

**UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS**

LUIZA SARMENTO DIVINO

**CONTRIBUIÇÕES DA LINGUÍSTICA DE CORPUS PARA A DESCRIÇÃO DOS
NÍVEIS DE PROFICIÊNCIA ESCRITA NO EXAME CELPE-BRAS: UM ESTUDO
SOBRE LÉXICO**

PORTO ALEGRE

2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS
PROGRAMA DE PÓS GRADUAÇÃO EM LETRAS
ÁREA DE CONCENTRAÇÃO: ESTUDOS LINGUÍSTICOS
LINHA DE PESQUISA: LINGUÍSTICA APLICADA

LUIZA SARMENTO DIVINO

**CONTRIBUIÇÕES DA LINGUÍSTICA DE CORPUS PARA A DESCRIÇÃO DOS
NÍVEIS DE PROFICIÊNCIA ESCRITA NO EXAME CELPE-BRAS: UM ESTUDO
SOBRE LÉXICO**

Dissertação de Mestrado apresentada como requisito parcial para a obtenção de título de Mestre em Linguística Aplicada pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.
Orientadora: Juliana Roquele Schoffen

PORTO ALEGRE

2024

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Carlos André Bulhões Mendes

Vice-reitora: Patricia Pranke

INSTITUTO DE LETRAS

Diretora: Carmem Luci da Costa Silva

Vice-diretora: Márcia Montenegro Velho

PROGRAMA DE PÓS-GRADUAÇÃO EM LETRAS

Coordenadora: Elaine Barros Indrusiak

Coordenadora Substituta: Elisa Battisti

CIP - Catalogação na Publicação

Sarmiento Divino, Luiza
CONTRIBUIÇÕES DA LINGUÍSTICA DE CORPUS PARA A
DESCRIÇÃO DOS NÍVEIS DE PROFICIÊNCIA ESCRITA NO EXAME
CELPE-BRAS: UM ESTUDO SOBRE LÉXICO / Luiza Sarmiento
Divino. -- 2024.
195 f.
Orientador: Juliana Roquele Schoffen.

Dissertação (Mestrado) -- Universidade Federal do
Rio Grande do Sul, Instituto de Letras, Programa de
Pós-Graduação em Letras, Porto Alegre, BR-RS, 2024.

1. Avaliação de proficiência. 2. Exame Celpe-Bras.
3. Português como Língua Adicional. 4. Corpus de
língua adicional. 5. Linguística de Corpus. I. Roquele
Schoffen, Juliana, orient. II. Título.

LUIZA SARMENTO DIVINO

**CONTRIBUIÇÕES DA LINGUÍSTICA DE CORPUS PARA A DEFINIÇÃO DE
NÍVEIS DE PROFICIÊNCIA ESCRITA NO EXAME CELPE-BRAS**

Dissertação de Mestrado apresentada como requisito parcial para a obtenção de título de Mestre em Linguística Aplicada pelo Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.
Orientadora: Juliana Roquele Schoffen

Porto Alegre, 21 de Março de 2024.

Resultado: Aprovado com conceito “A com louvor”.

BANCA EXAMINADORA:

Profa. Dra. Juliana Roquele Schoffen (orientadora)
Universidade Federal do Rio Grande do Sul

Profa. Dra. Ana Eliza Pereira Bocorny
Universidade Federal do Rio Grande do Sul

Profa. Dra. Elisa Marchioro Stumpf
Universidade Federal do Rio Grande do Sul

Profa. Dra. Tanara Zingano Kuhn
Universidade de Coimbra

À Jurita e à Gertrudes.

AGRADECIMENTOS

Esta dissertação foi escrita com apoio de muitas mãos. Gostaria de agradecer a cada mão que fez com que este trabalho se concretizasse:

À Universidade Federal do Rio Grande do Sul (UFRGS), por todas as oportunidades, pela excelente educação e pela possibilidade de estudar gratuitamente.

Ao Programa de Pós Graduação em Letras (PPG-Letras), pela qualidade, suporte e investimento financeiro, tornando possível a realização deste estudo.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa, permitindo que a experiência da pesquisa fosse vivida com mais dedicação.

À Juliana, que despertou em mim o interesse pela pesquisa acadêmica e me orientou desde o início do meu percurso, por trabalhar incansavelmente para o desenvolvimento de seus orientandos, pela parceria ao longo de todos esses anos e por acreditar em mim.

Aos colegas de pesquisa do Grupo Avalia, especialmente à Ellen Nagasawa, à Gabrielle Sirianni, à Kaiane Mendel, à Margarete Schlatter e à Simone Kunrath, que me acolheram quando me tornei membro do grupo, em 2017, e à Giovana, e à Isabel, com quem divido anseios e vivências também fora da academia, e pelo companheirismo e pelo senso de pertencimento.

À ramificação do Grupo Avalia responsável pela compilação do Corpus do Celpe-Bras (CorCel) e aos bolsistas de digitação e revisão, por tornaram possível a concretização deste trabalho.

Às professoras Ana Beatriz Fontes, Ana Eliza Pereira Bocorny, Juliana Roquele Schoffen, Karen Pupp Spinassé, Luciene Juliano Simões, Rozane Rodrigues Rebechi, Simone Sarmento e ao professor William Kirsch, pelas disciplinas de excelente qualidade, de grande contribuição para a minha formação como pesquisadora e professora.

Aos meus colegas de pós-graduação, pela oportunidade de participar em discussões aprofundadas e de compartilhar o conhecimento. Ao Camilo, por tornar a experiência do mestrado mais leve e divertida.

À Deise e à Marine, por terem me guiado, ainda na graduação, pelo universo de Linguística de Corpus, e pela motivação nos momentos finais da escrita deste trabalho.

A todas as pessoas para quem já dei aula, por me fazerem ter a certeza de que escolhi o caminho certo.

À banca examinadora, pela certeza da leitura atenta e cuidadosa, que com certeza serão de grande contribuição para este trabalho. À Ana Bocorny, por apresentar a autores que me permitiram expandir meu conhecimento sobre a área da Linguística de Corpus e me capacitaram para a escrita deste trabalho. À Elisa Stumpf, pelo trabalho conjunto e valiosas discussões que enriqueceram minha compreensão da Linguística de Corpus e contribuíram significativamente para o desenvolvimento deste estudo. À Tanara Kuhn, por sua contribuição como uma das referências utilizadas neste trabalho, enriquecendo minha compreensão da área.

À Simone, minha mãe, por ser a minha maior inspiração, por não medir esforços em me ajudar em qualquer situação, por me dar forças para realizar meus sonhos.

Ao Vitor, meu irmão, pela amizade e pela cumplicidade sempre. À Camila, minha cunhada, pelos momentos de descontração e a certeza de boas risadas mesmo em períodos de estresse.

À Luci e ao Juca, meus avós, por todo o amor e cuidado que sempre tiveram comigo.

Ao João, meu pai, por sempre tentar me entender e me apoiar, e por ter me apresentado a maioria das músicas que fizeram parte da trilha sonora que acompanhou a escrita deste trabalho. À minha madrastra, por toda a preocupação e pelas mensagens de carinho.

Ao Leandro e às minhas irmãs de Campinas, que sempre vibraram com as minhas realizações, pelo conforto e atenção.

Às minhas amigas, com quem compartilho angústias e alegrias, por sempre me darem o suporte e o incentivo que eu precisei para seguir em frente. Ao Gabo, meu amigo e colega de apartamento, com quem compartilho o cotidiano, por tornar os meus dias mais divertidos e por sempre me manter alimentada.

Ao Pissetti, meu amor, por ser a pessoa mais paciente, carinhosa e atenciosa em todos os momentos, por toda a ajuda, por tornar este período mais leve.

À Neni e ao Flávio, meus companheiros felinos, que viveram comigo as madrugadas de escrita.

A todos aqueles que, de alguma forma, contribuíram para que eu pudesse realizar este trabalho.

Convite

*Poesia
é brincar com palavras
como se brinca
com bola, papagaio, pião.*

*Só que
bola, papagaio, pião
de tanto brincar
se gastam.*

*As palavras não:
quanto mais se brinca
com elas
mais novas ficam.*

*Como a água do rio
que é água sempre nova.*

*Como cada dia
que é sempre um novo dia.*

Vamos brincar de poesia?

(Paes, 1989).

RESUMO

Desde a virada do milênio, percebe-se o aumento do interesse no uso da Linguística de Corpus para informar a validação de exames de língua (Cushing, 2017; 2022), havendo um aumento significativo no uso de corpora de línguas adicionais em pesquisas voltadas para avaliações de proficiência (Wisniewski, 2017). Resultados destas pesquisas são úteis para um maior entendimento das classificações de proficiência em um nível empírico (Callies; Díez-Bedmar; Zaytseva, 2014; Barker; Salamoura; Saville; 2015; Callies; Götz, 2015). O Celpe-Bras, Certificado de Proficiência em Língua Portuguesa para Estrangeiros, é um exame brasileiro que busca avaliar as práticas linguísticas em Português como Língua Adicional em contextos variados (Brasil, 2020). Com o intuito de dar seguimento a trabalhos que se debruçaram sobre os níveis de proficiência escrita avaliados no Celpe-Bras (Sirianni, 2016; Mendel, 2019; Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023), esta dissertação se propõe a responder à pergunta: Como as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência em tarefas de leitura e escrita no exame Celpe-Bras? Para isto, foi analisado um corpus não anotado de 2.293 textos produzidos por examinandos em resposta à Tarefa 4 da edição de 2015/2 do Celpe-Bras, avaliados com notas 0, 1, 2, 3, 4 e 5 (CorCel20152t4). Os objetivos específicos se dividem em três grupos: o primeiro deles diz respeito à extensão dos textos, incluindo análises das quantidades totais, médias e normalizadas de *types*, *tokens* e sentenças, bem como o cálculo *Type-Token-Ratio*; o segundo grupo é voltado para a relação do léxico utilizado em cada subcorpus com o material de insumo, com análises de *n-grams*, *keywords* e *wordlists* no *Sketch Engine* (Kilgarriff *et al.*, 2004); e o terceiro grupo se refere à incidência de termos relacionados à configuração da relação de interlocução dentro do gênero do discurso proposto, com foco na ferramenta *concordance* do *Sketch Engine*. Os resultados sugerem que a extensão dos textos é um fator importante no que diz respeito à proficiência, com textos mais avançados sendo, em geral, maiores. Quanto aos termos relacionados à construção do gênero, estes são mais recorrentes conforme a nota aumenta, mas não estão presentes em 100% dos textos, confirmando que há diversas possibilidades de se cumprir adequadamente a tarefa. Os dados também desmistificam a ideia de que a cópia é realizada apenas nos níveis mais baixos: apesar da quantidade de cópias de trechos mais longos ser, proporcionalmente, mais alta nos corpora n3, n2 e n1 do que nos corpora n5 e n4, foram encontrados *n-grams* com cópia de fragmentos do material de insumo em todas as notas. Foi identificado ainda que o corpus n5 apresenta mais palavras diferentes das que constam no material de insumo, mas que são morfologicamente semelhantes às apresentadas neste material. Com isto, infere-se que estes examinandos conseguem transitar mais na língua portuguesa ao demonstrarem conhecer seu funcionamento.

Palavras-chave: Avaliação de proficiência; Exame Celpe-Bras; Português como Língua Adicional; Corpus de língua adicional; Linguística de Corpus.

ABSTRACT

Since the turn of the millennium, there has been a growing interest in using Corpus Linguistics to inform the validation of language exams (Cushing, 2017; 2022), with a significant increase in the use of additional language corpora in research focused on proficiency assessments (Wisniewski, 2017). Findings from these studies are valuable for a deeper understanding of proficiency classifications at an empirical level (Callies; Díez-Bedmar; Zaytseva, 2014; Barker; Salamoura; Saville; 2015; Callies; Götz, 2015). The Celpe-Bras, Certificate of Proficiency in Portuguese for Foreigners, is a Brazilian exam designed to assess language practices in Portuguese as an Additional Language in various contexts (Brazil, 2020). Building on previous work that examined written proficiency levels assessed in the Celpe-Bras (Sirianni, 2016; Mendel, 2019; Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023), this dissertation aims to address the question: How can lexical analysis conducted with Corpus Linguistics tools contribute to describing proficiency levels in reading and writing tasks in the Celpe-Bras exam? To achieve this, an unannotated corpus of 2,293 texts produced by exam takers in response to Task 4 of the 2015/2 edition of Celpe-Bras, evaluated with scores ranging from 0 to 5, was analyzed (CorCel20152t4). The specific objectives are divided into three groups: the first group concerns the text's length, including analyses of total, average, and normalized quantities of types, tokens, and sentences, as well as Type-Token-Ratio calculation; the second group focuses on the relationship between the lexicon used in each subcorpus and the input material, with analyses of n-grams, keywords, and wordlists in the Sketch Engine (Kilgarriff *et al.*, 2004); and the third group refers to the incidence of terms related to the configuration of the interlocution relationship within the proposed discourse genre, with a focus on the concordance functionality of the Sketch Engine. The results suggest that text length is an important factor regarding proficiency, with more advanced texts generally being longer. As for the terms related to genre construction, they become more frequent as the score increases, but they are not present in 100% of the texts, confirming that there are different ways to adequately complete the task. The data also debunk the notion that copying is only done at lower levels. Despite the proportionally higher amount of longer passage copying in corpora n3, n2, and n1 than in corpora n5 and n4, n-grams copying fragments from the input material were found at all levels. In addition, corpus n5 presented more different words than those in the input material, but morphologically similar to those in that material. Therefore, it can be inferred that these examinees can navigate more freely in the Portuguese language by demonstrating an understanding of its functioning.

Keywords: Proficiency assessment; Celpe-Bras exam; Portuguese as a Foreign Language; Additional language corpus; Corpus linguistics.

LISTA DE FIGURAS

Figura 1 - Número de examinandos anuais do Exame Celpe-Bras.....	28
Figura 2 - Escala de proficiência.....	32
Figura 3 - Enunciado da Tarefa 4 de 2015-2.....	41
Figura 4 - Design da Tarefa 4 de 2015-2.....	42
Figura 5 - Pesos da avaliação do Avaliador Observador na Parte Oral.....	46
Figura 6 - Níveis de certificação por faixa de nota.....	48
Figura 7 - Resultado da ferramenta <i>Word Sketch</i> no <i>Sketch Engine</i>	59
Figura 8 - Resultado da ferramenta <i>Thesaurus</i> como lista no <i>Sketch Engine</i>	60
Figura 9 - Resultado da ferramenta <i>Thesaurus</i> como gráfico em bolhas no <i>Sketch Engine</i>	60
Figura 10 - Resultado da ferramenta <i>Thesaurus</i> como nuvem de palavras no <i>Sketch Engine</i> .	61
Figura 11 - Resultado do recurso <i>Collocations</i> no <i>Sketch Engine</i>	63
Figura 12 - Resultado da ferramenta <i>Wordlist</i> no <i>Sketch Engine</i>	64
Figura 13 - Resultado da ferramenta <i>N-grams</i> no <i>Sketch Engine</i>	65
Figura 14 - Resultado da ferramenta <i>Keywords</i> no <i>Sketch Engine</i>	66
Figura 15 - Fórmula do cálculo <i>Type-Token-Ratio</i>	90
Figura 16 - Distribuição dos textos em cada nota no CorCel2015t4.....	94
Figura 17 - Combinações entre termos de saudação/endereçamento.....	136
Figura 18 - Colocações utilizados para identificação.....	142
Figura 19 - Colocações de “atenciosamente” e “cordialmente”.....	144

LISTA DE QUADROS

Quadro 1 - Estrutura da Parte Escrita.....	36
Quadro 2 - Especificações da Parte Escrita.....	37
Quadro 3 - Estrutura da Parte Oral.....	43
Quadro 4 - Classificação de Corpora.....	68
Quadro 5 - <i>N-grams</i> do corpus n5 separados por grupos de informações.....	103
Quadro 6 - <i>N-grams</i> do corpus n4 separados por grupos de informações.....	104
Quadro 7 - <i>N-grams</i> do corpus n3 separados por grupos de informações.....	106
Quadro 8 - <i>N-grams</i> do corpus n2 separados por grupos de informações.....	110
Quadro 9 - <i>N-grams</i> do corpus n1 separados por grupos de informações.....	112
Quadro 10 - Extensão dos trechos encontrados nos <i>n-grams</i> em cada corpus.....	114
Quadro 11 - Ordem de relevância e <i>Relative DOCF</i> (%) das primeiras 15 <i>keywords</i> em cada subcorpus.....	120
Quadro 12 - <i>Keywords</i> que não estão no material de insumo ordenadas por <i>Relative DOCF</i> (%).....	121
Quadro 13 - Palavras morfologicamente semelhantes a palavras do material de insumo.....	123
Quadro 14 - <i>LogDice</i> dos primeiros colocados mais frequentes dos pronomes possessivos em cada corpus.....	147

LISTA DE TABELAS

Tabela 1 - Número de textos, número de <i>types</i> , número de <i>tokens</i> e <i>TTR</i> dos subcorpora do CorCel20152t4.....	95
Tabela 2 - Número de <i>types</i> , número de <i>tokens</i> e <i>TTR</i> dos subcorpora do CorCel20152t4 normalizado.....	96
Tabela 3 - Média de <i>tokens</i> , texto com o menor número de <i>tokens</i> e texto com o maior número de <i>tokens</i> em cada subcorpus do CorCel20152t4.....	97
Tabela 4 - Número total de sentenças e média de sentenças por texto em cada subcorpus do CorCel20152t4.....	99
Tabela 5 - Palavras que aparecem em pelo menos 5% dos textos em cada corpus.....	125
Tabela 6 - Relação entre as palavras que estão em pelo menos 5% dos textos e o número total de <i>types</i>	127
Tabela 7 - Palavras que aparecem em pelo menos 10% dos textos em cada corpus.....	128
Tabela 8 - Palavras que aparecem em pelo menos 20% dos textos em cada corpus.....	129
Tabela 9 - Palavras que aparecem em pelo menos 50% dos textos em cada corpus.....	130
Tabela 10 - <i>Relative DOCF</i> de palavras de conteúdo com alta frequência nos corpora.....	131
Tabela 11 - Palavras de conteúdo com aumento da <i>Relative DOCF</i> do corpus n5 ao corpus n1.....	133
Tabela 12 - <i>Relative DOCF</i> de palavras que remetem a saudação/endereçamento.....	135
Tabela 13 - Frequências absolutas de “prefeitura” e “municipal” e <i>logDice</i> entre ambas.....	136
Tabela 14 - Frequências absolutas de “prezada prefeitura” e “prezada prefeitura municipal” e <i>logDice</i> das colocações.....	137
Tabela 15 - Frequências absolutas de “senhores” e “prefeitura” e <i>logDice</i> entre ambas.....	137
Tabela 16 - Frequências absolutas de “prezados senhores” e “prezados senhores da prefeitura” e <i>logDice</i> das colocações.....	138
Tabela 17 - Frequências absolutas de “prefeito” e “municipal” e <i>logDice</i> entre ambas.....	138
Tabela 18 - Frequências absolutas de “senhor”, “sr.” e “sr” e <i>logDice</i> de cada termo com “prefeito”.....	139
Tabela 19 - Frequências absolutas e <i>logDice</i> dos termos de saudação associados a “prefeito”...	140

