dificultam a escrita autônoma, intencional e reflexiva. Escrever é existir!

Embora os resultados desta pesquisa sejam tangíveis, é imprescindível reconhecer algumas limitações inerentes ao processo investigativo, especialmente no que tange ao quantitativo significativo de métricas dispostas na ferramenta NILC-Metrix. Quase nunca conseguimos responder a tudo o que um *corpus* pode propiciar; é exatamente por isso, que podemos acreditar que uma pesquisa nunca se encerra totalmente. A presente pesquisa, por exemplo, não se encerra em si mesma; pelo contrário, instiga novas investigações que possam, de algum modo, aprofundar a compreensão dos aspectos essenciais que foram analisados nesta tese.

Conclusivamente, espero que os achados desta pesquisa não apenas enriqueçam o debate acadêmico, mas também fundamentem reflexões e aplicações práticas em linguística aplicada, especialmente, no ensino de produção textual. Reforço a ideia de que o ensino de redação deve contemplar tanto os critérios avaliativos quantificados e analisados por métricas, nesta investigação, quanto a ampliação das habilidades discursivas que permitam que os alunos construam textos sobre os quais consigam refletir.

Somado a isso, reafirmo a relevância desta tese ao demonstrar que, no tocante ao ensino e avaliação da escrita de textos, ferramentas computacionais, como a NILC-Metrix podem auxiliar na análise sistemática de textos e na formulação de estratégias pedagógicas mais eficazes a partir de evidências. A combinação entre dados

quantitativos e a interpretação qualitativa me proporcionaram uma visão mais ampla, científica e acurada dos padrões linguísticos que caracterizam uma redação de alto desempenho.

Finalmente, como perspectivas futuras, incentivo um olhar contínuo sobre a investigação iniciada. Sugiro pesquisas semelhantes sobre a mesma temática, que contribuam com o aprofundamento científico e com a aplicabilidade social. Uma possibilidade promissora de continuidade desta pesquisa reside na investigação qualitativa e quantitativa das redações localizadas nas extremidades da curva de distribuição das métricas analisadas. Embora todas as produções do *corpus* tenham alcançado a nota máxima no Enem, observa-se, estatisticamente, a presença de textos com variações que se afastam da média — tanto por apresentarem valores mínimos quanto máximos em determinadas métricas. A análise desses casos poderia revelar características discursivas, estilísticas ou estruturais que diferenciam tais produções dos modelos mais recorrentes. Seria pertinente, por exemplo, investigar se essas redações representam formas mais criativas de argumentação, soluções discursivas menos convencionais ou ainda usos linguísticos atípicos, mas igualmente eficazes. Assim, investigações futuras poderiam se debruçar sobre esses textos "fora da curva", com o intuito de aprofundar a compreensão sobre a diversidade textual possível mesmo dentro de uma avaliação altamente normatizada.

Como encaminhamento futuro da pesquisa, e em alinhamento com os princípios de ciência aberta, pretendo disponibilizar publicamente o *corpus* utilizado neste estudo por meio da plataforma *Hugging Face*, amplamente reconhecida na comunidade de Processamento de Linguagem Natural. A sugestão foi apresentada durante a minha qualificação de tese, pelo professor Sidney Leal (avaliador da banca), um dos desenvolvedores do NILC-Metrix, e já está em andamento. A publicação será feita no formato de *dataset* aberto, com descrição técnica e metadados, permitindo que outros pesquisadores tenham acesso aos textos analisados. Ressalto que uma das principais contribuições desta pesquisa é justamente a organização sistematizada e curada do *corpus* especializado, que representa uma base textual profícua para futuros estudos sobre escrita, avaliação, métricas linguísticas e ensino de produção textual. Assim, a disponibilização do *corpus* reforça o compromisso da minha pesquisa com a replicabilidade e o avanço do conhecimento na interface entre os estudos linguísticos e o PLN.

Diante dos desafios existentes nos cenários de ensino de escrita, urge a necessidade da ampliação desse escopo de análise para diferentes contextos acadêmicos, a fim de que seja aprofundada a compreensão acerca das relações das influências de métricas textuais na qualidade de escrita e de ensino de produção textual, não somente no Ensino Médio, mas em todos os níveis escolares.