Tabela 20 - <i>Relative DOCF</i> de palavras que remetem a despedida e assinatura.....	141
Tabela 21 - Frequências absolutas de “bibiana” e “cambará” e <i>logDice</i> entre ambas.....	142
Tabela 22 - Frequências absolutas e <i>logDice</i> de termos de identificação sem nome próprio.	143
Tabela 23 - Frequência absolutas e <i>logDice</i> de colocações com os termos “atenciosamente” e “cordialmente”	145
Tabela 24 - <i>Relative DOCF</i> possessivos no singular e no plural.....	146

LISTA DE ANEXOS

Anexo A - Tarefa IV (Azulejos Valiosos) da edição de 2015/2.....	177
Anexo B - Parâmetros de Avaliação da Parte Escrita.....	178
Anexo C - Tutorial de Digitações do Corpus de Textos de Examinandos do Celpe-Bras	179
Anexo D - <i>N-grams</i> do corpus n4 separados por grupos de informações.....	182
Anexo E - <i>N-grams</i> do corpus n3 separados por grupos de informações.....	184
Anexo F - <i>N-grams</i> do corpus n2 separados por grupos de informações.....	187
Anexo G - <i>N-grams</i> do corpus n1 separados por grupos de informações.....	190

LISTA DE ABREVIATURAS E SIGLAS

AI - Avaliador-Interlocutor

AO - Avaliador-Observador

CAL2 - Corpus de Aquisição de L2

CAPLE - Centro de Avaliação de Português Língua Estrangeira

CELGA - ILTEC - Centro de Estudos de Linguística Geral e Aplicada da Universidade de Coimbra

Celpe-Bras - Certificado de Língua Portuguesa para Estrangeiros

CLUL - Centro de Linguística da Universidade de Lisboa

CLUNL - Centro da Universidade NOVA de Lisboa

COPLE2 - Corpus de Português Língua Estrangeira/Língua Segunda

CorCel - Corpus do Celpe-Bras

CQL - Corpus Query Language

DOCF - Frequência por Documento

e-SIC - Sistema Eletrónico do Serviço de Informação ao Cidadão

FLUC - Faculdade de Letras da Universidade de Coimbra

FLUL - Faculdade de Letras da Universidade de Lisboa

IA - Inteligência Artificial

IC - Iniciação Científica

ICLP - Instituto de Cultura e Língua Portuguesa

IES - Instituições de Ensino Superior

IL - Instituto de Letras

IsF - Idiomas sem Fronteiras

Inep - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

LA - Linguística Aplicada

LC - Linguística de Corpus

LL - LogLikelihood

Macaws - Multilingual Academic Corpus of Assignments - Writing and Speech

MEC - Ministério da Educação

MERCOSUL - Mercado Comum do Sul

Nele - Núcleo de Ensino de Línguas em Extensão

PALOP - Países Africanos de Língua Oficial Brasileira

PEAPL2 - Produções Escritas de Aprendentes de PL2

PE - Parte Escrita

PEG-G - Programas de Convênio de Graduação

PEC-PG - Programas de Convênio de Pós-Graduação

PLA - Português como Língua Adicional

PLM - Português como Língua Materna

PO - Parte Oral

PPE - Programa de Português para Estrangeiros

PPGLET - Programa de Pós-Graduação em Letras

PUCRS - Pontifícia Universidade Católica do Rio Grande do Sul

Relative DOCF - Frequência Relativa por Documento

SESu - Secretaria de Educação Superior

SIPLE - Sociedade Internacional de Português Língua Estrangeira

SkE - Sketch Engine

UFRGS - Universidade Federal do Rio Grande do Sul

TCC - Trabalho de Conclusão de Curso

TTR - Type-Token-Ratio

Unicamp - Universidade Estadual de Campinas

UFPE - Universidade Federal de Pernambuco

UFRGS - Universidade Federal do Rio Grande do Sul

UFRJ - Universidade Federal do Rio de Janeiro

UnB - Universidade de Brasília

APOIO DE FINANCIAMENTO CAPES

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código 001.

SUMÁRIO

CARTA ABERTA: Introdução.....	19
1. PATRIMÔNIO HISTÓRICO, ARTÍSTICO E CULTURAL: Celpe-Bras: elaboração, histórico, concepções teóricas e estrutura.....	24
1.1 Contexto de elaboração do Celpe-Bras.....	28
1.2 Especificações do exame Celpe-Bras.....	28
1.2.1 A concepção teórica do exame.....	29
<i>1.2.1.1 Gêneros do discurso e a noção de proficiência.....</i>	<i>31</i>
1.2.2 Estrutura do exame.....	34
1.2.2.1 <i>Parte Escrita.....</i>	<i>35</i>
1.2.2.1.1 <i>Integração de habilidades.....</i>	<i>38</i>
1.2.2.2 <i>Parte Oral.....</i>	<i>42</i>
1.2.3 Avaliação do exame.....	44
1.2.3.1 <i>Processo de avaliação da Parte Escrita.....</i>	<i>44</i>
1.2.3.2 <i>Processo de avaliação da Parte Oral.....</i>	<i>46</i>
1.2.3.3 <i>Resultado final.....</i>	<i>47</i>
1.3 Validade, confiabilidade, autenticidade, impacto e praticidade.....	48
1.4 Resumo do capítulo.....	52
2. CASARÕES HISTÓRICOS: Pressupostos teóricos.....	53
2.1 Campo de estudos da Linguística de Corpus.....	53
2.1.1 Como pode ser utilizado um corpus?.....	56
2.1.2 Abordagens de pesquisa.....	57
2.1.3 Técnicas de manuseio.....	59
2.1.4 Categorização e discussão terminológica: problemática em “corpora de aprendizes”.....	67
2.1.5 Corpora de línguas adicionais e avaliação de proficiência.....	72
2.1.6 Corpora de PLA e contribuições para o ensino e a avaliação.....	75
2.2 Resumo do capítulo.....	81
3. MEDIDAS IMEDIATAS: Procedimentos metodológicos.....	82

3.1 Perguntas de pesquisa.....	82
3.2 Objetivos do trabalho.....	83
3.3 CorCel: o corpus da pesquisa.....	84
3.4 Apresentação da tarefa.....	85
3.5 Compilação do corpus.....	86
3.6 Procedimentos para as análises.....	88
3.6.1 Análise da extensão dos corpora.....	88
3.6.2 Análise do léxico utilizado nos corpora.....	90
3.7 Resumo do capítulo.....	93
4. AZULEJOS VALIOSOS: Análises dos dados.....	94
4.1 Extensão dos subcorpora do CorCel20152t4.....	94
4.2 Léxico utilizado no CorCel20152t4.....	101
4.2.1 N-grams.....	102
4.2.2 Keywords.....	118
4.2.3 Wordlist.....	124
4.2.4 Concordance.....	134
4.3 Resumo do capítulo.....	149
5. PROCESSOS DE TOMBAMENTO: Discussão dos resultados.....	151
5.1 O que a extensão dos textos dos subcorpora informa sobre os níveis de proficiência?.....	151
5.2 Qual a relação do léxico usado em cada subcorpus com o material de insumo da tarefa?.....	153
5.3 Quais semelhanças e diferenças na utilização de termos relacionados ao gênero do discurso solicitado na tarefa podem ser encontradas nos subcorpora?.....	160
5.4 Como as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência em tarefas de leitura e escrita no exame Celpe-Bras?.....	162
6. MERCADO DE AZULEJOS: Considerações finais e limitações do trabalho.....	165
REFERÊNCIAS BIBLIOGRÁFICAS.....	170

ANEXOS.....	119
Anexo A - Tarefa IV (Azulejos Valiosos) da edição de 2015/2.....	179
Anexo B - Parâmetros de Avaliação da Parte Escrita.....	180
Anexo C - Tutorial de Digitações do Corpus de Textos de Examinandos do Celpe-Bras	181
Anexo D - <i>N-grams</i> do corpus n4 separados por grupos de informações.....	186
Anexo E - <i>N-grams</i> do corpus n3 separados por grupos de informações.....	188
Anexo F - <i>N-grams</i> do corpus n2 separados por grupos de informações.....	191
Anexo G - <i>N-grams</i> do corpus n1 separados por grupos de informações.....	194

CARTA ABERTA: Introdução

Brincar de profissões é algo comum entre crianças pequenas. Atravessar a bolsa do pai no peito pra brincar de carteiro, pendurar o estetoscópio da mãe no pescoço para brincar de médico, colocar o capacete do pai na cabeça pra brincar de engenheiro. No meu faz de conta infantil, antes mesmo de saber ler, eu destacava trechos de textos, que eu não sabia nem do que se tratavam, com marca texto amarelo. Minha mãe, pesquisadora e professora universitária da área de Letras, concluiu o mestrado e o doutorado quando eu era ainda criança, de modo que, na minha infância, a vi incontáveis vezes utilizando marca textos amarelos para destacar trechos de textos. Na época, minha mãe lecionava na Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), e me lembro de acompanhá-la em algumas de suas aulas. Gosto de brincar que ingressei cedo na academia.

Como não consigo determinar com precisão quando ocorreram os fatos no passado, sei que entre 10 a 15 anos depois prestei vestibular para Licenciatura em Letras com ênfase em Português e Alemão na Universidade Federal do Rio Grande do Sul (UFRGS), iniciando, oficialmente, minha jornada universitária, em 2017. No mesmo ano, comecei a frequentar os Seminários de Formação de Professores do Programa de Português para Estrangeiros (PPE). Nos seminários, tive meu primeiro contato com o exame para obtenção do Certificado de Língua Portuguesa para Estrangeiros (Celpe-Bras) por meio da análise em grupo de algumas produções escritas em resposta a tarefas do exame para fomentar a discussão sobre níveis de proficiência. O contato com o PPE me possibilitou ter minha primeira experiência como professora no Curso de Espanhol - Português para Intercâmbio (CEPI), coordenado, na época, pela Professora Dra. Gabriella Bulla. Nesta ocasião, pude preparar materiais voltados para o uso da língua em contexto acadêmico e interagir com estudantes de países de língua espanhola, que viriam fazer intercâmbio na UFRGS.

Ainda em 2017 entrei, como bolsista voluntária de Iniciação Científica (IC), no projeto de pesquisa intitulado “Exame Celpe-Bras: análise do acervo de provas já aplicadas, manuais, legislação e estudos realizados”, coordenado pela Professora Dra. Juliana Schoffen, que me orientou durante toda a minha trajetória acadêmica. Desde então, faço parte do Grupo Avalia - Avaliação de Uso da Linguagem¹, atuante no Instituto de Letras (IL) da UFRGS, onde conheci pesquisadoras e pesquisadores com os quais tive a oportunidade de adentrar o mundo da pesquisa acadêmica, participando de discussões teóricas, comparecendo a

¹ O grupo Avalia está cadastrado no Diretório dos Grupos de Pesquisa no Brasil do CNPq desde 2016. O site do grupo está disponível em: <https://www.ufrgs.br/grupoavalia/>. Acesso em: 31 de outubro de 2023.

congressos e sendo co-autora em produções acadêmicas ainda na graduação (Schoffen *et al.*, 2018; Nagasawa, 2019). Meu trabalho como bolsista de IC começou com a divulgação do número de examinandos por nível de certificação de todas as edições do Celpe-Bras entre 1999/1 e 2017/1, trabalho que consistiu na contagem de quantos examinandos foram avaliados em cada nível de proficiência nas listas divulgadas no Diário Oficial da União. Os resultados foram disponibilizados no Acervo de Provas e Documentos Públicos do Celpe-Bras (Acervo Celpe-Bras)², publicizado em 2014, um banco de dados que reúne todos os documentos oficiais sobre o exame produzidos pelo MEC e pelo Inep, bem como as provas já aplicadas, pesquisas realizadas e dados relacionados ao exame (Schoffen *et al.*, 2017). A existência do Acervo Celpe-Bras é fruto do projeto de pesquisa intitulado “Resgatando a história do Exame Celpe-Bras: desenvolvimento e análise de um banco de dados reunindo documentos públicos, provas aplicadas e estudos realizados sobre o Exame”, coordenado pela Professora Dra. Juliana Schoffen, uma iniciativa muito importante com objetivo de democratizar o acesso às informações relacionadas ao exame, tendo impacto positivo na vida de professores, pesquisadores, avaliadores e estudantes.

Ao longo desta trajetória, também fui aprendendo a ser professora. Concomitante à minha participação no grupo de pesquisa, tive experiências em sala de aula com a língua alemã, sendo professora-bolsista no Idiomas sem Fronteiras (IsF) (UFRGS) e, posteriormente, bolsista-ministrante no Núcleo de Ensino de Línguas em Extensão (Nele) (UFRGS), projeto com o qual mantenho vínculo até hoje. Estas duas práticas contínuas de preparação de materiais e contato com estudantes no ambiente universitário foram um grande incentivo para que eu me sentisse confiante na posição de docente. Ao final da graduação, participei como professora-bolsista em cursos de português para estrangeiros na Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), atuando em um contexto diferente de ensino, voltado para a área da saúde, e também auxiliando examinandos a se prepararem para fazer o Celpe-Bras. Ao terminar a graduação, tive a oportunidade, na UFCSPA, de trabalhar na aplicação do Celpe-Bras como avaliadora da Parte Oral, momento em que pude colocar em prática os conhecimentos que tinha adquirido como professora e pesquisadora nesta área, aprimorados no curso de Formação de Avaliadores da Parte Oral do Exame Celpe-Bras. Desde que finalizei a graduação, trabalhei tanto na aplicação do exame quanto na avaliação da Parte Oral e da Parte Escrita, participando, em cada edição, dos cursos de formação para a avaliação de ambas as partes.

² Disponível em: <https://www.ufrgs.br/acervocelpebras/acervo/>

Depois de ingressar no mestrado, lecionei em cursos preparatórios ao Celpe-Bras na UFRGS em projetos do PPE voltados para o ingresso de refugiados na universidade. Com o intuito de apresentar aos meus alunos exemplos autênticos de produções em resposta às tarefas do exame, conscientizando-os sobre o que é avaliado em cada nível, me vali de documentos oficiais, como o Manual do Examinando (Brasil, 2013a) e o Caderno de Tarefas Comentadas Edição 2019/2 (Brasil, 2020), e também preparei materiais baseados em resultados de pesquisas acadêmicas, como Mendel (2019), que analisou textos em resposta a duas tarefas de leitura e escrita na edição de 2015/2. Ter acesso a estas informações me permite preparar meus alunos de maneira mais adequada, pois posso possibilitar que eles se familiarizem com as práticas avaliativas reais do exame. Por este motivo, busco contribuir para que professores como eu possam ter acesso a cada vez mais informações relacionadas aos níveis de proficiência do exame Celpe-Bras, tópico que foi, ao longo dos anos, despertando meu interesse.

Além de Mendel (2019), muitos outros trabalhos se debruçaram sobre os níveis de proficiência escrita no Celpe-Bras (Sidi, 2002; Schoffen, 2009; Gomes, 2009; Evers, 2013; Dell’Isola; Tosatti, 2015; Sirianni, 2016; 2020; Kunrath, 2019; Divino, 2021; Tosatti, 2021; Vicentini, 2022; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023). Meu Trabalho de Conclusão de Curso (TCC) (Divino, 2021) teve o intuito de responder à pergunta “Quais índices lexicais de análise são relevantes para a caracterização de diferentes níveis de proficiência na Tarefa 4 da edição de 2015-2 do Celpe-Bras?” sob a luz da Linguística de Corpus (LC), compondo o pequenos grupo de trabalhos que olharam para os níveis de proficiência do Celpe-Bras também utilizando a LC (Kunrath, 2019; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023). No TCC, analisei textos nivelados com as notas 2 e 5 em resposta à Tarefa 4 da edição de 2015/2, intitulada “Azulejos Valiosos”.

Com o intuito de dar seguimento ao meu TCC, analisando, além dos textos nota 2 e 5, também os textos de notas 0, 1, 3 e 4, com intenção de contribuir para o aumento do número de trabalhos relacionados aos níveis de proficiência do Celpe-Bras, sobretudo daqueles que se utilizam da LC, busco, neste trabalho, responder à pergunta “*Como as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência em tarefas de leitura e escrita no exame Celpe-Bras?*”. Para responder a esta pergunta, este trabalho foi organizado em 6 capítulos, cujos nomes fazem alusão à tarefa que serviu de ponto de partida para a produção dos textos que deram origem ao corpus aqui analisado.

No capítulo **1. PATRIMÔNIO HISTÓRICO, ARTÍSTICO E CULTURAL: Celpe-Bras: elaboração, histórico, concepções teóricas e estrutura**, trago informações sobre o exame Celpe-Bras, iniciando com o contexto em que se encontrava a área do Português como Língua Adicional (PLA) no momento de sua elaboração. Abordo, neste capítulo, as especificações do exame, como seu alinhamento teórico e concepção de proficiência, além de sua estrutura, descrevendo a Parte Escrita e a Parte Oral, bem como o processo de avaliação. Para finalizar, são discutidos os conceitos de validade, confiabilidade, autenticidade, impacto e praticidade, termos fundamentais uma vez que este trabalho pode trazer resultados com potencial para impactar na validade e na confiabilidade do exame, bem como na preparação dos examinandos e formação de professores.

O capítulo **2. CASARÕES HISTÓRICOS: Pressupostos teóricos** tem como principal objetivo falar sobre a área de pesquisa interdisciplinar entre Avaliação de línguas e LC, denominada Avaliação de línguas informada por corpus (Nagasawa, 2023). Para isto, são apresentados pressupostos da LC, como as características que constituem um corpus e para que pode ser utilizado, abordagens de pesquisa e algumas técnicas de manuseio, de modo a já apresentar o *Sketch Engine (SkE)*, software de análise utilizado neste trabalho. Apresento, também, os diferentes tipos de corpora a partir da taxonomia proposta por Viana (2011), levantando a discussão acerca do termo “corpora de aprendizes” com base em Jenkins (2005) e Mauranen (2011), que discordam da utilização deste termo para generalizar falantes não nativos de uma língua³. As últimas duas seções deste capítulo tratam de corpora de línguas adicionais e sua relação com a avaliação de proficiência e das contribuições de corpora de PLA para o ensino e a avaliação. Por fim, para situar este trabalho neste campo interdisciplinar de pesquisa, são apresentados os resultados de trabalhos com foco na descrição dos níveis de proficiência do Celpe-Bras.