A elaboração desta pesquisa e a escrita laboriosa da minha tese de doutorado não vislumbraram a descoberta de respostas definitivas, tampouco, o engessamento de algumas questões; ao contrário disso, considerando que toda pesquisa é, em essência, um convite à busca, por mais que se chegue a algumas respostas, é na inquietação investigativa que reside o verdadeiro conhecimento.

REFERÊNCIAS

ADAM, J. M. **A linguística textual:** introdução à análise textual dos discursos. São Paulo: Cortez, 2008.

AJAY, H. B.; TILLET, P.; PAGE, E. B. **Analysis of essays by computer (AEC-II)**. Storrs, CT: Univeristy of Connecticut, 1973.

ALEX B.; HENRY, T. *Corpus*-based study of language and teacher education. *In:* M. Bigelow & J. Ennser-Kananen (eds), **The Routledge Handbook of Educational Linguistics**. New York: Routledge, p. 301-312, 2014.

ALUÍSIO, S. M.; ALMEIDA, G. M. O que é e como se contrói um Córpus?: Lições aprendidas na compilação de vários córpus para pesquisa lingüística. **Calidoscópio**, São Leopoldo, v. 4, n. 3, p. 155-177, 2006.

ALUÍSIO, S.; GASPERIN, C. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. (T. Solorio, T. Pedersen, Eds.) Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas. **Anais.** Los Angeles, California: Association for Computational Linguistics, jun. 2010. Disponível em: Acesso em: 13 out. 2024.

AMORIM, G. **Retratos da leitura no Brasil.** São Paulo: Instituto Prólivro/Imprensa Oficial do Estado de São Paulo, 2012. Disponível em: RLBhttp://cbl.org.br/pesquisas_de_mercado_categoria/4-retratos-da-leitura-no-brasil/ Acesso em: 25 out. 2024.

Panorama da reciclagem de papel no Brasil – *2023.* São Paulo: ANAP, 2023. Disponível em: https://www.anap.org.br/. Acesso em: 28 jun. 2025.

ANTUNES, Irandé. **Território das palavras:** estudo do léxico em sala de aula. São Paulo: Parábola Editorial, 2012.

SALVADOR, A.; SQUARISI, D. **Escrever Melhor:** guia para passar os textos a limpo. Editora Contexto, 2008.

AUTHIER-REVUZ, J. Heterogeneidades enunciativas. *In:* **Cadernos de estudos lingüísticos,** 19. Campinas: IEL, 1990. p. 25-42.

BAPTISTA, L. M. T. R. **Manobras e estratégias de autoria:** a singularidade do sujeito na produção escrita em língua espanhola. 2005. 322f. Tese (Doutorado) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2005.

Barreto, J. P. **A redação de vestibular sob uma perspectiva multidimensional:** uma abordagem da linguística de corpus. 2016. 208 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) - Programa de Estudos Pós-

Graduados em Linguística Aplicada e Estudos da Linguagem, Pontifícia Universidade Católica de São Paulo, São Paulo, 2016.

BAZERMAN, C. Gênero, agência e escrita. In: DIONÍSIO, Angela Paiva; HOFFNAGEL, Judith Cortesão (orgs.). **Gênero, discurso e ensino**. São Paulo: Cortez, 2006.

BERBER SARDINHA, T. Como usar a Linguística de Córpus no ensino de língua estrangeira. Ou: Por uma linguística de Córpus educacional brasileira. I*n:* TAGNIN, S.; VIANA, V. (org.). **Córpus no ensino de línguas estrangeiras.** São Paulo: Hub Editorial. p. 301-356, 2010.

BERBER SARDINHA, T. **Linguística de Córpus.** Barueri: Manole, 2004.

BERBER SARDINHA, T. Linguística de *corpus*: histórico e problemática. **Revista D.E.L.T.A**., São Paulo, v. 16, n. 2, p. 323-367, 2000.

BIBER, D. Representativeness in *corpus* design. **Literary Linguist Computing,** Arizona, v. 8 n. 4. p. 244-257, 1993.

BIBER, D., CONRAD, S., & CORTES, V. If you look at …: Lexical bundles in university teaching and textbooks. **Applied Linguistics**, v. 25, p. 371-405, 2004.