O capítulo **3. MEDIDAS IMEDIATAS: Procedimentos metodológicos** trata de explicar o aprofundamento da pergunta de pesquisa e dos objetivos desta. Depois disso, o corpus e a tarefa que originou seus textos são descritos, incluindo detalhes do acesso aos textos e do processo de compilação. Os procedimentos realizados em cada etapa das análises são descritos, sendo apresentadas as ferramentas *N-grams*, *Keywords*, *Wordlist* e *Concordance* do *SkE*, escolhidas para comporem as análises deste trabalho.

³ Ao longo desta dissertação, os termos “falante nativo” e “falante não nativo” são utilizados para manter a terminologia utilizada pelos autores que os usam. Ressalta-se que esses termos não refletem nosso posicionamento teórico.

No capítulo **4. AZULEJOS VALIOSOS: Análises dos dados**, são apresentados os resultados obtidos a partir dos procedimentos descritos no capítulo 3. Primeiro, são apresentados dados referentes à extensão dos subcorpora de cada nota e, depois, são descritas as análises realizadas com cada uma das ferramentas.

No capítulo **5. PROCESSOS DE TOMBAMENTO: Discussão dos resultados**, os resultados apresentados no capítulo 4 são discutidos e organizados de forma a responder todas as perguntas de pesquisa. A discussão envolveu comparações com os resultados obtidos por Sirianni (2016) e Mendel (2019), que analisaram a mesma tarefa aqui pesquisada.

Por fim, no capítulo **6. MERCADO DE AZULEJOS: Considerações finais e limitações do trabalho**, relaciono os resultados com o aporte teórico apresentado no trabalho, apontando algumas das limitações desta pesquisa. Proponho, também, algumas sugestões de estudos que podem ser realizados a partir do que foi descoberto, com o objetivo de incentivar que mais pesquisas sejam realizadas nesta área.

1. PATRIMÔNIO HISTÓRICO, ARTÍSTICO E CULTURAL: Celpe-Bras: elaboração, histórico, concepções teóricas e estrutura

Neste capítulo, será fornecido um breve histórico da área de Português como Língua Adicional (PLA)⁴ no Brasil até o surgimento do exame Celpe-Bras, justificando sua criação. Além disso, será apresentado um panorama histórico do Exame desde sua concepção até os dias atuais, abordando seu construto teórico, estrutura e evolução ao longo do tempo.

1.1 Contexto de elaboração do Celpe-Bras

A área de PLA começou a se consolidar no Brasil mais fortemente a partir da década de 1960, momento em que vários cursos de português começaram a ser ofertados nos Estados Unidos e quando houve parcerias com consultores brasileiros para a elaboração de livros didáticos⁵ (Schlatter; Bulla; Costa, 2020). O marco inicial na área de PLA no país é ainda anterior a isto, em 1942, com a publicação do primeiro livro didático de português brasileiro para estrangeiros, “A Língua Portuguesa para estrangeiros”, escrito por Hermine Topker, seguido por Mercedes Merchant, que, em 1957, publicou já o segundo livro neste contexto, intitulado “Português para Estrangeiros” (Fernandes, 2020). A partir da década de 1970, universidades do sudeste⁶ começaram a ofertar cursos de PLA e, nas duas décadas seguintes, “foram criados vários programas de PLA [em diversas universidades espalhadas pelo país⁷], com o objetivo de ofertar cursos, formar professores e desenvolver pesquisa na área” (Schlatter; Bulla; Costa, 2020, p. 491). O cenário para o ensino do português brasileiro era bastante otimista. A necessidade de formar professores e de realizar pesquisas na área de PLA, bem como de elaborar materiais didáticos, surgiu como consequência de parcerias com universidades no exterior, que expandiam seus programas de intercâmbio com o Brasil.

⁴ Neste trabalho, utiliza-se o termo Língua Adicional por não fazer “referência a categorizações aprioristas sobre as relações afetivas entre o estudante e a língua, o contexto e as condições de aprendizagem da língua ou os objetivos de aprendizagem” (Bulla; Kuhn, 2020, p. 7). O termo *adicional* implica a ideia de ampliação das possibilidades que um indivíduo tem de participação, entendendo que uma nova língua é “adicionada” ao seu repertório linguístico (Schlatter; Garcez, 2009).

⁵ “Português Contemporâneo” (Rameh; Abreu, 1967) e “Modern Portuguese: a Project of the Modern Language Association” (Ellison, 1971) (Schlatter, 2020, p. 490 apud Matos, 1995, p. 9).

⁶ Universidade de São Paulo (USP) e Universidade do Estado de Campinas (Unicamp).

⁷ Universidade Federal de Pernambuco (UFPE), Universidade Federal Fluminense (UFF), Universidade Federal do Rio Grande do Sul (UFRGS), Universidade de Brasília (UnB), Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal de Minas Gerais (UFMG) e Universidade Federal do Paraná (UFPR).

O Brasil direcionava, na década de 90, esforços para estimular o aprendizado do português. Isso ocorria por meio da manutenção da Rede Brasil Cultural⁸, da disponibilização de cursos de língua portuguesa para estrangeiros em diversas instituições nacionais e do desenvolvimento de programas estruturados para a promoção da cultura nacional (Dell’Isola *et al.*, 2003). Foram realizadas inúmeras ações para valorizar o ensino da língua e a presença da cultura brasileira, tanto no território nacional quanto no exterior, como, por exemplo, a criação da Sociedade Internacional de Português Língua Estrangeira (SIPLE) em 1992 (Schlatter; Bulla; Costa, 2020). Além disso, a publicação de artigos acadêmicos “abordando conhecimentos teórico-práticos sobre ensinar PLA, programas e metodologias de ensino, materiais didáticos e aprendizagem de diferentes aspectos do português” (Schlatter; Bulla; Costa, 2020, p. 491) teve início nessa época, juntamente com a publicação de livros didáticos⁹ (Schlatter; Bulla; Costa, 2020). Os movimentos iniciais de internacionalização nas universidades brasileiras foram um dos motores que impulsionaram a institucionalização da área de PLA no meio acadêmico, com protagonismo de professores universitários de Português como Língua Materna (PLM), Inglês e outras línguas estrangeiras (Bulla; Kuhn, 2020).

Em meio à crescente difusão e promoção do português brasileiro e, como reação à necessidade da criação de um exame padronizado para ser aplicado para estudantes estrangeiros interessados em estudar nas universidades brasileiras (Schlatter; Scaramucci, 2021), foi desenvolvido o exame Celpe-Bras (Dell’Isola *et al.*, 2003). Em junho de 1993 foi instituída, pelo Ministério da Educação (MEC), uma Comissão para a Elaboração do Exame de Proficiência de Português para Estrangeiros (Schlatter, 2014; Schlatter; Bulla; Costa, 2020) e, em 1994, foi constituída a Comissão Permanente (Schlatter, 2014; Schlatter; Bulla; Costa, 2020), com representantes de universidades de diversas partes do Brasil, “para concluir a elaboração do Exame de Português para Estrangeiros, administrar a aplicação, credenciar instituições para a aplicação do exame e assessorar a SESu [Secretaria de Educação Superior] nas questões relativas ao ensino de português para estrangeiros” (Schlatter, 2014, n.p.). As reuniões foram, desde o início, realizadas nas universidades em que atuavam os membros da

⁸ A Rede Brasil Cultural é formada por Centros de Estudos Brasileiros, Leitorados e Núcleos de Estudos, configurando-se como “um instrumento do Ministério das Relações Exteriores para a promoção da língua portuguesa e da cultura” (Schlatter *et al.*, 2020, p. 493 apud Rede Brasil Cultural, s.d.)

⁹ “Falando... Lendo... Escrevendo...” (Lima; Iunes, 1981), “Português para Falantes de Espanhol (Edição Experimental)” (Lombello; Baleeiro, 1983), “Tudo Bem” (Ramalhete, 1984), “Avenida Brasil: Curso básico para estrangeiros” (Lima, 1991); “De Conversa em Conversa. Aprendendo o Português do Brasil I e II” (Santos, 1994; 1995) (Schlatter *et al.*, 2020).

comissão e em algumas outras que, ou ofereciam ensino de Português como Língua Estrangeira (PLE), ou tinham interesse em oferecer (Schlatter; Scaramucci, 2021). O exame foi projetado para ser uma ação de política linguística para o incentivo à promoção da área de português para falantes de outras línguas e, por esse motivo, foi interessante que a organização fosse concentrada em universidades que “seriam potenciais centros de referência de português para estrangeiros e postos aplicadores do exame” (Schlatter; Scaramucci, 2021, p. 120).

Das tarefas delegadas à comissão, além da própria criação do exame, estava a de realizar a “testagem de uma versão piloto, seguida de análise e tratamento estatísticos para uma validação inicial, e [a] sistematização de procedimentos necessários para a validade e a confiabilidade da avaliação” (Schlatter; Scaramucci, 2021, p. 120). Foram deliberadas também questões relacionadas às especificações do exame, características das tarefas da Parte Escrita, procedimentos para as interações da Parte Oral (PO), critérios de avaliação com seus respectivos descritores, perfil e formação dos avaliadores, critérios para credenciamento de centros aplicadores, e a logística necessária para a realização do exame, tanto no Brasil, quanto no exterior (Schlatter; Scaramucci, 2021).

A criação de um exame de proficiência em português brasileiro situava-o como um idioma importante e estratégico para a comunicação no cenário mundial, “reforça[ndo] as estruturas existentes e estabiliza[ndo] o corpo docente de todos os centros difusores da língua portuguesa, bem como promove[ndo] a possibilidade de criação de cursos específicos de preparação para a realização do exame” (Dell’Isola *et al.*, 2003, p. 155). Neste sentido, algumas das ações que possibilitaram o fortalecimento entre programas de intercâmbio e mobilidade já existentes e ações políticas de fomento à língua portuguesa e à cultura brasileira foram: o credenciamento, desde o primeiro ano de sua aplicação, de instituições parte da Rede Brasil Cultural como postos aplicadores do Celpe-Bras; e a exigência do Exame, a partir de 1999, para candidatos aos Programas de Estudante Convênio de Graduação (PEC-G)¹⁰ e Pós-Graduação (PEC-PG)¹¹ (Schlatter; Bulla; Costa, 2020). Desde o início do

¹⁰ “O PEC-G destina-se a estudantes de países em desenvolvimento com os quais o Brasil mantém acordo de cooperação cultural e/ou educacional e/ou ciência e tecnologia, que possibilita ao estudante desses países estudar nas Instituições de Ensino Superior (IES) brasileiras” (Carvalho; Schlatter, 2011, p. 267).

¹¹ “Por meio do PEC-PG, são oferecidas anualmente vagas em cursos de pós-graduação (Mestrado e Doutorado) em IES brasileiras (...). Dentre os objetivos do Programa estão cooperar na formação avançada de nações em desenvolvimento e enriquecer o conhecimento de professores e estudantes brasileiros de alto nível com as experiências acadêmicas, culturais e de vivência de alunos estrangeiros” (Carvalho; Schlatter, 2011, p. 268).

desenvolvimento do sistema de avaliação Celpe-Bras, o MEC¹² tinha o objetivo de integrar a elaboração e aplicação do exame com a criação de uma rede de ensino de português para estrangeiros, que teria como finalidade preparar os candidatos ao PEC-G para sua admissão nas universidades brasileiras (Schlatter; Scaramucci, 2021)¹³.

O Celpe-Bras teve sua primeira aplicação anos depois de sua concepção, em 1998, com 127 examinandos em 8 Postos Aplicadores (Brasil, 2020) no Brasil - nas cinco universidades representadas na Comissão Técnica na época (UFRGS, Unicamp, UFRJ, UFPE e UnB) - e no exterior - em instituições de países pertencentes ao MERCOSUL (Instituto Cultural Uruguaio-Brasileiro, no Uruguai, Centro de Estudos Brasileiros de Assunção, no Paraguai, e Fundação Centro de Estudos Brasileiros de Buenos Aires, na Argentina) (Carvalho; Schlatter, 2011). Inicialmente, com apenas dois níveis (Primeiro Certificado e Segundo Certificado) (Scaramucci, 2000), o Exame parecia “atender às exigências e necessidades de certificação do público-alvo, composto majoritariamente de alunos estrangeiros que vinham estudar em nossas universidades” (Scaramucci; Santos Junior; Diniz, 2019, p. 168). A necessidade de uma certificação baseada em um nível de proficiência que evidenciasse habilidades para a realização de estudos universitários foi um dos fatores que influenciaram a decisão de criar um único exame, composto por duas etapas distintas: uma Parte Escrita (PE) e uma Parte Oral (PO) (Schlatter; Scaramucci, 2021).

Atualmente, o Celpe-Bras é aplicado em mais de 130 Postos Aplicadores no Brasil e no exterior, “que reúnem universidades, institutos federais, representações diplomáticas, missões consulares e outras instituições que oferecem programas de ensino de português” (Schlatter; Scaramucci, 2021, p. 119) contabilizando 42 países nos continentes americano, africano, europeu e asiático (Schlatter; Scaramucci, 2021; Nagasawa, 2023). Ao levar em conta que a condição para se tornar um posto aplicador é comprovar a oferta de cursos na instituição, o aumento contínuo do número de postos aplicadores ao longo dos anos indica uma expansão na área e na rede de profissionais, essa rede agora conecta iniciativas em todos os cinco continentes e compartilha um padrão de referência comum (Schlatter; Scaramucci, 2021).

¹² A partir de 2009, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) passou a ser responsável pelo exame (Portaria nº 856, de 04/09/2009) (Schlatter; Scaramucci, 2021).

¹³ Desde o início de sua aplicação até 2013, o Celpe-Bras era exigido como comprovação de proficiência apenas a candidatos “não-lusófonos” ao PEC-G. Sua exigência foi estendida, por meio do Decreto Nº 7.948, de 12 de março de 2013, a todos os estudantes estrangeiros, incluindo aqueles que vêm dos Países Africanos de Língua Oficial Portuguesa (PALOP) (Brasil, 2013b). Esta dissertação se debruça sobre os textos da segunda edição de 2015, de modo que há probabilidade de haver, no corpus, textos escritos por examinandos oriundos destes países.

Figura 1 - Número de examinandos anuais do Exame Celpe-Bras

Fonte: Nagasawa, 2019.

O crescimento no número de interessados em comprovar seu nível de proficiência em português por ano (Figura 1) é evidente, tendo tido, em 2019, 11.266 examinandos homologados, colocando o Celpe-Bras como um exame de alta relevância, pois seus resultados têm impacto na vida das pessoas que o realizam (Schlatter; Scaramucci, 2021).

1.2 Especificações do exame Celpe-Bras

A definição das especificações de um exame são indispensáveis para seu processo de desenvolvimento e validação.

Se quisermos usar as pontuações de um exame de língua para fazer inferências sobre a capacidade linguística de um indivíduo e, possivelmente, para tomar várias decisões, devemos ser capazes de demonstrar como o desempenho no exame está relacionado ao uso do idioma em situações específicas que não o idioma em si. Para podermos demonstrar essa relação, precisamos de uma estrutura conceitual que nos permita tratar o desempenho em um exame como uma instância específica do uso da língua. Ou seja, precisamos de uma estrutura que nos permita usar as mesmas características para descrever o que acreditamos ser as características críticas, tanto

do desempenho em um exame, quanto do uso do idioma fora dele (Bachman; Palmer, 1996, p. 10).¹⁴

Douglas (2000) descreve, em uma lista de oito componentes, as principais características que deve ter um exame de língua:

1. **Descrição do propósito do exame**, explicitando as decisões tomadas a respeito das inferências sobre habilidade linguística e capacidade de uso da língua;
2. **Descrição das situações e tarefas futuras de uso da língua**, determinando a esfera de uso da língua sobre a qual se deseja fazer inferências, bem como que usos serão feitos nessa esfera;
3. **Descrição das características dos falantes/examinandos**, precisando a população para qual o exame é/foi elaborado;
4. **Definição de um construto a ser medido**, tornando explícita a natureza da habilidade que se quer medir, providenciando uma descrição de quais aspectos particulares de habilidade e performance linguística se tem o interesse em fazer inferências sobre;
5. **Descrição do conteúdo do exame**, definindo que tipos de tarefa serão incluídos no exame com base na definição do construto;
6. **Descrição dos critérios de correção**, providenciando uma descrição de como as respostas serão avaliadas com base em uma escala, bem como como a pontuação será calculada;
7. **Apresentação de exemplos de tarefas**, demonstrando os tipos de tarefa que poderão ser solicitadas;
8. **Desenvolvimento de um plano para avaliar as qualidades de uma boa avaliação**, incluindo validade, confiabilidade, autenticidade, impacto e praticidade.

Alinhadas aos 8 componentes descritos por Douglas (2000), serão apresentadas, a seguir, as especificações do exame Celpe-Bras.

¹⁴ No original: *If we want to use the scores from language test to make inferences about individual's language ability, and possibly to make various types of decisions, we must be able to demonstrate how performance on the language test is related to language use in specific situations other than the language itself. In order to be able to demonstrate this relationship, we need a conceptual framework that enables us to treat performance on a language test as a particular instance of language use. That is, we need a framework that enables us to use the same characteristics to describe what we believe are the critical features of both language test performance and non-test language use* (Bachman; Palmer, 1996, p. 10)

1.2.1 A concepção teórica do exame

Como dito anteriormente, para a estruturação de uma avaliação de proficiência, é importante que o desempenho exigido nessa avaliação e o uso real que se vai fazer da língua fora do contexto avaliativo sejam correspondentes (Bachman; Palmer, 1996). E, mais do que isso, “qualquer avaliação de língua deve ser desenvolvida com um propósito específico, para um grupo específico de examinandos e para um contexto específico de uso da língua” (Bachman; Palmer, 1996, p. 18)¹⁵, ou seja, o reconhecimento das particularidades dos usuários ou potenciais usuários de uma avaliação é um ponto crucial para a sua concepção (Bachman; Palmer, 1996). No que corresponde à **1. Descrição do propósito do exame**, o Celpe-Bras foi elaborado, para aferir “o potencial [dos examinandos] para ler, escrever, ouvir e falar em interações da vida cotidiana” (Schlatter, 2014, s.n.), tendo como objetivo avaliar a proficiência através da realização de tarefas que simulem situações reais de comunicação, isto é, através de tarefas autênticas, que compartilham características com as tarefas que serão realizadas em situações reais de uso (Douglas, 2000).

Aprofundando a **3. Descrição das características dos falantes/examinandos**, o público-alvo pretendido eram os alunos estrangeiros que vinham às universidades brasileiras para estudar e chegavam aqui com um nível baixo de proficiência em português, o que afetava seu desempenho acadêmico (Brasil, 2020). Além de sua função principal de certificar a proficiência, o objetivo era que o exame também desempenhasse um papel transformador no ensino do português, promovendo parâmetros para o uso da língua, em vez de apenas conhecimentos sobre a língua (Schlatter; Scaramucci, 2021). O Celpe-Bras foi desenvolvido, portanto, com intenção de criar um instrumento que pudesse direcionar o ensino do português no Brasil e no exterior (Scaramucci, 2008) para “práticas mais contemporâneas, com foco no uso da língua” (Scaramucci; Santos Junior; Diniz, 2019, p. 105).

A partir dos objetivos acima descritos, o exame foi pensado de tal forma que: (a) a competência do examinando fosse verificada por meio de tarefas relacionadas a interações cotidianas; (b) se buscasse aferir a capacidade de uso da língua, dando atenção especial às competências linguísticas do examinando; (c) fossem levados em conta, na avaliação, aspectos socioculturais também; (d) todos os materiais utilizados no exame fossem autênticos e contextualizados; (e) a avaliação fosse qualitativa e holística, levando em conta o desempenho na realização da tarefa como um todo; (f) os resultados fossem expressos em

¹⁵ No original: *[...] any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain [...]* (Bachman; Palmer, 1996, p. 18).

descritores de competência e desempenho (Dell’Isola *et al.*, 2003); e (g) “os parâmetros de correção [tivessem] como base os próprios objetivos das tarefas e os recursos discursivos exigidos para sua realização” (Schlatter, 2014, n.s.).

Uma das justificativas para a implementação do CELPE-Bras foi o efeito potencial que esse exame poderia ter como instrumento redirecionador do ensino de PLE no Brasil e no exterior, sinalizando não apenas a importância em se estudar português mas, principalmente, como deveria ser essa proficiência (Scaramucci, 2000), que aspectos deveriam ser considerados ou que habilidades, competências e conteúdos são importantes para o candidato fazer frente às situações de uso da língua. O exame, dessa forma, forneceria aos candidatos, instituições e professores responsáveis pela preparação de candidatos elementos e informações sobre o conteúdo e objetivos que devem ser priorizados nesse ensino, ressaltando, principalmente, a necessidade de reverem sua abordagem, sua visão de linguagem e de aprender línguas, ou, em outras palavras, seu **conceito de proficiência** (Dell’Isola *et al.*, 2003, p. 162, grifo nosso).

Na época da elaboração, a comissão responsável pensou que seria relevante propor uma avaliação que verificasse a capacidade dos examinandos de “usar adequadamente a língua para desempenhar ações no mundo” (Brasil, 2020, p. 29), e, principalmente, de se envolver em práticas sociais no ambiente universitário brasileiro como uma tentativa de aumentar a probabilidade de que os examinandos realizassem as tarefas do exame da mesma forma que a tarefa seria realizada na situação-alvo real (Douglas, 2000), sendo esta a **2. Descrição das situações e tarefas futuras de uso da língua**. Levando em consideração o principal público-alvo - estudantes e/ou futuros estudantes universitários - o exame foi concebido para abranger habilidades de leitura, escrita, compreensão e expressão oral relacionadas às práticas do dia a dia e acadêmicas (Schlatter; Scaramucci 2021). Dois conjuntos de características de interesse central no que diz respeito ao uso da língua e ao desempenho no exame são, primeiro, as características específicas dos examinandos, relevantes para a validade de qualquer inferência que se faça sobre o desempenho no uso da língua, e as características das tarefas, relevantes para dimensionar o alcance da generalização dessas inferências (Bachman; Palmer, 1996).

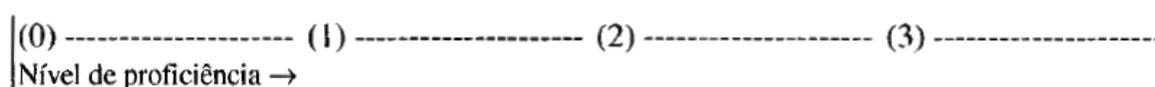
1.2.1.1 Gêneros do discurso e a noção de proficiência

Exames são avaliados e interpretados de acordo com as expectativas do que os examinandos devem conseguir desempenhar em determinado nível de proficiência (Callies; Götz, 2015). Segundo Bachman e Palmer (1996), “O principal objetivo de um exame de língua é fornecer uma *medida* que possa ser interpretada como um indicador da habilidade

linguística de um indivíduo” (Bachman; Palmer, 1996, p. 23)¹⁶ e, para isso, é preciso que esse exame seja pensado a partir de uma base conceitual, que sirva de pilar para sua elaboração, aplicação e avaliação. Essa base conceitual é denominada construto teórico, um conjunto de definições específicas que incorpora a visão dos elaboradores sobre como abordar a avaliação de proficiência linguística (Barker, 2010) e fornece subsídios para se ter com clareza o que está sendo avaliado e de que forma os resultados serão interpretados (Bachman; Palmer, 1996). O construto teórico de um exame não pode ser observado diretamente, sendo operacionalizado em itens ou tarefas (Brasil, 2020).

No que tange a **4. Definição de um construto a ser medido**, há o conceito de proficiência que, de acordo com Scaramucci (2000), pode ser definida como um compromisso com o construto teórico, que depende das especificidades de cada contexto comunicativo. A autora conceitualiza proficiência como sendo “relativa”, pois além de ser contextualizada, incluindo aspectos relacionados ao propósito da situação de uso, é representada como uma graduação, não havendo um ponto de corte entre **proficiência** e **não proficiência**. Nessa acepção mais ampla, ilustrada na Figura 2, que foge à ideia de uma proficiência “única, absoluta [e] monolítica” (Scaramucci, 2000, p. 14), os falantes são considerados mais ou menos proficientes a depender do seu desempenho para se adequar a determinado contexto de comunicação (Scaramucci, 2000).

Figura 2 - Escala de proficiência



Fonte: Scaramucci, 2000.

A proficiência, sendo definida como “a capacidade de participação em diversas práticas em diferentes contextos de uso da língua portuguesa” (Brasil, 2020, p. 31), só pode ser avaliada adequadamente por um exame de desempenho, que implica colocar o examinando em situações nas quais a proficiência pode ser demonstrada de maneira direta (Brasil, 2020).

¹⁶ No original: *The primary purpose of a language test is to provide a measure that we can interpret as an indicator of an individual's language ability* (Bachman; Palmer, 1996, p. 23).

A noção de proficiência adotada no exame Celpe-Bras pressupõe o uso da língua em uma relação dialógica (Bakhtin, 2003) entre falantes, pensada como “uma relação (de sentido) que se estabelece entre enunciados na comunicação verbal” por meio de práticas sociais, e esta relação “só é possível entre enunciados concluídos, proferidos por sujeitos falantes distintos” (Bakhtin, 2003, p. 345). Ao participar de práticas sociais, as pessoas utilizam repertórios de significados, conhecimentos, valores, comportamentos e visões de mundo construídos a partir de suas experiências prévias em interação com o contexto social e histórico atual (Schlatter; Scaramucci, 2021) ou seja, com a cultura.

Interagir com aspectos culturais no exame Celpe-Bras significa, portanto, estar atento a diferentes pontos de vista sobre como o mundo pode ser compreendido e atribuir sentidos considerando as ações dos participantes e seus usos da linguagem nas práticas sociais propostas (Schlatter; Scaramucci, 2021, p. 124/125).

De acordo com Bakhtin (2003), “a língua penetra na vida através de enunciados concretos que a realizam, e é também através dos enunciados concretos que a vida penetra na língua” (Bakhtin, 2003, p. 282). As práticas sociais estão inseridas em diferentes esferas da atividade humana, sempre relacionadas à utilização da língua (Bakhtin, 2003), que materializa as ações dos participantes. Essa materialização se dá por meio de enunciados concretos, que “reflete[m] as condições específicas e as finalidades de cada uma dessas esferas [de atividade humana]” (Bakhtin, 2003, p. 279), sendo caracterizado por seu conteúdo, seus recursos lexicais, fraseológicos e gramaticais, e sua construção composicional que, juntos, “são marcados pela especificidade de uma esfera da comunicação” (Bakhtin, 2003, p. 279). Partindo desta perspectiva dialógica de língua, que se dá por meio de enunciados concretos e compartilhados em uma comunidade/cultura, “a coletividade linguística [e] a multiplicidade de interlocutores são fatos que não podem ser ignorados” (Bakhtin, 2003, p. 290), ou seja, é dada importância à relação entre os participantes da interação, visto que cada indivíduo é social e historicamente situado em relação a outro. Bakhtin (2003) não vê o receptor/ouvinte como um sujeito passivo na interação e entende a percepção do enunciado pelo receptor/ouvinte como uma “atitude *responsiva ativa*” (Bakhtin, 2003, p. 290), que nada mais é do que a “fase inicial e preparatórios para uma resposta” (Bakhtin, 2003, p. 291), que pode ser verbal ou não. Para Dell’Isola e Pordeus (2021), o texto é entendido como “um território onde se materializam, em situações de uso e de acordo com um dado propósito comunicativo, processos linguísticos e estratégias comunicativas”, em que o leitor constrói, nesse espaço de

interação, significados, “atuando como coautor nesse processo” (Dell’Isola; Pordeus, 2021, p. 76).

Cada enunciado é, segundo Bakhtin (2003), individual e único, porém, as diferentes esferas de utilização da língua apresentam tipos relativamente estáveis de enunciados, denominados por Bakhtin como **gêneros do discurso**, através dos quais materializam-se as práticas de linguagem (Dell’Isola; Pordeus, 2021). Sob a perspectiva bakhtiniana, afirma Schoffen (2009), o gênero é parte do processo que envolve o “somatório das relações dialógicas estabelecidas em um determinado contexto de comunicação, em que se relacionam o falante, o ouvinte, o propósito do enunciado e a esfera de atividade humana em que a comunicação ocorre” (Schoffen, 2009, p. 91), e não somente o produto final. Os gêneros são, portanto, entendidos como “instrumentos de mediação de uma situação comunicativa, podendo tanto determiná-la bem como estabelecer o comportamento dos envolvidos” (Dell’Isola; Pordeus, 2017). No Celpe-Bras, a integração desses componentes se dá por meio de tarefas, caracterizadas como um “convite para interagir com o mundo, usando a linguagem com um propósito social” (Brasil. MEC, 2006), operacionalização que é consistente com a definição de proficiência proposta, visto que o examinando é exigido a demonstrar diferentes habilidades no uso da língua a depender da situação apresentada (Brasil, 2020).

A tarefa pressupõe a realização, por meio da língua, de uma ação, materializada em um texto escrito, cuja estrutura, organização e convenções são de ordem sociocomunicativa. Assim, uma tarefa determina uma ação com um propósito claro de comunicação – planejada por um enunciador e direcionada a um ou mais interlocutores –, que deve orientar a produção de um determinado gênero discursivo por parte do participante (Brasil, 2020, p. 32).

Pode-se afirmar que as práticas sociais de uso da língua propostas no exame Celpe-Bras são, portanto, situadas e contextualizadas, baseada em propósitos sociais, em que a construção do sentido é “construída social e localmente por seus participantes” (Brasil, 2020, p. 28), e a leitura e compreensão de um texto são entendidas como uma atividade colaborativa (Dell’Isola; Pordeus, 2021). Esta atividade envolve “a identificação de informações de base textual, geração de hipóteses e inferências, a partir das competências e habilidades que o leitor traz para esse processo” (Dell’Isola; Pordeus, 2021, p. 76).

1.2.2 Estrutura do exame

Como mencionado anteriormente, o Celpe-Bras é composto de uma PE e uma PO. Para receber uma certificação, o examinando precisa completar ambas as partes

satisfatoriamente, não havendo certificações específicas para alguma das partes de maneira independente. No Documento Base do Exame Celpe-Bras (Brasil, 2020), documento oficial destinado a apresentar o construto do exame e sua operacionalização, incluindo a **5. Descrição do conteúdo do exame**, há um esclarecimento detalhado de como são implementadas ambas as partes, qual sua duração e demais especificações a respeito das habilidades avaliadas e dos textos de insumo fornecidos.

1.2.2.1 Parte Escrita

A PE é a primeira parte da avaliação e ocorre no primeiro dia do exame, com horário padronizado, iniciando e finalizando no mesmo horário para todos os examinandos de um mesmo posto¹⁷¹⁸. A PE do Celpe-Bras tem duração de três horas, conforme apresentado no Quadro 1, e é composta de quatro tarefas que integram compreensão (oral, imagética e/ou escrita) e produção escrita. A Tarefa 1 tem como material de insumo um vídeo, que pode ser de distintos gêneros, como “[r]eportagens, entrevistas, documentários, esquetes cênicas, vídeos publicitários, entre outros” (BRASIL, 20202, p. 58) que circularam previamente na mídia brasileira. O vídeo é exibido duas vezes, em sequência, e o examinando tem 30 minutos para a realização dessa tarefa. Após passados os 30 minutos, os examinandos são convidados a ler o enunciado da Tarefa 2 e, então, é reproduzido o áudio, texto de insumo da Tarefa 2, também duas vezes, em sequência. Assim que o áudio termina, os examinandos têm o restante do tempo para a organização das leituras das Tarefas 3 e 4 e produção escrita dos quatro textos. Os primeiros 30 minutos são direcionados para a Tarefa 1 apenas por uma questão de organização, mas o examinando pode retornar a este texto em qualquer momento na prova. Assim como no caso dos vídeos, os áudios também provém de fontes públicas de circulação midiática brasileiras, com gêneros comuns a quem acessa esses meios de informação.

Quanto à Tarefa 3 e à 4, ambas têm como material de insumo textos escritos, retirados de “[j]ornais, revistas, panfletos, informativos, livros, redes sociais, mídia eletrônica em geral, entre outros” (Brasil, 2020, p. 58), que abarca gêneros como

[r]eportagens, entrevistas, artigos de opinião, editoriais, crônicas, colunas assinadas, notícias, informativos, panfletos, propagandas, cartuns, quadrinhos, anúncios,

¹⁷ Por questões de logística e, a depender do tamanho do Posto Aplicador, o exame pode iniciar com alguns minutos de diferença em locais diferentes.

¹⁸ No ato da inscrição, o examinando que necessitar de atendimento especializado pode solicitar no ato da inscrição e, a depender do caso, poderá ter direito ao tempo adicional de 60 minutos, previsto no edital (Brasil, 2023).

poesias, resenhas, relatórios, currículos, biografias, editais, formulários, questionários, instruções, mapas, roteiros, cardápios, letras de canções, manifestações, cartas do leitor, entre outros (Brasil, 2020, p. 58).

Quadro 1 - Estrutura da Parte Escrita

Tarefas	Habilidades envolvidas	Tempo total
1	Compreensão oral e imagética (vídeo) + produção escrita	3h
2	Compreensão oral (áudio) + produção escrita	
3	Leitura + produção escrita	
4	Leitura + produção escrita	

Fonte: Brasil, 2020.

Não há, nos editais recentes, qualquer tipo de informação a respeito da idade mínima ou grau mínimo de escolarização para a realização da prova, porém as temáticas e os textos utilizados no exame são voltados para um público jovem e adulto escolarizado (Schlatter; Scaramucci, 2021). Em estudo descritivo da PE, realizado por Schoffen *et al.* (2018), chegou-se a algumas conclusões a respeito das especificações da PE, posteriormente incorporadas ao Documento Base, apresentadas no Quadro 2, que serviram para se ter mais detalhes a respeito do que é solicitado dos examinandos nesta parte do exame. Isto, todavia, não significa que não há possibilidade de ser solicitada uma produção inserida em um novo eixo temático, com novos propósitos materializados em um gênero do discurso diferente, com uma relação de interlocução distinta do que já aparece em edições anteriores. No entanto, este mapeamento permitiu que se tivesse uma descrição mais detalhada das tarefas da PE, tendo mais conhecimento das suas especificações.

Cada tarefa do Celpe-Bras apresenta um enunciado, que indica as condições de produção específicas daquela tarefa, ou seja, o contexto, que não é simplesmente um conjunto de características impostas ao falante, mas é construído pelos participantes do evento comunicativo (Douglas, 2000). Para que o examinando possa adequar sua produção ao contexto de comunicação esperado, são determinados, no enunciado, “um propósito claro de

comunicação (escrever um texto para reclamar, informar, discordar etc.), um enunciador (morador de um determinado bairro, gerente de uma empresa, internauta etc.) e um ou mais interlocutores (leitores de um jornal, o chefe, o prefeito da cidade etc.)” (Brasil, 2020, p. 36). Para se comunicar da maneira desejada, o examinando precisa saber o que é esperado dele, o que está acontecendo, onde ele está, com quem ele está se comunicando, qual a sua função e qual é o assunto (Douglas, 2000). Assim, o examinando tem em mãos o que precisa para saber que informações dos textos de insumo são relevantes para o cumprimento da tarefa, de forma a poder fazer a leitura desses textos de maneira objetiva e direcionada. Se as informações dadas a ele forem suficientes para que ele compreenda o contexto desejado, há uma probabilidade maior das interpretações do desempenho como um indicador de habilidade de uso da língua serem mais válidas (Douglas, 2000).

Quadro 2 - Especificações da Parte Escrita

Características dos textos solicitados	
Gêneros dos textos solicitados	Abaixo-assinado, anúncio de emprego, apresentação, artigo, carta, e-mail, artigo de opinião, capítulo de livro, carta do leitor, convite, depoimento, diário de viagem, editorial, guia de orientações, item de catálogo, notícia panfleto, propaganda, texto para quadro de avisos, crônica, carta aberta, carta de apresentação, roteiro turístico, instruções, relatório, resumo, seção de guia, entre outros.
Relação de interlocução nos textos solicitados	Cidadão para órgão do governo, órgão do governo para a população, empresa para clientes, cliente para empresa, funcionário para gestor, gestor para funcionário, funcionário para colega de trabalho, amigo para amigo, morador para vizinhos, leitor/ouvinte para veículo de comunicação, empresa para consumidores, membro de associação para empresas, integrante de comunidade educacional para comunidade educacional, colonista para leitores, empresa para candidato a emprego, instituição para a população, indivíduo para profissional, integrante de grupo social para seus pares, organização para associados, entre outras.
Propósitos dos textos solicitados	Aconselhar, apresentar, divulgar, incentivar, indicar, informar, orientar, posicionar-se, propor, recomendar, relatar experiência, solicitar, sugerir, entre outros.
Temáticas envolvidas	
Eixos temáticos avaliados	Ambiente, ciência e tecnologia, consumo, esporte, estilos de vida, educação, comunicação, mundo do trabalho, habitação, transporte, serviços, compras, alimentação, saúde, lazer, mídia, arte, ambiente, patrimônio cultural, políticas e cidadania, turismo, entre outros.*

Fonte: Brasil (2020) com base em Schoffel *et al.* (2018).