BIBER, D.; CONRAD, S.; REPPEN, R. ***Corpus* Linguistics**: investigating language structure and use. New York: Cambridge University Press, 1998.

BIDERMAN, M. T. C. Conceito linguístico de palavra. **Palavra,** Rio de Janeiro, v. 1, n. 5, 1999.

BIDERMAN, M. T. C. Dicionários do português: da tradição à contemporaneidade. **ALFA: Revista de Linguística**, São Paulo, v. 47, n. 1, 2003. Disponível em: https://periodicos.fclar.unesp.br/alfa/article/view/4232. Acesso em: 28 out. 2024.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2019:** cartilha do participante. Brasília, 2019.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2023:** cartilha do participante. Brasília, DF: INEP, 2023.

BUZATO, D.; PERUGINI, D.; MACHADO, E.; TEIXEIRA, I. Operadores argumentativos SCIENTIA PRIMA Feliz, v. 7, n. 1: e-97, 2021. 22 em aprendizes: panorama do ensino médio. **Anais do Congresso Nacional Universidade EAD e Software Livre**. Belo Horizonte: FALE/UFMG, 2021.

CASELI, H.M.; NUNES, M.G.V. (org.) **Processamento de Linguagem Natural:** Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao.

CARDOSO, J. V. S. **Argumentação linguística e polifonia na análise de redações "nota mil" do Enem**. 2022. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, Porto Alegre, 2018.

CELANI, M. A. A; Maximina M. F; Ramos, R. C. G. (org.). **A Abordagem Instrumental no Brasil:** um projeto, seus percursos e seus desdobramentos. Campinas: Mercado de Letras; São Paulo: EDUC, 2009. Coleção As Faces da Linguística Aplicada. v.10. ISBN 978-85-7591-105-1. 184p.

CHARAUDEAU, P. ***Linguagem e discurso****. Modos de organização.* Contexto : São Paulo, 2009.

COBB, T. Computing the vocabulary demands of L2 reading. **Language Learning & Technology,** Nation, p. 38–63., 2007.

COBB, T.; BOULTON, A. Classroom applications of córpus analysis. *In:* D. Biber & R. Reppen (eds), **Cambridge Handbook of English Córpus Linguistics**. Cambridge: Cambridge University Press, p. 478-497, 2015.

COSTA, J. C.; SILVA, E.B. Uma análise quantiqualitativa de redações em língua portuguesa: algumas reflexões para o ensino. *In:* GONÇALVES, R. B.; SILVA, E. B. **Recortes linguísticos sob uma perspectiva intercultural.** Maringá: Uniedusul, 2020. cap. 3. P, 36-40.

COXHEAD, A. J. **An academic word list.** English Language Institute Occasional Publication, Wellington: Victoria University of Wellington, 1998.

DORNO, T. W. **Dialética negativa***.* Tradução de Marco Antonio Casanova. Rio de Janeiro: Jorge Zahar, 2009.

DUBOIS, J.; GIACOMO, M.; GUESPIN, L.; MARCELLESI, C.; MARCELLESI, J.B.; MEVEL, J-P. **Dicionário de Linguística.** 10. ed. São Paulo: Editora Cultrix, 1998.

DUCROT, O. e TODOROV, T. **Dicionário enciclopédico das ciências da linguagem.** 3ª ed., São Paulo, Perspectiva, 2001.

ELLIS N. C.; SIMPSON-VLACH, R. **An Academic Formulas List (AFL):** *Corpus* Linguistics, Psycholinguistics, and Education. *In*: Proceedings of TaLC8 - The 8th Teaching and Language Corpora Conference, Lisbon, 2008.

EVERS, A. A redação engaiolada: padrões lexicais e ensino de redação em cursos pré-vestibulares populares. **Tese de Doutorado em Letras**. Porto Alegre, Universidade Federal do Rio Grande do Sul, 2018.

FLOWERDEW, J. Li, Y.-Y. Plagiarism and second language writing in an electronic age. **Annual Review of Applied Linguistics 27**, p. 161–83, 2009.

GARCIA , W. D; PINTO, P. T. A elaboração de uma atividade didática de língua inglesa direcionada por dados de aprendizes. **Estudos Linguísticos,** São Paulo, v. 49, n. 1, p. 528-549, 2020.