1.2.2.1.1 Integração de habilidades

Uma importante etapa no desenvolvimento das tarefas contempladas em um exame é, além da definição do propósito da avaliação e do construto a ser avaliado, a especificação das “características do insumo, da resposta, e da **relação entre insumo e resposta**” (Bachman; Palmer, 1996, p. 173, grifo nosso). No Celpe-Bras, esta relação é direta: cada tarefa da PE é elaborada com base em um material de insumo, prevendo uma avaliação que integra as habilidades de leitura ou compreensão oral e produção escrita (Mendel; Schoffen, 2017), como acontece em situações cotidianas, em que “ler um texto para discuti-lo ou ainda para escrever outro, assistir a uma palestra e fazer anotações são práticas comuns” (Brasil, 2020, p.

32). Os propósitos de leitura são determinados pelos propósitos de escrita (Scaramucci, 2016), ou seja, reafirmam a “noção de responsividade [proposta por Bakhtin (2003)], na medida em que condicionam a produção do examinando a uma compreensão ativamente responsiva perante os textos oferecidos como material de insumo” (Mendel, 2019, p. 47).

O enunciado instrui, por meio do que Mendel (2019) denomina como *pistas contextuais*, sobre o conteúdo informacional indispensável para o cumprimento do propósito e, desta forma, direciona a leitura do texto de insumo. Isto é, para avaliar a compreensão de leitura, é imprescindível que o enunciado contenha os elementos do texto de insumo que são necessários para a sua execução (Dell’Isola; Pordeus, 2021). Cada tarefa propõe especificações particulares, solicitando uma ampla variedade de gêneros discursivos, que mesclam distintos propósitos, interlocutores e temáticas (Mendel; Schoffen, 2017), o que pressupõe uma relação, ou integração de níveis diferentes de leitura e escrita a depender da tarefa (Scaramucci, 2016; Pileggi, 2017). Mencionar o material de insumo no enunciado é considerado por Vicentini (2019) como uma condição essencial para que o examinando utilize as informações contidas neste material. A depender do enunciado, é possível que o texto a ser produzido seja inteiramente dependente da leitura e compreensão do material de insumo (Scaramucci, 2016) e, por mais que a simples cópia não seja suficiente para a construção do texto, paráfrases, cópias e colagens são incentivadas em tarefas deste tipo (Pileggi, 2017).

Para além da compreensão global e localização de informações específicas do texto de insumo, o examinando precisa demonstrar sua capacidade de fazer inferências sobre o que é relevante para sua produção escrita, assumindo o papel de “leitor ativo[...] capaz de construir sentidos e usar conhecimento prévio” (Scaramucci, 2016, p. 405) a partir do que é solicitado pela tarefa. Para que a tarefa seja adequadamente cumprida, o examinando precisa realizar a articulação das informações do texto de insumo com as demais exigências do enunciado (Pileggi, 2017), pois a “compreensão envolve não apenas o conteúdo temático do texto, mas também a situação de produção em que o texto a ser [recontextualizado] está inserido” (Mendel, 2019, p. 52). Em outros termos, pode-se afirmar que a integração de habilidades se dá, no exame, por meio da recontextualização de informações do material disponibilizado, apoiada nas expectativas de compreensão e produção determinadas no enunciado da tarefa (Mendel, 2019).

Busca-se, com as tarefas propostas no exame, fornecer atividades que sejam realistas e desafiadoras, simulando eventos em que a leitura precede a escrita, partindo do pressuposto de que a competência comunicativa é global e requer um uso pragmático (Scaramucci, 2016).

Esta forma de avaliar “pauta-se em uma visão de linguagem como uma ação conjunta de participantes com um propósito social” (Dell’Isola; Pordeus, 2021), estando em oposição à avaliação isolada de aspectos da língua (Pileggi, 2017), o que impacta diretamente na preparação dos examinandos, minimizando algumas práticas mecânicas na preparação ao direcioná-la à leitura, escrita e reflexão (Scaramucci, 2016). Mendel e Schoffen (2017), afirmam que as tarefas integradas

proporcionam uma simulação de uso da língua de forma mais válida, isto é, a produção de um texto pertencente a determinado gênero do discurso em resposta a enunciados anteriores se aproxima de situações cotidianas de uso da língua mais do que as tarefas que testam uma única habilidade de forma isolada, posto que, nas atividades do dia-a-dia, não utilizamos as habilidades de compreensão e produção separadamente (Mendel; Schoffen, 2017, p. 152).

Os textos orais, escritos e multimodais que embasam as situações de uso da língua no exame são autênticos, “reais, não fabricados, fornecidos como fonte na qual os examinandos baseiam suas produções” (Vicentini, 2019, p. 571), relacionando língua e cultura por meio de experiências de mundo e práticas compartilhadas por uma comunidade (Schlatter; Scaramucci, 2021). Isto é, língua e cultura são vistos como elementos indissociáveis, fundamentados num conceito de proficiência que tem a interação verbal como a união do conjunto de conhecimentos necessários que leva o falante a desempenhar ações no mundo (Dell’Isola; Pordeus, 2021). É importante estar atento a diferentes pontos de vista sobre como o mundo pode ser compreendido e atribuir significados considerando as ações dos participantes e seus usos da linguagem nas práticas sociais propostas (Schlatter; Scaramucci, 2021).

Há, no Celpe-Bras, uma recorrência de reportagens, tanto televisivas, quanto radiofônicas, disponibilizadas como material de insumo (nas Tarefas 1 e 2). Vicentini (2019) afirma que, apesar destes gêneros serem roteirizados e, por conta disto, favorecerem a formalidade da língua falada, os trechos selecionados para serem incluídos no material sempre incluem entrevistas, favorecendo a avaliação da língua falada natural, que apresenta pausas, hesitações, reestruturações etc. Neste sentido, o que está em jogo vai muito além de aspectos linguísticos, como conhecimentos gramaticais, sendo centralizada a compreensão oral por meio de conhecimentos discursivos e sociolinguísticos, como coerência e coesão, sotaques etc (Vicentini, 2019). Geralmente, as Tarefas 1 e 2 exigem que o examinando apresente/descreva/exponha características de algo/alguém/algum local, havendo uma ênfase na recuperação de informações explícitas do material de insumo (vídeo e áudio,

respectivamente) (Vicentini, 2019). Quanto às Tarefas 3 e 4, há uma demanda de que o leitor se posicione em relação ao material de insumo (Mendel, 2019). Isto se dá pelas diferentes naturezas dos materiais de insumo, de modo que, como apontam os estudos, a integração de habilidades se dá de maneiras distintas nestes dois conjuntos de tarefas.

Para ilustrar como essa integração de habilidades se dá no exame, será realizada a **7. Apresentação de exemplo de tarefa**, com as especificações da Tarefa 4 da 2ª Edição de 2015, analisada neste trabalho, na Figura 3. Entre as informações, são determinados o locutor, um **morador de Belém**, e o interlocutor projetado, a **prefeitura municipal**. Como complemento da descrição do papel do examinando neste contexto comunicativo, há a informação de que ele está **inconformado com a situação dos casarões históricos da cidade**, o que também complementa o propósito de **explicar o problema e argumentar sobre a necessidade de se tomarem medidas imediatas para solucioná-lo**. O gênero de produção também é explicitado no enunciado, sinalizando que a interação está circunscrita em uma **carta aberta** que será **publicada em jornais locais**.

Figura 3 - Enunciado da Tarefa 4 de 2015-2

Você é morador de Belém e está inconformado com a situação dos casarões históricos da cidade. Com base na matéria “Azulejos valiosos”, escreva uma carta aberta endereçada à prefeitura municipal, para ser publicada em jornais locais. Seu texto deverá explicar o problema e argumentar sobre a necessidade de se tomarem medidas imediatas para solucioná-lo.

Fonte: Acervo Celpe-Bras, 2015.

As condições de produção desta tarefa também indicam que ela deve ser feita **com base na matéria “Azulejos valiosos”**, a reportagem que serve de material de insumo, como pode ser observado na Figura 4. O texto em questão, publicado no Jornal Em Dia, relata a situação de casarões históricos na cidade de Belém, que vêm sendo depredados e têm seus azulejos sendo alvo de roubo (Anexo A). No texto, são abordados, ainda, a possível existência de um mercado ilegal de azulejos e o processo de tombamento de um dos casarões. Com as

especificações dadas no enunciado, o examinando dispõe das condições necessárias para apropriar-se das informações do texto de insumo e articulá-las para cumprir com a tarefa.

Figura 4 - Design da Tarefa 4 de 2015-2

Você é morador de Belém e está inconformado com a situação dos casarões históricos da cidade. Com base na matéria "Azulejos valiosos", escreva uma carta aberta endereçada à prefeitura municipal, para ser publicada em jornais locais. Seu texto deverá explicar o problema e argumentar sobre a necessidade de se tomarem medidas imediatas para solucioná-lo.

Azulejos valiosos

Quatro casarões do século XIX são alvo de roubos e depredações em Belém.

A capital paraense já foi considerada uma das cidades brasileiras com maior variedade de azulejos, que cobriam as fachadas e o interior de residências. Boa parte deles foi importada da Europa, principalmente na virada do século XIX para o XX, surge da produção de borracha. Da década de 1970 para cá, no entanto, mais de 50% dos azulejos se perderam. Este ano, a situação parece ter se agravado. Desde fevereiro, pelo menos quatro casarões foram alvo de vandalismo. O assunto vem se espalhando pela capital paraense, e há quem suspeite de encomenda de roubos.

Uma das construções depredadas é o Palácio Visor Maria da Silva, batizado com o nome de seu antigo dono, inspetor de obras do estado do Pará no governo Augusto Montenegro (1901-1909).

Os azulejos foram encontrados dias depois, em cacos, e estão no Laboratório de Conservação e Restauração da UTPA (Lacor): "Recebemos aqui no laboratório mais de 1.000 fragmentos de azulejos e estamos montando o quebra-cabeça para ver a que patência pertencem. Vamos limpar e organizar o material até o fim de junho. Só depois será decidido o que pode ser restaurado ou refeito", explica Thais Sarajed, coordenadora do Lacor.

Há cerca de um ano, o Departamento do Patrimônio Histórico, Artístico e Cultural (Dphac) iniciou o processo de tombamento do casarão. Segundo a diretora Thais Toscano, o procedimento é demorado, por ser necessário documentar detalhes arquitetônicos e históricos da

construção. "No caso deste imóvel, os detalhes se tornam mais elaborados, dado o nível artístico dos painéis de azulejo. Mas o local já foi interditado".

A proteção do palacete parece encaminhada, mas a situação na cidade causa preocupação, já que outros três casarões tiveram azulejos do século XIX furtados. "Foram roubos pontuais muito comuns. O Palácio Visor Maria da Silva tem um dos interiores mais bonitos da cidade, mas por fora é muito simples, não chama atenção. As pessoas que invadiram devem ter sido encarregadas de roubar azulejos. Ou então foi uma tentativa de desqualificação da propriedade, para que se possa fazer o que quiser com o patrimônio", suspeita a arquiteta e urbanista Cláudia Nascimento. A superintendente do Iphan no Pará, Maria Dorotéia Lima, concorda: "Tudo indica que há um mercado de azulejos na cidade, até porque os exemplares fora das áreas tombadas não têm qualquer proteção, o que pretendemos fazer em breve", disse. Enquanto as investigações não forem concluídas, os poucos exemplares de azulejos que ainda restam aumentam cada vez mais de valor.

Fonte: Acervo Celpe-Bras, 2015.

Com base no conceito de validade, que será discutido na seção **1.3 Validade, confiabilidade, autenticidade, impacto e praticidade**, a “integração de habilidades contribui para que as tarefas simulem situações de uso da língua de forma mais válida” (Mendel; Schoffen, 2017), uma vez que o exame mede “aquilo que ele deve medir” (Schlatter *et al.*, 2005, p. 15). O Celpe-Bras, portanto, é um “exame que se preocupa em operacionalizar uma visão de língua(gem) ampla, de uso da língua e de habilidades integradas” (Vicentini, 2019, p. 577).

1.2.2.2 Parte Oral

A PO tem duração de aproximadamente 20 minutos por examinando, e é dividida em duas partes: os primeiros cinco minutos da interação são reservados para que o examinando fale um pouco sobre si; nos 15 minutos restantes, fala-se sobre três assuntos diversos, sendo aproximadamente 5 minutos para cada assunto. A interação ocorre entre o examinando e o

Avaliador Interlocutor (AI), e “parte de situações de comunicação próximas às reais, procurando simular, na medida do possível, conversas do dia a dia” (Brasil, 2020, p. 41) de maneira que “se aproxime de uma conversa, em que o examinando seja convidado a se expressar com a liberdade e a espontaneidade que a situação permite” (Brasil, 2020, p. 42).

Quadro 3 - Estrutura da Parte Oral

Etapa	Conteúdo da interação	Práticas envolvidas	Tempo
1	Conversa sobre interesses pessoais do examinando com base nas informações do formulário de inscrição	Compreensão oral e produção oral	5 minutos
2	Conversa sobre tópicos do cotidiano e de interesse geral com base em três elementos provocadores	Compreensão oral, compreensão escrita e produção oral	15 minutos (cinco minutos para cada elemento provocador)

Fonte: Brasil, 2013; 2020.

A PO, conduzida pelo AI, é pautada no questionário de inscrição do Celpe-Bras de cada examinando que, além das perguntas específicas sobre a realização do exame, contém, também, perguntas relacionadas à vida do examinando, por exemplo: “*O que levou você a se interessar pela língua portuguesa?*”, “*Quais aspectos específicos da cultura brasileira chamam sua atenção? Por quê?*”, “*Como é um dia típico de sua rotina?*” ou “*Comente sobre suas preferências (hobbies, comidas, lembranças da infância, algo que gostaria de aprender, desejos para o futuro, pessoa famosa que você admira ou algo de seu país de que você se orgulhe)*”. No primeiro momento da PO, o examinador busca fazer perguntas relacionadas ao questionário do examinando, com o intuito de que o examinando expanda as respostas previamente dadas, relacionadas à vida do examinando no geral, à sua família, ao seu trabalho e atividades no geral, aos seus interesses, à cultura brasileira etc. Dessa forma, o examinando tem a possibilidade de falar sobre assuntos com os quais está mais habituado, criando uma atmosfera mais propícia à interação que seguirá.

Também com base nas respostas que o examinando dá no questionário de inscrição, são escolhidos os temas sobre os quais se vai falar no segundo momento da interação. Essa parte ocorre com base nos chamados Elementos Provocadores (EPs), textos curtos e, muitas vezes, multimodais aos quais o examinando tem acesso durante essa etapa. Os EPs

“compreendem textos ou recortes de textos de diferentes gêneros discursivos (reportagens, notícias, panfletos, propagandas, cartuns, tirinhas, gráficos, mapas etc.) que circulam ou circularam no Brasil retirados de distintos suportes (jornais, revistas, sites, livros etc.)” (Brasil, 2020, p. 43). Iniciada a segunda etapa da interação, o examinando tem um minuto para fazer a leitura silenciosa do primeiro EP e, então, a interação segue um roteiro específico, que fica sob posse do avaliador apenas. Esse procedimento se repete para o segundo e o terceiro tema e, por fim, está encerrada a PO.

1.2.3 Avaliação do exame

Nesta seção, será tratado dos procedimentos de avaliação da PE e da PO, ou seja, da **6. Descrição dos critérios de correção**, bem como da avaliação final. Os procedimentos e os parâmetros de avaliação de ambas as partes serão descritos conforme consta no Documento Base do Celpe-Bras (Brasil, 2020).

1.2.3.1 Processo de avaliação da Parte Escrita

Para garantir a confiabilidade no processo de avaliação da PE, os avaliadores passam por uma capacitação, que envolve discussão sobre textos produzidos por examinandos na edição em questão para que os ajustes nos parâmetros de avaliação sejam devidamente realizados. Os avaliadores são divididos em quatro grupos, cada um responsável por uma tarefa específica. Cada grupo conta com um coordenador, responsável por acompanhar todo o processo de avaliação da tarefa da qual está encarregado. Além disso, o coordenador, juntamente com os coordenadores das outras tarefas, têm a tarefa de ler e discutir uma amostra estratificada e representativa de textos produzidos pelos examinandos para cada tarefa para redigir as especificações para a avaliação de cada tarefa. A partir dessa amostra, são selecionados os textos a serem discutidos com cada grupo, apresentando as possíveis dificuldades para a avaliação e os comentários feitos pelos coordenadores, a fim de balizar os parâmetros e refinar as especificações da tarefa.

O desempenho linguístico varia de acordo com o contexto e a tarefa e, portanto, as interpretações da habilidade linguística de um examinando também variam de acordo com o desempenho (Douglas, 2000). O intuito de separar os avaliadores por tarefas é de diminuir as discrepâncias e aumentar a confiabilidade, uma vez que, ao se especializarem nas especificações e na grade de avaliação de uma única tarefa, aumentam as chances de que ele

utilize os mesmos critérios de avaliação para os textos de todos os examinandos. Esses textos são avaliados de forma independente por dois avaliadores, distribuídos de maneira que haja rotatividade entre os avaliadores e a quantidade de textos seja equivalente entre eles. Cada avaliador atribui uma nota, que vai 0 a 5, sem o conhecimento da nota atribuída pelo outro avaliador, e a nota final é calculada pelo sistema, sendo a média aritmética das notas dos dois avaliadores. Caso a diferença entre as notas dos dois avaliadores seja maior do que um ponto, considera-se como um caso de discrepância, sendo necessária uma nova avaliação. A nota final da tarefa, nesse caso, é a média entre a terceira nota e a nota mais próxima atribuída anteriormente¹⁹. Quanto à nota final da PE, ela é calculada com base nas quatro tarefas, sendo a média aritmética entre as notas finais de cada uma. Segat (2023), em seu estudo sobre as atribuições de notas na PE em três edições do Celpe-Bras, concluiu que o consenso entre os avaliadores da PE do Celpe-Bras é considerado aceitável, com valores inclusive “mais próximos do que seria considerado um consenso alto entre os avaliadores” (Segat, 2023, p. 138).

Por mais que cada tarefa apresente especificações próprias, relacionadas às condições de produção explicitadas no enunciado, todas baseiam a sua avaliação em parâmetros de avaliação comuns (Anexo B). Esses parâmetros reúnem informações sobre gênero, interlocutor, propósito, conteúdo informacional, organização textual e recursos linguísticos. Como mencionado anteriormente, a avaliação da PE é holística, uma vez que o construto do exame é “incompatível com a ideia de avaliar o conhecimento ou a acurácia dos recursos linguísticos de forma independente do uso da língua” (Brasil, 2020, p. 38), ou seja, os elementos presentes nos textos são avaliados à medida em que são adequadamente empregadas para o cumprimento da tarefa, inserida em determinado contexto comunicativo.

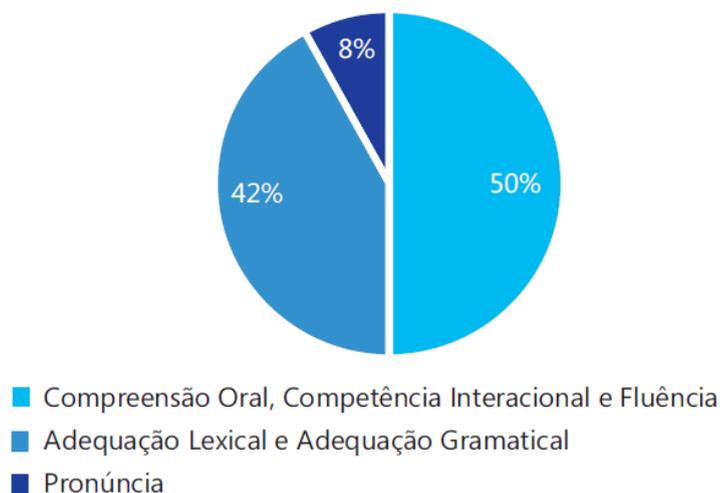
1.2.3.2 Processo de avaliação da Parte Oral

Na PO, como explicado anteriormente, há um AI e um Avaliador Observador (AO), que dispõem, respectivamente, de uma grade holística e uma grade analítica, devendo avaliar cada examinando com base nesses diferentes critérios. Ambos os avaliadores avaliam, com notas que vão de 0 a 5, a interação sob os mesmos seis aspectos, com a diferença de que o AI atribui uma única nota e o AO atribui uma nota para cada um dos seis aspectos. As seis notas

¹⁹ Caso a terceira nota seja equidistante das duas novas atribuídas anteriormente, ela é considerada a nota final da tarefa.

atribuídas pelo AO não têm o mesmo peso, sendo divididas conforme apresentado na Figura 5.

Figura 5 - Pesos da avaliação do Avaliador Observador na Parte Oral



Fonte: Brasil, 2020.

A nota do AO é, portanto, a média ponderada das seis notas com base nesses pesos. A nota final da PO é a média aritmética das notas do AI e do AO. Os casos de discrepância na PO ocorrem quando a diferença entre as notas atribuídas entre pelos dois avaliadores da PO seja de 1,5 ou mais, ou a diferença entre a nota final da PO e a da PE seja maior de 2,0 pontos. Neste caso, há a necessidade de realizar uma reavaliação, que só é possível de ser realizada porque as interações são gravadas do início ao fim, permitindo a conferência da avaliação e contribuindo para aumentar a confiabilidade do exame.

Da mesma forma como ocorre na capacitação da PE, os coordenadores da PO discutem uma amostra aleatória de áudios e selecionam alguns deles para serem utilizados com os avaliadores que participarão da reavaliação da PO, com a diferença de que estes não são divididos em grupos. Durante o processo de reavaliação, cada áudio é ouvido e avaliado independentemente por dois avaliadores, utilizando cada um uma das grades de avaliação (holística e analítica). A nota final é, então, a média aritmética das duas notas atribuídas pelos dois avaliadores. Quando há um novo caso de discrepância entre as notas, ocorre uma nova reavaliação, nos mesmos moldes da primeira. A nota final da PO, neste caso, é, portanto, a “média aritmética das notas convergentes para cada uma das grades, considerando tanto as

notas atribuídas pelos avaliadores nos postos aplicadores quanto às notas atribuídas por eles durante o processo de reavaliação” (Brasil, 2020, p. 77)²⁰.

1.2.3.3 Resultado final

Com base na avaliação de ambas as partes, o examinando é alocado em um dos níveis de proficiência relacionados às faixas de nota acima. Não há a possibilidade de obtenção de certificados separados para Parte Escrita e Parte Oral. Cada nível de certificação está atrelado a uma descrição, representativa do desempenho do examinando no exame. As descrições “foram determinadas com base na experiência de certificação do desempenho dos participantes ao longo dos anos” (Brasil, 2020, p. 79). A nota final do Celpe-Bras é baseada no menor resultado entre ambas as partes, sendo necessário atingir, no mínimo, o nível Intermediário. Ou seja, se o examinando atingiu um desempenho equivalente ao nível Avançado na Parte Oral, mas, na Parte Escrita, alcançou o nível Intermediário Superior, seu Certificado de Proficiência em Língua Portuguesa será de nível Intermediário Superior. Na Figura 6, são apresentadas as relações das notas finais com os níveis de certificação.

Figura 6 - Níveis de certificação por faixa de nota

Nível	Faixa de notas
Sem Certificação	De 0,00 a 1,99
Intermediário	De 2,00 a 2,75
Intermediário Superior	De 2,76 a 3,50
Avançado	De 3,51 a 4,25
Avançado Superior	De 4,26 a 5,00

Fonte: Brasil, 2020.

Como mencionado na seção **1.2.3.1 Processo de avaliação da Parte Escrita**, Segat (2023) conduziu um estudo sobre as atribuições de notas na PE em três edições do Celpe-Bras, as edições de 2016/1, 2016/2 e 2017/1. Ao analisar os dados referentes ao número

²⁰ “Até a edição 2019/2, as discrepâncias decorrentes da reavaliação da Parte Oral do Celpe-Bras eram solucionadas com uma reavaliação feita por apenas um avaliador que, utilizando a grade holística, atribuía uma nota única, a qual substituía todas as demais.” (Brasil, 2020, p. 77)

de examinandos que obtiveram a nota final de proficiência oriunda da nota da PE, Segat (2023) chegou a resultados que apontam para a relevância de realizar estudos que buscam o refinamento dos níveis avaliados na PE. Na edição de 2016/1, 84,42% dos examinandos tiveram sua nota final de proficiência proveniente da PE, enquanto nas edições de 2016/2 e 2017/1, este foi o caso de, respectivamente, 79,06% e 71,52%. Em todas as edições, o percentual supera os 70%, justificando que mais pesquisas que se debrucem sobre o refinamento da descrição dos níveis da Parte Escrita.

1.3 Validade, confiabilidade, autenticidade, impacto e praticidade

Incluídos no **8. Desenvolvimento de um plano para avaliar as qualidades de uma boa avaliação**, item 8 da lista proposta por Douglas (2000), estão os conceitos de **validade, confiabilidade, impacto, autenticidade e praticidade**, que serão explicados a seguir.

A **confiabilidade** pode ser compreendida como a acurácia (Douglas, 2000) e a “consistência de medição” (Bachman; Palmer, 1996, p. 19) de um exame. Em outras palavras, a confiabilidade é a definição do grau em que a pontuação reflete de fato as habilidades dos examinandos (Yan; Fan, 2022). Segat (2023) sugere que uma das variáveis que contribui para a acurácia de um exame é a “operacionalização adequada do construto [...] nas tarefas e nas grades de avaliação” (Segat, no 2023, p. 24), e, segundo a autora, faz-se necessária a definição minuciosa do construto que orienta o teste e que será operacionalizado nas tarefas, especificações e grade de avaliação de modo a permitir que se compreenda e estude de que forma os resultados do exame são quantificados. A pergunta para a confiabilidade, neste caso, centra-se na estabilidade e precisão dos processos e resultados da avaliação (Segat, no 2023), ou seja, supondo que um mesmo examinando apresente a mesma performance em um mesmo exame em duas ocasiões diferentes, este examinando deve, de acordo com Bachman e Palmer (1996), receber a mesma nota em ambas as ocasiões. Ou ainda, seguem os autores, se um exame possui mais de uma versão que, supostamente, podem ser utilizadas de maneira intercambiável, um mesmo examinando deve receber a mesma nota em ambas as versões, caso as realize. Yan e Fan (2022), no entanto, argumentam que o conceito de confiabilidade pode ser considerado contraintuitivo, pois é composto, simultaneamente, pelo resultado, ou seja, a nota que o examinando deve, teoricamente, receber, e a medição do erro, ou seja, o grau em que uma medição apresenta (ou não) erros. Neste sentido, o problema da confiabilidade é caracterizado pela identificação dos possíveis erros ou fontes de variância em

contraste com a constância e acurácia dos processos e resultados da avaliação (Segat, 2023). Para Segat (2023), “erro de mensuração” refere-se a “uma fonte de variância inesperada e indesejada na obtenção dos resultados de um teste, e não [...] um erro propriamente dito” (Segat, 2023, p. 25). Dirigindo o foco para as notas atribuídas por avaliadores humanos, depara-se com a pergunta: “quando diferentes avaliadores dão a mesma nota, eles querem dizer a mesma coisa?” (McNamara, 2000, p. 55). Muitas vezes, sim, mas muitas vezes, não. Para diminuir as inconsistências na avaliação ou garantir que a interpretação dos resultados seja semelhante entre avaliadores, e, conseqüentemente, contribuir para a confiabilidade, pode-se investir, por exemplo, no treinamento de avaliadores e na especificação dos parâmetros de avaliação.

Relacionado à confiabilidade está a **validade**, ou validade de construto (Bachman; Palmer, 1996), conceituada como um processo que diz respeito à pertinência e à adequação das interpretações feitas dos resultados (Douglas, 2000) ou seja, refere-se ao grau em que se pode interpretar os resultados como indicadores das habilidades ou do construto que se quer medir (Bachman; Palmer, 1996). Discussões mais atualizadas acerca da validade ampliam o conceito, sugerindo que ela abrange o aporte teórico e prático, ou seja, é baseado nas evidências que apoiam as interpretações dos resultados de um exame (Chapelle; Lee, 2022). Os resultados, de acordo com Chapelle e Lee (2022), são utilizados como indicadores da habilidade linguística dos examinandos, que contêm uma ampla gama de conhecimentos e capacidades complexas e, por este motivo, a interpretação que será feita dos resultados sobre o que está sendo compreendido como saber uma língua precisa ser bem especificada. As autoras afirmam que o desafio que está posto é o de recontextualizar a perspectiva de língua que se pretende avaliar, ou seja, da língua em uso, em uma definição do construto que sirva para a interpretação dos resultados. Fala-se, neste sentido, sobre uma **validade baseada em argumento**, descrita por Chapelle e Lee (2022) como ferramentas e conceitos precisos para a definição de construtos linguísticos a partir de **garantias** e **suposições** subjacentes à extrapolação das inferências.

As **garantias** são declarações sobre o que precisa ser apoiado por evidências empíricas ou fundamentos teóricos, e as **suposições** são declarações sobre a teoria e as evidências específicas que fornecem esse apoio. Esses requisitos podem ser úteis para ajudar a estabelecer a conexão entre a pesquisa de validação e as declarações

teoricamente fundamentadas sobre a interpretação do construto pretendido²¹ (Chapelle; Lee, 2022, p. 24, grifo nosso).

Kane (2022) sugere que para especificar detalhadamente a interpretação e os usos propostos, é necessário estabelecer uma rede de inferências que relacione os resultados às interpretações e usos propostos dos resultados. O autor afirma que diferentes tipos de inferência requerem diferentes tipos de evidência para sua avaliação e, nesse sentido, Ross (2022) afirma que inferências feitas sobre a validade de um exame se baseiam em abordagens tanto empíricas quanto interpretativas. Estudos que se propõem a averiguar os resultados de exames, triangulando informações a respeito das notas atribuídas, dos critérios de avaliação e do aporte teórico no qual o exame tal é pautado são uma forma de buscar a validade. Scaramucci (2011) afirma que “um teste não tem como ser válido sem antes ser confiável (consistente e estável), uma vez que para ser válido necessita avaliar com precisão e de forma consistente” (Scaramucci, 2011, p. 105). A confiabilidade é, portanto, compreendida como uma condição de validade (Yan; Fan, 2022). No entanto, pode existir uma tensão entre ambos os termos, visto que, se não forem estabelecidos critérios elucidativos para a correção de uma tarefa ou se os avaliadores não forem treinados para avaliá-la, a tarefa e o teste podem ser válidos, porém não confiáveis (Scaramucci, 2011). Nestes casos, deveria-se otimizar tanto o uso quanto a interpretação da pontuação em vez de priorizar um em detrimento do outro (Yan; Fan, 2022). Utilizar resultados de pesquisas que se debruçam sobre a definição dos níveis de proficiência para o refinamento dos critérios de avaliação pode ser um dos caminhos para diminuir a tensão existente entre a confiabilidade e a validade, uma vez que tornar os critérios de avaliação mais robustos auxilia os avaliadores a compreenderem o que está sendo avaliado, dando-lhes mais subsídios na hora da correção.

Dando enfoque às consequências sociais, uma característica importante é o **impacto**, que pode ser definido como o efeito de um exame em escala macro, ou seja, na sociedade e no sistema educacional, e em escala micro, nos indivíduos envolvidos (Bachman; Palmer, 1996), sejam eles estudantes, examinandos ou professores. Dentro do **impacto**, há o chamado **efeito retroativo**, que considera os “impactos sociais, éticos e políticos [...] em contextos de ensino/aprendizagem de línguas, [...] reconhe[cendo] o poder exercido pela avaliação nos participantes e na sociedade em geral” (Scaramucci, 2004, p. 204) e deve “ser avaliado com

²¹ No original: *Warrants are statements about what needs to be supported by empirical evidence or theoretical rationales, and assumptions are statements about the specific theory and evidence that provide the support. These requirements may be useful in helping make the connection between validation research and theoretically grounded statements about the intended construct interpretation* (Chapelle; Lee, 2022, p. 24).

referência às variáveis, objetivos e valores da sociedade, do sistema educacional em que é usado, assim como o resultados potenciais de seu uso” (Scaramucci, 2004, p. 204). Ou seja, caracteriza-se como a influência que um exame pode exercer “no ensino, na aprendizagem, no currículo, na elaboração de materiais didáticos e nas atitudes das pessoas envolvidas” (Scaramucci, 2011, p. 109). Quando percebe-se, por exemplo, uma mudança no ensino em cursos de língua, que passa a ser pautado em práticas sociais de uso da língua estabelecidas em um exame, há sinais de efeito retroativo.

Quanto à **autenticidade**, esta pode ser compreendida como o grau de correspondência entre a performance exigida no exame e as situações reais fora dele, ou seja, relaciona as tarefas do exame à possibilidade de generalização dos resultados no que se refere aos usos futuros da língua (Bachman; Palmer, 1996). Por último, há, também, a **praticidade**, que se remete à questão do grau em que as demandas propostas nas especificações podem ser atendidas dentro dos limites dos recursos existentes (Bachman; Palmer, 1996), como, por exemplo, dinheiro, tempo, equipe e políticas educacionais (Douglas, 2000). Bachman e Palmer (1996) sugerem que a união e o equilíbrio dessas qualidades refletem na **utilidade** do exame, que deve ser o ponto mais importante da sua elaboração.

1.4 Resumo do capítulo

Neste capítulo, foi apresentado o exame Celpe-Bras. O capítulo iniciou com um panorama geral do contexto de elaboração do exame, seguido de suas concepções teóricas, especificamente a noção de proficiência relacionada aos gêneros do discurso. Depois, foi descrita a estrutura do exame, sendo apresentadas a Parte Escrita, com atenção especial à integração de habilidades, e a Parte Oral. Na sequência, foram descritos os processos de avaliação de ambas as partes, incluindo o cálculo do resultado final e, por fim, foram apresentados conceitos basilares no que diz respeito à qualidade de uma boa avaliação.

No próximo capítulo, será apresentada a base teórica que fundamenta este trabalho: os estudos informados por corpus.

2. CASARÕES HISTÓRICOS: Pressupostos teóricos

A área de estudos sobre avaliação de proficiência linguística é, por natureza, interdisciplinar (Götz, 2022). Esse campo de estudos se debruça sobre a avaliação de indivíduos em uma variedade de contextos e para uma série de finalidades, sendo o principal objetivo o de medir características que permitam inferir sobre a proficiência linguística em uma habilidade específica ou sobre a competência linguística geral (Barker, 2010).

Avaliação de línguas informada por corpus é como se denomina a área de pesquisa interdisciplinar que abrange a Avaliação de línguas e a Linguística de Corpus (Nagasawa, 2023). Tem-se, nesta área, o corpus como fonte de referência para fornecer evidências sobre exames de língua, como, por exemplo, informações sobre o conteúdo ou validação da avaliação de avaliadores humanos (Callies; Díez-Bedmar; Zaytseva, 2014). Nagasawa (2023) sugere uma série de possíveis aplicações de corpora no aprimoramento de exames, bem como análises que viabilizariam estas aplicações. Análises que possibilitam a detecção de padrões que podem informar tanto grades de avaliação analíticas, holísticas e pontos de corte de níveis de proficiência, quanto algoritmos de correção automática, geração automática de feedback e até de identificação de plágio ou cópia são algumas possibilidades de aplicações sugeridas pela autora a partir do trabalho com corpora de textos produzidos por examinandos. Pode-se, por exemplo, prover informações práticas sobre os usos feitos em diferentes níveis de proficiência (Götz, 2022), ou seja, sobre o que é avaliado em cada nível (Barker, 2010), que podem ser incorporadas à grade de avaliação. Nesta seção, serão operacionalizados conceitos fundamentais da Linguística de Corpus e a sua relação com a avaliação de proficiência linguística.

2.1 Campo de estudos da Linguística de Corpus

A Linguística de Corpus (LC) concentra-se na coleta e análise de corpora, coleções de textos em formato eletrônico, compilados cuidadosamente com critérios específicos (Sinclair, 1999), com o objetivo de servirem para a pesquisa de determinados aspectos linguísticos por meio de processamento computacional (Sardinha, 2000). Biber (2012) define a LC como uma metodologia de pesquisa que apoia investigações empíricas sobre variação e uso da língua, resultando em descobertas de pesquisa que têm maior capacidade de generalização e validade do que seria possível de alguma outra forma. O autor argumenta, no entanto, que a LC é muito mais do que uma abordagem metodológica, pois suas inovações metodológicas

possibilitaram aos pesquisadores formular perguntas substancialmente diferentes, o que alterou radicalmente a perspectiva sobre variação e uso da língua em relação às abordagens de pesquisa anteriores.

Sardinha (2000) sintetiza os pontos mais importantes constituintes de um corpus:

- (a) A origem: Os dados devem ser autênticos
- (b) O propósito: O corpus deve ter a finalidade de ser um objeto de estudo lingüístico
- (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 (Sardinha, 2020, p. 338).