GENOUVRIER, E.; PEYTARD, J. **Linguística e ensino do português***.* Coimbra: Livraria Almedina, 1985.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GRAMA, D. F. WordSmith Tools: para uma análise da coesão sequencial em redações dissertativas argumentativas. **Estudos Linguísticos,** São Paulo, 2016.

GRIES, STEFAN Th. **Frequency, dispersion, association, and keyness:** revising and tupleizing corpus-linguistic measures. Amsterdam & Philadelphia: John Benjamins, 2024.

GUERRA, M. M. e ANDRADE, K. de S. O léxico sob perspectiva: contribuições da Lexicologia para o ensino de línguas. **Domínios da Linguagem**. v. 6, n° 1, 1° Semestre, 2012.

FINATTO, Maria José Bocorny; CASELI, Helena de Medeiros; LOPES, Lucelene; RASSI, Amanda. Sequência de caracteres e palavras: morfologia e morfossintaxe. *In:* CASELI, H. M.; NUNES, M. G. V. (org.). **Processamento de linguagem natural:** conceitos, técnicas e aplicações em português. 2. ed. [S. l.]: BPLN, 2024. cap. 4. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/parte-palavras/cap-caracteres-palavras/cap-caracteres-palavras.html. Acesso em: 4 set. 2024.

HALLIDAY, M. A. K. Quantitative studies and probabilities in grammar. *In:* M. HOEY (ed.) **Data, description, discourse** - papers on the English language in honour of John McH Sinclair on his sixtieth birthday. London: HarperCollins, 1993.

HOFFMANN, L. Possibilidades de aplicação e a aplicação atual de métodos estatísticos 115 na pesquisa de linguagens especializadas. **Cadernos de Tradução,** Porto Alegre, nº 20, p. 61-76, 1998.

HUNSTON, S. Córpus in Applied Linguistics. London: Cambridge University Press, Johns, T. Whence and whither classroom concordancing? *In* T. Bongaerts, P. de Haan, S. Lobbe, and H. Wekker (eds.), **Computer applications in language learning**, 9–27, 1998.

INAF/INSTITUTO PAULO MONTENEGRO. Indicador de Alfabetismo Funcional (Inaf) – Principais Resultados. [S.*l.*]. 2011.

INDURKHYA, N.; DAMERAU, F. J. **Handbook of Natural Language Processing.** 2nd. ed. [S.l.]: Chapman & Hall/CRC, 2010.

Kaplan, R. B. **The Oxford handbook of applied linguistics.** New York: Oxford University Press, 2002.

LEAL, S. E. et al. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **Language Resources and Evaluation,** 2023.

LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular.** Tese de doutorado—[s.l.] Universidade de São Paulo, 2021.

LEAL, S. E**.** Simpligo ranking. **NILC,** 2020. Disponível em: http://fw.nilc.icmc.usp.br:23380/simpligo-ranking. Acesso em: 20 out. 2024.

Leal, S. E., Duran, M. S., and Aluísio, S. M. (2018). A nontrivial sentence *corpus* for the task of sentence readability assessment in Portuguese. *In:* **Proceedings of the 27th International Conference on Computational Linguistics**, pages 401–413.

LEAL, S. E.; LUKASOVA, K.; CARTHERY-GOULART, M. T.; ALUISIO, S. M. Rastros project: Natural language processing contributions to the development of an eyetracking *corpus* with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation**, v.56(4), p.1333– 1372, 2022b. DOI https://doi.org/10.1007/s10579-022-09609-0

LEAL, S. E.; SCARTON, C.; CUNHA, A.; HARTMANN, N.; DURAN, M.; ALUÍSIO, S. **NILC-Metrix,** NILC, 2022a. Disponível em: http://fw.nilc.icmc.usp.br:23380/nilcmetrix. Acesso em: 20 fev. 2024.

LEWIS, C. S. **The pilgrim's regress:** an allegorical apology for Christianity, reason and romanticism. London: Geoffrey Bles, 1956.

LORENTE, M. A lexicologia como ponto de encontro entre a gramática e a semântica. *In:* ISQUERDO, A. N. e KRIEGER, M. G. **As ciências do léxico.** vol. 2. Campo Grande: UFMS, 2004. p. 19 – 30.