Quando a língua é produzida para fins de comunicação e não com o intuito de ser usada para análise linguística, então se tem dados autênticos (Stefanowitsch, 2020), ou seja, o propósito de estudo, que estabelece o que exatamente se quer averiguar, é uma etapa posterior à produção dos dados. O propósito é de fundamental importância, pois é essencial que o corpus apresente em sua composição as características linguísticas que o estudo pretende analisar, de modo que é preciso que os critérios de compilação do corpus permitam que os objetivos sejam atingidos (Kuhn; Ferreira, 2020). As pesquisas linguísticas são realizadas por meio de análises computacionais (Viana, 2011), utilizando *softwares* específicos (Evinson, 2010) de processamento automatizado da linguagem, de modo que o corpus precisa estar no formato adequado para permitir que seja analisado.

Atrelado ao propósito, Kuhn e Ferreira (2020) sugerem que o tamanho do corpus pode ser estabelecido com base no número total de palavras que o corpus vai precisar para se adequar ao propósito de pesquisa estabelecido. Além disso, os autores apontam uma segunda perspectiva para determinar o tamanho do corpus, que pode levar em conta a “quantidade de textos disponíveis representantes da linguagem de interesse” (Kuhn; Ferreira, 2020, p. 6). Compreende-se, dessa forma, que o tamanho do corpus pode ser considerado subjetivo, e o mesmo ocorre quanto à representatividade: tanto a extensão quanto a representatividade do corpus são subjetivas, não havendo uma determinação exata do que pode ser considerado representativo ou extenso o suficiente, porém, pode-se afirmar que um corpus deve ser grande o suficiente para conter amostras suficientes para análise do fenômeno que se deseja investigar, sendo, assim, representativo (Stefanowitsch, 2020).

Outro ponto importante na constituição de um corpus é o equilíbrio, que se refere “à distribuição de textos de acordo com o tipo de texto que compõe o corpus” (Kuhn; Ferreira, 2020, p. 6). Há algumas maneiras de garantir o equilíbrio: ou se pode optar por balancear os textos por gênero, por exemplo, de forma equilibrada, ou se pode determinar um número relativamente parecido de palavras, ou até mesmo de textos, para que os subcorpora sejam do mesmo tamanho.

É comum que um corpus, depois de compilado, seja anotado (marcado)²² (Cushing, 2022). Há, de acordo com Aluísio e Almeida (2006), basicamente dois níveis de anotação: **anotação estrutural** e **anotação linguística**. A anotação estrutural inclui dados externos, como tipo de autoria, gênero do texto, nível de proficiência (Cushing, 2022), e dados internos, como

a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) marcação da estrutura de subparágrafos – elementos que são de interesse linguístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc (Aluísio; Almeida, 2006, p. 161).

Estes são os chamados “metadados”, e sua adição contribui para que se faça uma análise mais avançada do corpus (Kuhn; Ferreira, 2020). Não há um padrão de tipo ou quantidade de informações a serem adicionadas como metadados, isto vai depender da quantidade de informações que foram coletadas juntamente com os textos (Kuhn; Ferreira, 2020).

Já a anotação linguística pode ser realizada em diversos níveis, como “morfofossintático, semântico, discursivo, etc., sendo inserida manualmente (por linguistas), automaticamente (por ferramentas de Processamento de Língua Natural - PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas” (Aluísio; Almeida, 2006, p. 161). Este tipo de anotação permite que os dados sejam analisados de forma mais precisa (Aluísio; Almeida, 2006). Quando os textos de um corpus contêm inadequações ortográficas, é comum que a anotação envolva a normalização das inadequações (del Rio *et al.*, 2016), caso contrário, algumas análises podem apresentar distorções devido às diferentes grafias de uma mesma palavra. Como este processo é realizado por códigos, é possível manter a versão original, inserindo, como marcação, a versão regularizada.

²² Em inglês: *annotated (tagged)*.

A pesquisa com corpora reforça a ideia de que a variação na linguagem é sistemática e pode ser analisada por meio de métodos empíricos e quantitativos, sendo a sua maior contribuição a documentação da presença de construtos linguísticos que não haviam sido reconhecidos nas teorias linguísticas existentes (Biber, 2012). A seguir, serão apresentadas técnicas de manipulação e pesquisa de corpora, bem como diferentes tipos de corpora e discussões terminológicas.

2.1.1 Como pode ser utilizado um corpus?

Para além do refinamento nas descrições a respeito do léxico e dos estudos sobre aspectos gramaticais, a LC pode ser utilizada como ferramenta de pesquisa em diversas áreas, como ensino, aprendizado e avaliação de línguas, análise do discurso, estilística literária, linguística forense, pragmática, sociolinguística e comunicação, entre outras (McCarthy; O’Keeffe, 2010). Como ponto de partida, observa-se repetições e padronizações no uso de determinados elementos linguísticos, que ocorrem na língua de maneira inesperada (Hunston, 2010), e um corpus pode servir como uma fonte valiosa para revelar os aspectos repetidos e padronizados da língua que, de outra forma, não seriam percebidos (Reppen, 2010). Embora corpora não tenham, por si só, a capacidade de fornecer informações sobre o uso da língua, o fato de serem coleções de textos eletrônicos viabiliza a utilização de *softwares* específicos (Evinson, 2010), que encontram e organizam esses dados (Hunston, 2010).

A LC é, portanto, caracterizada como um tipo de pesquisa que envolve, principalmente, a análise da frequência de determinados fenômenos linguísticos (Cushing, 2022). No entanto, a simples contagem de palavras ou a exibição de suas ocorrências não são suficientes para se tirar conclusões sobre o uso da língua (Evinson, 2010), pois a “interpretação é uma atividade humana” (Hunston, 2010, p. 158), sendo necessário realizar análises posteriores dos dados apresentados pelo *software*, podendo ser tanto quantitativas, quanto qualitativas, ou até mesmo uma combinação entre ambas. Assim dizendo, cabe ao pesquisador o indispensável papel de transformar as informações advindas da análise do corpus em resultados significativos e concretos sobre o objeto pesquisado.

2.1.2 Abordagens de pesquisa

Na LC as duas principais abordagens de pesquisa são a pesquisa baseada em corpus (*corpus-based research*) e a pesquisa guiada por corpus (*corpus-driven research*). É sobre a

descrição do uso de cada uma delas em pesquisas na área de avaliação linguística que se deterá nesta seção.

A **pesquisa baseada em corpus** é fundamentada na validação de formas e estruturas linguísticas advindas da teoria, tendo como principal objetivo a análise de padrões sistemáticos de variação e uso de características linguísticas predefinidas (Biber, 2012). Ou seja, é comumente utilizada para confirmar ou mesmo refutar hipóteses previamente levantadas, muitas vezes (mas não sempre) comparando produções escritas em diferentes níveis de proficiência em língua adicional a padrões de uso encontrados em corpora produzidos por falantes nativos (Callies; Díez-Bedmar; Zaytseva, 2014). O intuito dessas comparações é, a partir das semelhanças e diferenças de uso, tirar conclusões sobre características linguísticas em cada um dos níveis de proficiência (Callies; Díez-Bedmar; Zaytseva, 2014), buscando descritores/características capazes de melhor diferenciar esses níveis (Götz, 2022). Além da comparação com corpora de falantes nativos, ela também pode ser conduzida com base na réplica de estudos anteriores, averiguando a aplicabilidade dos resultados desses estudos em corpora diferentes.

A **pesquisa guiada por corpus** é mais indutiva, e os fenômenos linguísticos a serem analisados emergem a partir da própria análise do corpus (Biber, 2012). Nesta segunda abordagem, faz-se poucas suposições *a priori* em relação às características linguísticas que devem ser aplicadas nas análises e os padrões de co-ocorrência, descobertos a partir da análise, servem de base para as descrições posteriores. As evidências encontradas a respeito dos níveis de proficiência são baseadas nas análises estatísticas automáticas a partir dos testes no *software* (Götz, 2022), e o pesquisador se utiliza dessas técnicas para a escolha dos fenômenos linguísticos a serem posteriormente avaliados (Callies; Díez-Bedmar; Zaytseva, 2014).

Por mais que se tivesse, aqui, o intuito de dar seguimento a estudos previamente realizados (Divino, 2021; Schoffen; Divino, 2023), como a descrição de elementos utilizados em cada nota para configurar o gênero discursivo solicitado pela tarefa, muitos dos resultados aqui descritos partiram de testes realizados no *software*, sem muitas suposições *a priori*. Isto significa dizer que, embora algumas das características linguísticas de análise estivessem predefinidas, boa parte dos fenômenos linguísticos analisados e descritos neste trabalho se fizeram relevantes a partir da própria análise. Pode-se considerar que esta pesquisa é, majoritariamente, **guiada por corpus**, pois alguns dos fenômenos linguísticos foram escolhidos depois de serem realizadas as análises automáticas no *software*, ainda que

apresente traços que poderiam se encaixar na descrição da **pesquisa baseada em corpus**, visto que algumas das análises foram conduzidas com o objetivo de confirmar (ou refutar) hipóteses previamente levantadas.

2.1.3 Técnicas de manuseio

Quanto ao manuseio de um corpus, existem diversas técnicas e possibilidades de análise, as técnicas primárias iniciam em algum *software* específico de análise de corpora, como o *SkE* ou o *Wordsmith*, ou até mesmo *softwares* utilizados para análise de bancos de dados, como o *RStudio*, ou o *SAS*. O *software* utilizado para as análises deste trabalho é o *SkE*, um conjunto de ferramentas *online* com uma gama de funções altamente flexíveis para manusear corpora (Kunilovskaya; Koviagina, 2017). Com o *SkE*, é possível gerar resumos automáticos do comportamento gramatical e da colocação de palavras, encontrar palavras semelhantes em termos de sinônimos e diferenças de usos de uma mesma palavra, bem como realizar busca mais manual de itens no corpus (Kilgarriff *et al.*, 2004). A seguir, serão apresentadas algumas das ferramentas oferecidas pelo *SkE*, sendo elas: *Word Sketch*, *Thesaurus*, *Concordance*, *Wordlist*, *N-Grams* e *Keywords*. Para ilustrar, serão apresentadas imagens-exemplo de cada uma com um dos corpora analisados neste trabalho, que será apresentado na seção **3.3 CorCel: o corpus da pesquisa**. O corpus se chama CorCel e o subcorpus utilizado aqui neste trabalho se chama CorCel20152t4n5r.

Word Sketch é uma ferramenta que permite que se tenha um **resumo do comportamento gramatical ou contextual de uma palavra** (Kilgarriff *et al.*, 2004). Para isto, é necessário escolher o termo sobre o qual se deseja obter tais informações para fazer a busca no *software*. Como resultado, obtém-se, por exemplo, de que maneiras o termo foi usado como sintagma preposicional, ou quando antecede ou sucede um verbo, um adjetivo, um substantivo, etc. Desta forma, pode-se ter um panorama geral do comportamento do termo no corpus. Para ilustrar, o termo “azulejo” foi buscado no CorCel20152t4n5r, como apresentado na Figura 7.

Figura 7 - Resultado da ferramenta *Word Sketch* no *Sketch Engine*

The screenshot shows the Word Sketch interface with the search term 'azulejo' and 783 occurrences. The results are organized into four panels:

Panel	Category	Results
1	sintagma preposicional	...de azulejo ...de+os azulejo azulejo em+a substantivo azulejo de+o substantivo ...com azulejo
2	azulejo + adjetivo participial	roubar dos azulejos roubados importar com azulejos importados da Europa furtar tiveram azulejos furtados
3	azulejo + adjetivo	valioso azulejos valiosos antigo de azulejos antigos histórico azulejos históricos
4	azulejo + verbo	perder dos azulejos se perderam ir azulejos foram colorir de azulejos que coloriam as farfaldas e

Fonte: Sketch Engine.

O *Thesaurus* apresenta as palavras utilizadas no corpus que compreendem alguma relação com o termo específico que se deseja pesquisar (Kilgarriff *et al.*, 2004). Esta ferramenta gera uma lista produzida com base no contexto de uso (colocações) em que os termos aparecem no corpus, ou seja, parte do pressuposto de que palavras que são utilizadas em contextos semelhantes no corpus estão no mesmo campo semântico. Nesta lista, são incluídos apenas substantivos, verbos, advérbios e adjetivos (Sketch Engine, s.d.). Para que funcione bem, o *Thesaurus* precisa que o corpus tenha, pelo menos, 100.000 palavras, pois depende das colocações para determinar a relação entre as palavras (Sketch Engine, s.d.). Como exemplo, a Figura 8 apresenta os 10 primeiros termos que apareceram na lista do *Thesaurus* da palavra “azulejo”.

Figura 8 - Resultado da ferramenta *Thesaurus* como lista no *Sketch Engine*

	Lempos	Frequency ?
1	casarão	469 ...
2	patrimônio	270 ...
3	cidade	670 ...
4	situação	214 ...
5	vandalismo	183 ...
6	valor	133 ...
7	roubo	264 ...
8	prefeitura	169 ...
9	tombamento	44 ...
10	ano	147 ...

Fonte: Sketch Engine.

A lista gerada contém informações sobre a frequência com que as palavras aparecem no corpus. Também é possível obter o resultado em forma de gráfico de bolhas, como na Figura 9.

Figura 9 - Resultado da ferramenta *Thesaurus* como gráfico em bolhas no *Sketch Engine*



Fonte: Sketch Engine.

No gráfico em bolhas, o tamanho do círculo depende da frequência da palavra no

corpus, e a distância da palavra do centro é baseada na relação que estas duas palavras apresentam. Se a relação é de proximidade, a palavra estará mais próxima. Se não é tão próxima, ficará mais distante. Esta mesma informação também pode ser visualizada como nuvem de palavras, como apresentado na Figura 10.

Figura 10 - Resultado da ferramenta *Thesaurus* como nuvem de palavras no *Sketch Engine*



Fonte: Sketch Engine.

Na nuvem de palavras, o tamanho de cada palavra indica sua proximidade com a palavra “azulejo”.

A ferramenta *Concordance* é caracterizada pela **busca de palavras nas linhas de concordâncias** e se ocupa da pesquisa por alguma palavra específica, obtendo, como resultado, todos os exemplos da palavra pesquisada juntamente com o contexto em que foi utilizada, como demonstrado na Figura 10, também com a escolha de “azulejo” como exemplo. A análise das linhas de concordância é uma técnica importante uma vez que permite que um grande número de exemplos de uma palavra seja reunido em um único lugar, apresentando seus contextos de uso. Todas as ocorrências da palavra pesquisada aparecem em forma de lista, juntamente com um número definido de caracteres para a esquerda e a direita, o que pode fornecer evidências dos significados mais frequentes de palavras, ou das palavras que mais co-ocorrem, permitindo a identificação de padrões de uso nas ocorrências (Evinson, 2010).

Figura 10 - Resultado da ferramenta *Concordance* no *Sketch Engine*

The screenshot shows the Sketch Engine Concordance tool interface. At the top, the search term 'azulejo' is entered, resulting in 784 matches. The interface includes a search bar, a list of concordance lines, and navigation options like 'Left context', 'KWIC', and 'Right context'. The concordance lines are numbered 1 through 7, and each line shows the search term highlighted in red within a sentence. The tool also displays statistics such as '14,655.58 per million tokens' and '1.5%'.

Line	Context	Search Term
1	doc#0 a da nossa história está sendo depredata e roubada.	Azulejos
2	doc#0 mo urbano.	azulejos
3	doc#0 : azulejos.	azulejos
4	doc#0 tá se tornando uma encomenda frequente porque o valor dos	azulejos
5	doc#0 o valor dos azulejos está aumentando.	azulejos
6	doc#0 trimônio da cidade para desqualificar a propiedade ou vender	azulejos
7	doc#0 ider azulejos no mercado clandestino.	azulejos

Fonte: Sketch Engine.

Para possibilitar que se acesse quais itens co-ocorrem, pode-se acessar as *Collocations*, disponível no ícone da barra de funcionalidades acima das linhas de concordância, representado por uma linha com três círculos. Esta função permite que se acesse termos que co-ocorrem com o adicional das informações sobre a frequência de vezes que eles ocorrem conjuntamente e isoladamente, além dos resultados do *MI-Score*, *T-Score* e *logDice*, medidas estatísticas de associação que identificam a relação colocacional entre palavras (Brezina, 2018). Nas colocações, pode-se ter o interesse de olhar para as palavras que estão logo antes ou logo depois do termo pesquisado, mas também se pode ter o interesse em palavras que estão em colocações mais distantes. No *SkE*, o alcance é de até 5 palavras para a esquerda e 5 palavras para a direita do termo pesquisado. Na Figura 11, são apresentados os resultados dos colocados que estão a 1 palavra para a direita do termo “azulejo”, ordenados por *logDice*.

Figura 11 - Resultado do recurso *Collocations* no *Sketch Engine*

The screenshot shows the Sketch Engine interface for the 'Collocations' tool. At the top, there is a search bar with the text 'corcel20152t4n5r'. Below the search bar, there are several icons for navigation and settings. The main content area displays a table of collocations for the word 'azulejo'. The table has columns for 'Word', 'Cooccurrences', 'Candidates', 'T-score', 'MI', and 'LogDice'. The results are sorted by LogDice score in descending order.

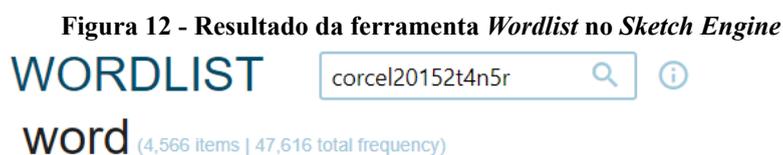
	Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓
1	<input type="checkbox"/> se	50	322	6.40	3.41	10.53 ...
2	<input type="checkbox"/> que	93	1,555	7.28	2.03	10.35 ...
3	<input type="checkbox"/> na	39	200	5.78	3.73	10.34 ...
4	<input type="checkbox"/> foram	39	223	5.72	3.58	10.31 ...
5	<input type="checkbox"/> são	29	182	4.89	3.44	9.94 ...
6	<input type="checkbox"/> .	78	2,148	5.27	1.31	9.77 ...
7	<input type="checkbox"/> valiosos	20	67	4.25	4.35	9.59 ...

Fonte: Sketch Engine.

Nas colunas da esquerda para a direita, tem-se as palavras ou sinais de pontuação que co-ocorrem com a palavra “azulejo” (*Word*), a frequência com que co-ocorrem (*Cooccurrences*), a frequência com que a palavra ou sinal de pontuação ocorre sozinho no corpus (*Candidates*), e as medidas de associação mencionadas anteriormente, com uma seta apontando para baixo na medida *logDice*, o que indica que a lista está ordenada de acordo com esta medida. Esta medida de associação é utilizada para identificar a força de atração entre dois termos (Rychlý, 2008). É baseado apenas na frequência do núcleo (“azulejo”) e do colocado (“se”, “que”, “na”, ...) e na frequência da colocação (“azulejo se”, “azulejo que”, “azulejo na”, ...). Como resultado, tem-se a associação estatística entre o núcleo e a colocação, que indica se aquela é uma “boa colocação” ou não. De acordo com Rychlý (2008), teoricamente, o valor máximo de *logDice* é 14, o que significa que, caso todas as ocorrências de “azulejo” co-ocorressem com “se” e todas as ocorrências de “se” co-ocorressem com “azulejo”, o resultado seria 14. Geralmente, os valores são inferiores a 10. O resultado não é afetado pelo tamanho do corpus, pois compara as frequências relativas de “azulejo se” apenas em relação a “azulejo” e “se” (Rychlý, 2008), sendo considerada uma boa medida para comparar os resultados da associação entre dois termos em um corpus com os resultados desta mesma associação em outros corpora (Sketch Engine, 2015).