MARTINS, T. B. F. *et al.* **Readability formulas applied** to **textbooks** in **brazilian portuguese**. São Carlos, SP: ICMSC-USP, 1996.

McENERY, T. e WILSON, A. **Córpus linguistics.** Edinburgh, Edinburgh University Press, 1996.

MCKAY, S. On notional syllabus. **Modern Language Journal,** v. 64, p. 179-186, 1980. Disponível em: https://doi.org/10.1111/j.1540-4781.1980.tb05182.x. Acesso em: 13 out. 2024.

MILLER, C. Estudos sobre gênero textual, agência e tecnologia. *In*: DIONÍSIO, A. P.; HOFFNAGEL, J. C. (Org.). **Estudos sobre Gênero Textual, Agência e Tecnologia de Carolyn R. Miller.** Recife: Universitária da UFPE, 2009.

NATION, I.S.P; BEGLAR, D. A vocabulary size test. ***The*** **Language Teacher**, v. 31, 7. ed., 2007.

NOVODVORSKI, A.; BOCORNY FINATTO, M. J. Linguística de Córpus no Brasil: uma aventura mais do que adequada. **Letras & Letras,** v. 30, n. 2, p. 7-16, 18 dez. 2014.

OLIVEIRA, F. C. C. **Um estudo sobre a caracterização do gênero redação do ENEM.** 2016. 166 f. Tese (Doutorado em Linguística) - Centro de Humanidades, Programa de Pós-Graduação em Linguística, Universidade Federal do Ceará, Fortaleza, 2016.

OLIVEIRA, L.P. Linguística de Córpus: teoria, interfaces e aplicações. **Revista Matraga**, v. 16, n. 24, RJ: Programa de Pós-graduação em Letras da UERJ, p. 48-76, 2009.

PAGE, E. B.; PETERSEN, N. S. The Computer Moves into Essay Grading: Updating the Ancient Test. **Phi Delta Kappan**, v. 76, p. 561–565, mar. 1995.

PÊCHEUX, M. **O discurso:** estrutura ou acontecimento. Tradução de Eni Pulcinelli Orlandi. 2.ed. São Paulo: Pontes, 1997a.

PERELMAN, C.; OLBRECHTS-TYTECA, L. **Tratado de argumentação:** nova retórica. Tradução de Maria Ermantina de Almeida Prado Galvão. São Paulo: Martins Fontes, 2005.

PINTO, P. T. *et al.* **Estratégias de Leitura em Língua Inglesa:** Nível básico. Araraquara: Letraria, 2021. 69 p. Disponível em: https://www.letraria.net/estrategias-de-leitura-emlingua-inglesa-nivel-basico/. Acesso em 23 nov. 2024.

PINTO, P. T.; CROSTHWAITE, P.; CARVALHO, C.; SPINELLI, F.; SERPA, T.; ORENHA-OTTAIANO, A. **Using language data to learn about language:** a teachers' guide to classroom *corpus* use. Brisbane: The University of Queensland, 2023. E-book: DOI: https://doi.org/10.14264/3bbe92d. Disponível em: https://uq.pressbooks.pub/using-language-data/. Acesso em: 3 nov. 2024.

PINTO, P. T.; GARCIA, W. D.; SERPA, T. **Caderno de atividades de aprendizagem movida por dados**. Campinas: Pontes Editores, 2022. E-book: 222 p. Disponível em: https://ponteseditores.com.br/loja3/pontes-editores-home-2__trashed/ensino-de-linguas /caderno-de-atividades-de-aprendizagem-movida-por-dados/. Acesso em: 14 out. 2024.

RANGER, S. The contribution of learner córpus to second language acquisition and foreign language teaching: A critical evaluation. *In:* K. Aijmer (ed.), **Córpus and language teaching**, p. 13–32. Amsterdam: John Benjamins, 2009.

REPPEN, R. Building a Córpus: What are the key considerations? *In:* A. O'Keeffe and M. McCarthy (eds.), **The Routledge handbook of Córpus linguistics**, p. 31–37. London: Routledge, 2015.

ROCCO, M. T. F. **Crise na linguagem:** a redação no vestibular. São Paulo: Mestre Jou, 1981.

SANTOS, L. V. DOS, PINTO, P. T. Data Driven Learning para aprendizes iniciantes de inglês: uma experiência em um curso de estratégias de leitura. **Revista Horizontes De Linguistica Aplicada**, 2024. Disponível em: https://doi.org/10.26512/rhla.v23i1.53170. Acesso em: 12 out. 2024.

SARMENTO, S. **Linguística de Córpus:** histórico, metodologia, campos de aplicação. Revista Trama, v. 6, n. 2, p. 87-107, 2010.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática**, v. 2, n. 1, p. 45-61, 2010.

SCARTON, C. *et al.* Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. **Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies**, p. 41–44, 2010.

SCARTON, C.; OLIVEIRA, M.; JÚNIOR, A. C.; GASPERIN, C.; ALUÍSIO, S. **Simplifica:** a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. Proceedings of the NAACL HLT 2010 Demonstration Session, 2010b. p. 41–44.

SCHLATTER, M.; GARCEZ, P. M. **Linguas adicionais na escola: aprendizagens colaborativas em inglês,** Erechim, RS: Edelbra, 2012.

SCHMITT, N. and SCHMITT, D. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. **Language Teaching,** Cambridge, v. 47, 4. ed. p. 484–503, 2014.

SINCLAIR, J. Córpus, **Concordance,** Collocation. Oxford: OUP, 1991.

SQUARISI, D.; SALVADOR, A.; **A arte de escrever bem:** um guia para jornalistas e profissionais do texto. 7. ed. São Paulo: Contexto, 2012.

STUBBS, M. **Words and Phrases.** Oxford: Blackwell, 2001.

SVARTVIK, J. Córpus are becoming mainstream. *In:* THOMAS, J. and SHORT, Thorndike, E. L. 1921. **Word knowledge in elementary school.** Teachers College Record, p. 334–370, 1921.

TESNIÈRE, L. **Éléments de Syntaxe Structruale**. Paris : C. Klincksieck, 1965.

THORNDIKE, E. L. **The fundamentals of learning.** New York: Teachers College, Columbia University, 1932.

TOGNINI-BONELLI, E. Towards Translation Equivalence from a *Corpus* Linguistics Perspective. *In:* SINCLAIR, J. et al. (Ed.). **Grammar patterns.** London: Collins COBUILD, 1996, p. 197-217.

TRASK, R.L. **Dicionário de Linguagem e Lingüística.** São Paulo, Contexto, 2004.

WEST, M. **A General Service List of English Words:** with semantic frequencies and a supplementary wordlist for the writing of popular science and technology. [S. *l.*]: Longman, 1953.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	JEANE CARDOSO COSTA 04/09/1987
Nacionalidade	brasileira
Nome em citações bibliográficas:	Costa, Jeane Cardoso. Costa, J. C.
Currículo Lattes	http://lattes.cnpq.br/9584646749205773
ORCID	https://orcid.org/0000-0001-6413-2959
outro identificador	URL
FORMAÇÃO ACADÊMICA	
0000/9999	Formação acadêmica (nome do curso e titulação) Instituição de ensino
0000/9999	Formação complementar (nome do curso e titulação) Instituição de ensino
PRODUÇÃO BIBLIOGRÁFICA	
SOBRENOME, N. ; SOBRENOME, N. Título: subtítulo. *In:* NOME DO EVENTO, 17., 2012, Cidade de realização. **Anais** […]. Cidade de publicação: Editora, 2012. p. 9-23.	
PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES	
Bancas de trabalhos de conclusão	
SOBRENOME, N.; SOBRENOME, N.; SOBRENOME, N. Participação em banca de Nome Completo do Aluno. **Título do trabalho**. 2021. Dissertação (Mestrado em Nome do curso) - Nome da Universidade, Cidade, 2021.	
Orientações	
SOBRENOME, N. **Título**: subtítulo. 2011.Orientador: Nome Completo do Orientador. 2011. Trabalho de Conclusão de Curso (Graduação em Nome do Curso) - Nome da Universidade, Cidade, 2011	
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	

NOME DO EVENTO, 17., 2012, (Cidade de realização). Título do trabalho apresentado: subtítulo. 2012. (Seminário).

NOME DO EVENTO, 7., 2017, (Cidade de realização). Mini-curso: Título do minicurso. 2017. (Seminário).