A ferramenta denominada *Wordlist*, é uma ferramenta que trata da **exploração de listas de frequência de palavras** e envolve a geração automática de uma lista completa de todas as palavras em um corpus, desde as mais frequentes, que podem aparecer milhões de

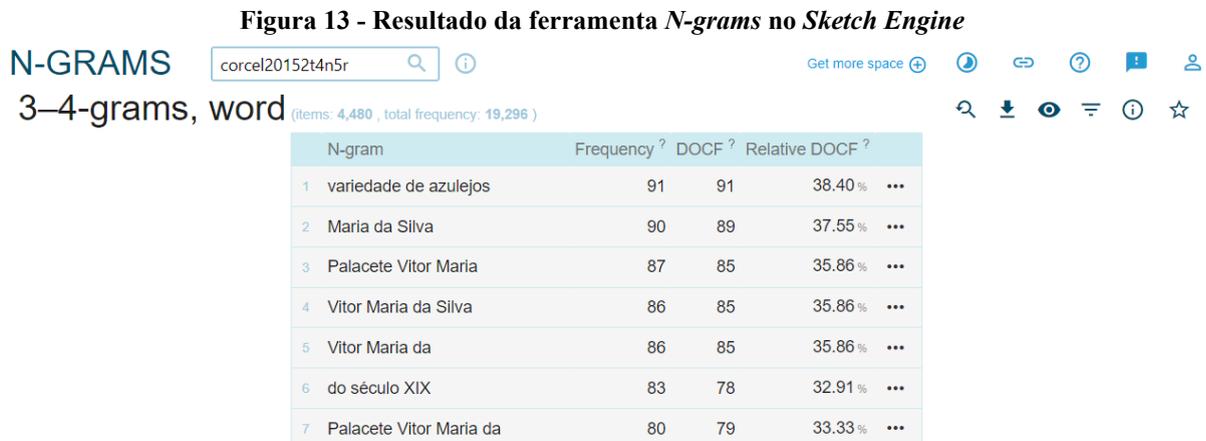
vezes em um corpus, até as mais incomuns, com apenas uma ocorrência (Evinson, 2010), como apresentado na Figura 12. Pode-se ter a intenção de pesquisar as palavras no geral, independente da classe, ou alguma classe específica de palavra, como substantivos, adjetivos, verbos etc.



Fonte: Sketch Engine.

O resultado padrão apresenta a frequência absoluta (*Frequency*) com que o termo aparece no corpus, a frequência por documento (*DOCF*) com que o termo aparece, que indica em quantos documentos do corpus aquela palavra pode ser encontrada, e a frequência relativa por documento (*Relative DOCF*), que apresenta o resultado da *DOCF* convertido em percentual, ou seja, em quantos % dos textos do corpus o termo aparece. A seta para baixo está em *Frequency*, indicando que a lista está ordenada a partir da frequência absoluta, no entanto é possível ordenar a lista automaticamente no *SkE* por qualquer uma das opções. Uma das possibilidades de análise é a comparação das frequências absolutas em que um termo aparece em mais de um corpus, ou pela *DOCF* ou *Relative DOCF*, tudo vai depender do interesse de pesquisa. Ainda é possível extrair a lista de *n-grams*, entendidos como sequências contíguas de 2 a 6 palavras. O padrão de busca do *SkE* para esta ferramenta é de 3 a 4 palavras, fazendo a diferenciação entre letras maiúsculas e minúsculas, sem a inclusão de outros sinais gráficos que não palavras, ou seja, sem a inclusão de pontuação ou de símbolos como “#”, “@”, “%”, etc. Como exemplo, a Figura 13 traz os resultados por uma busca padrão no corpus. Vale ressaltar que é possível alterar as configurações com facilidade caso se

deseje buscar uma sequência mais longa ou mais curta, bem como incluir elementos não alfa-numéricos.



Fonte: Sketch Engine.

A ordem padrão dos *n-grams* é feita com base na frequência absoluta, e não há possibilidade de alteração automática desta ordem.

Todas as ferramentas apresentadas até o momento são realizadas apenas com um corpus por vez, e suas frequências e medidas de associação se baseiam no que está contido exclusivamente dentro do corpus em questão, sem influência de qualquer fator externo ao corpus. A ferramenta *Keywords*, definida como a exploração de listas de palavras-chave, é baseada na comparação estatística de determinados termos entre mais de um corpus (Evinson, 2010), colocados em contraste para se obter resultados que indiquem o que ambos os corpora apresentam de diferenças entre si. Para que a ferramenta funcione, são necessários, pelo menos, dois corpora, sendo o corpus cujas palavras-chave se deseja explorar denominado **corpus de estudo**, e o corpus utilizado para comparação denominado **corpus de referência**²³. Nesta ferramenta, são comparadas as frequências normalizadas (frequências relativas) com que a palavra aparece no corpus de estudo e no corpus de referência. As palavras-chave não são necessariamente as palavras mais frequentes no corpus, mas sim aquelas que se mostram chave (típicas daquele corpus de estudo) em relação a um corpus de referência, e a sua identificação é feita automaticamente pelo *software*, como apresentado na Figura 14. Por exemplo, *em que se diferem dois níveis de proficiência no que diz respeito à escolha de palavras?* Esta ferramenta permite que se obtenha resultados neste sentido. Para que se

²³ Uma descrição mais detalhada dos tipos de corpora será feita na seção **2.1.4 Categorização e discussão terminológica: problemática em “corpora de aprendizes”**.

chegue a estes resultados, basta que se utilize o mesmo corpus de referência na busca pelas palavras-chave nos diferentes corpora de estudo e, depois, que se analise as listas resultantes destas comparações.

Figura 14 - Resultado da ferramenta *Keywords* no *Sketch Engine*

	Word	Frequency per million?		Document frequency?		Relative DOCF?		
		Focus	Reference	Focus	Reference	Focus	Reference	
1	casarões	8,972.80	0.85	200	265	84.39 %	0.02 %	...
2	azulejos	15,515.47	2.51	230	810	97.05 %	0.07 %	...
3	bibiana	3,832.13	0.08	169	36	71.31 %	< 0.01 %	...
4	cambará	3,626.51	0.59	160	134	67.51 %	0.01 %	...
5	atenciosamente	2,149.73	0.23	115	105	48.52 %	< 0.01 %	...
6	palacete	2,243.20	1.12	103	325	43.46 %	0.03 %	...
7	belem	785.12	0.07	22	31	9.28 %	< 0.01 %	...

Fonte: Sketch Engine.

A ordem da lista aqui é feita com base na relevância que aquelas palavras apresentam em relação ao corpus de referência, com os resultados de ambos. Por padrão, as frequências absolutas são normalizadas por milhão de palavras, tanto no corpus de estudo, quanto no corpus de referência, e é a partir disso que as palavras são automaticamente ranqueadas como mais ou menos relevantes pelo *SkE*. A lista de palavras-chave pode incluir tanto palavras significativamente frequentes (palavras-chave positivas), quanto significativamente raras (palavras-chave negativas). É comum, nesses casos, que se utilize algum cálculo matemático de chavidade, como o *Log-Likelihood* (LL), para averiguar se a frequência dos termos pode ser considerada estatisticamente relevante no corpus de estudo em relação ao corpus de referência.

Ainda é possível que uma técnica seja o caminho para a outra, ou seja, que a partir dos resultados obtidos na *wordlist* ou na lista de *keywords*, opte-se por fazer a busca de termos específicos nas ferramentas *Word Sketch*, *Thesaurus* e *Concordance*.

Segundo Hunston (2010), “alegações sobre a frequência são sempre relativas” (Hunston, 2010, p. 164), portanto, para possibilitar a comparação de corpora de tamanhos diferentes, os valores absolutos das ocorrências passam por um processo de normalização, que consiste em ajustar as frequências brutas das palavras nos corpora que estão sendo comparados de forma que possam ser representadas de acordo com uma mesma escala, como mil ou um milhão de palavras, obtendo-se as frequências relativas. Isso ocorre para evitar a representação excessiva de qualquer fenômeno linguístico (Cushing, 2022) e é muito importante quando a análise envolve corpora de tamanhos muito diferentes. Focalizar apenas as frequências brutas pode distorcer os resultados. A normalização pode ser realizada pela quantidade de palavras, averiguando quantas vezes a palavra teria aparecido nos corpora caso fossem iguais em quantidade de palavras, ou pela quantidade de textos, averiguando o percentual de textos em que determinado termo foi usado em cada corpus.

2.1.4 Categorização e discussão terminológica: problemática em “corpora de aprendizes”

De acordo com a taxonomia de corpora proposta por Viana (2011), um corpus pode ser classificado de acordo com critérios distintos, que o autor denominou: **abrangência**, **meio**, **tempo (número de períodos históricos e relação com a atualidade)**, **renovação**, **línguas (número e informantes²⁴)** e **emprego (estudo e referência)**.

A **abrangência** diz respeito ao grau de especificidade dos textos nele contidos, apresentando duas categorias: **geral**, quando o propósito do corpus é o de englobar textos de diversos gêneros, apresentando temáticas variadas; e **especializado**, quando os textos são de gêneros específicos, podendo, ou não, ser da mesma temática. Quanto ao **meio**, um corpus pode ser classificado em **oral** ou **escrito**, ou seja, categorizado de acordo com o meio em que os textos são veiculados. O fator **tempo** pode ser classificado de duas formas diferentes: pela **quantidade de períodos históricos representados** no corpus ou pela sua **relação com a atualidade**. A primeira delas, respectivamente, pode ser dividida entre **sincrônico**, quando os textos do corpus estão datados em um espaço de tempo determinado, ou **diacrônico**, quando o objetivo do corpus é documentar a linguagem utilizada em diversos períodos de tempo. A

²⁴ Viana (2011) utiliza “produtores” no lugar de informantes. Na seção **2.1.6 Corpora de PLA e contribuições para o ensino e a avaliação** são apresentados dois projetos de compilação de corpora de PLA que utilizam o termo “informantes”, de modo que se optou pela padronização com a utilização deste termo.

segunda diz respeito ao período específico em que os textos presentes no corpus foram produzidos, podendo ser classificado como **contemporâneo** ou **histórico**. A classificação quanto à **renovação** apoia-se na atualização do corpus: se o corpus é constantemente atualizado e alterado, ele é considerado **dinâmico**; caso contrário, o corpus é denominado **estático**. Outro critério para a diferenciação entre corpora é a **língua**, dividido entre **número de línguas** e seus **informantes**. Quanto ao número de línguas, um corpus pode ser **monolíngue**, significando que o corpus engloba apenas uma língua, ou **multilíngue**, abarcando mais de uma língua. Os corpora multilíngue podem ser **paralelos**, possuindo textos escritos em determinada língua com suas respectivas traduções para uma outra língua (e vice-versa), ou **não paralelos**, quando isto não ocorre. Os corpora paralelos também podem ser divididos entre **alinhados** ou **não alinhados** a depender da existência (ou não) de correspondência exata entre palavras, frases e/ou parágrafos no texto em ambas as línguas. Já os **informantes** são os falantes dessas línguas, ou seja, se o corpus é composto por textos produzidos por falantes nativos ou não. O autor também sugere a classificação de acordo com o **emprego** que se pretende fazer do corpus, ou seja, sua função na pesquisa. Por esse critério, um corpus pode ser classificado como corpus de **estudo**, que é um corpus cuja linguagem dos textos nele contido se deseja observar e descrever, ou corpus de **referência**, utilizado para contrastar com um corpus de estudo, de forma a fornecer uma norma com a qual se fará comparação das frequências do corpus de estudo. A classificação dos corpora, segundo Viana (2011), pode ser observada no Quadro 4.

Quadro 4 - Classificação de Corpora

Critérios		Corpus
Abrangência		Geral Especializado
Meio		Oral Escrito
Tempo	Nº de períodos históricos	Sincrônico Diacrônico
	Relação com a atualidade	Contemporâneo Histórico
Renovação		Dinâmico Estático

Línguas	Número	Monolíngue		
		Multilíngue	Paralelos	Alinhados Não alinhados
	Não paralelos			
	Informantes	De primeira língua (L1) De segunda língua (L2) De língua estrangeira		
Emprego		Estudo Referência		

Fonte: adaptado de Viana, 2011.

Segundo Sardinha (2000), quando um corpus não é composto por textos produzidos por falantes nativos, deve-se qualificá-los como um corpus de aprendizes (*learner corpus*). Embora a pesquisa na área da LC relacionada ao ensino de língua adicional tenha iniciado com a compilação de corpora de falantes nativos, a utilização de corpora de aprendizes começou, nos anos noventa, a fazer parte desse campo de estudo, tornando viável a identificação de padrões gerais de uso da língua produzida por esses falantes (Mauranen, 2011), podendo ser considerada uma revolução no ensino de idiomas por permitir que se deslocasse o foco dos corpora de falantes nativos (Mauranen, 2011). Estes corpora podem ser classificados como coleções eletrônicas de textos produzido por aprendizes estrangeiros de uma língua “que não é nem a sua primeira língua, nem uma língua adicional institucionalizada no país em que vivem” (Granger, 2008, p. 259). A princípio, fariam parte do grupo “aprendizes”, segundo Granger (2017), tanto alunos que aprendem o idioma em um país onde têm pouca exposição fora do contexto de sala de aula, como língua estrangeira, de acordo com a autora, ou até mesmo que adquirem o idioma em algum país onde esse idioma é a língua predominante de comunicação, como segunda língua.

Os dados obtidos para a compilação desses corpora são, prototipicamente, produzidos a partir de tarefas abertas, que permitem que os falantes escolham suas próprias palavras (Callies; Götz, 2015). De acordo com Granger (2009), entre as vantagens da pesquisa com corpora de aprendizes, estão suas contribuições para um maior detalhamento sobre proficiência linguística, trazendo para o campo do ensino e aprendizagem de língua adicional uma base empírica muito ampla para aplicações pedagógicas devido à possibilidade de acesso a uma maior quantidade de dados em relação a outras metodologias. A própria natureza dos

dados, ou seja, o “discurso contínuo contextualizado” (Granger, 2009, p. 16) e “representativo do uso contextualizado da língua dos aprendizes” (Callies; Götz, 2015, p. 3), oferece aos pesquisadores uma gama muito vasta de possibilidades, permitindo uma visão mais diversificada da língua dos aprendizes, de modo que estes corpora podem ser até mais efetivos no ensino de PLA do que corpora de falantes nativos (Sommer-Farias *et al.*, 2023).

Jenkins (2005) define o termo “língua de aprendizes” como um efeito da crença na ideia da existência de pureza linguística. Isto significa, segundo a autora, que os falantes nativos são tidos como os guardiões da língua e a variedade nativa escolarizada precisa ser protegida das corrupções provocadas por falantes não-nativos e, sob essa perspectiva, o contato com outras línguas torna a língua impura e os desvios da norma são tomados como erros, como se a língua falada por não-nativos fosse ilegítima (Jenkins, 2005). Nesse sentido, há autores, como Mauranen (2011), que não concordam em denominar qualquer corpus produzido por falantes não nativos como corpus de aprendizes. Mauranen (2011) diferencia corpora de aprendizes de corpora de língua franca, por exemplo, pois, segundo a autora, estes “diferem [um do outro] de forma significativa nos aspectos sociais e interacionais” (Mauranen, 2011, p. 155). Para Mauranen (2011), aprendizes em sala de aula compartilham, muitas vezes, a mesma língua materna, podem ter expectativas semelhantes em relação ao “idioma-alvo” e possuir o desejo da apropriação de aspectos culturais relacionados à língua, além de poderem ter um falante nativo como interlocutor projetado. De acordo com a autora, no caso da língua franca, o cenário é diferente, pois falantes de uma língua franca não necessariamente compartilham a primeira língua, a língua de comunicação é escolhida por conveniência, e os falantes, muitas vezes, pouco sabem sobre a bagagem cultural uns dos outros nem sobre sua proximidade com a língua utilizada. Por mais que a autora reconheça que os indivíduos podem alternar entre os dois papéis (de falantes e de estudantes) e, sob aspectos cognitivos, ambos medirem esforços para se fazer entender e produzir sentido, suas diferentes naturezas fazem com que corpora compilados nestes dois contextos distintos não possam ser denominados da mesma forma (Mauranen, 2011).

Nesta ocasião, faz-se relevante ressaltar que as expectativas dos falantes em relação ao “idioma-alvo” podem ser muito variadas, bem como as motivações que levam alguém a aprender uma língua. Além disso, não há garantia de que todos os estudantes em uma sala de aula terão a mesma língua materna. Em uma mesma sala de aula, é possível encontrar tanto aqueles que se encaixam na descrição de Mauranen (2011) do que seria um aprendiz, ou seja, um indivíduo que deseja se apropriar dos aspectos culturais relacionados à língua e tem o

falante nativo como interlocutor desejado, quanto aqueles que pretendem aprender a usar a língua para se comunicar com outros indivíduos que também falam aquela língua, independente da sua nacionalidade e cultura. Em suma, os perfis descritos por Mauranen (2011) do que seria um aprendiz e um falante de uma língua não são estáveis, ou seja, dois indivíduos pertencentes a um destes dois grupos não necessariamente carregam consigo a mesma língua materna nem as mesmas ambições, objetivos e expectativas relacionadas à língua. É possível, também, que um mesmo indivíduo transite entre os dois grupos dependendo de suas intenções e do contexto de uso da língua em que se encontra.

Há, também, casos em que, com os dados coletados, não é possível fazer a distinção entre falantes de uma língua adicional e pessoas que se encaixam na categoria de aprendizes dessa língua. Nessa ocasião, Gablasova, Brezina e McEney (2017) generalizam o falante como falante de L2, denominando o corpus compilado sob essas circunstâncias como corpus de L2, termo bastante utilizado na área (Staples; Fernández, 2019). Além de corpus de L2, encontra-se o termo corpus de língua estrangeira (Mauranen, 2011; Meunier, 2011) e, com menos frequência, corpus de língua adicional (Schepers, 2021), utilizados em contextos semelhantes aos de corpora de aprendizes, muitas vezes como sinônimos. Embora não seja comum que a escolha terminológica seja justificada em publicações da área, entende-se que ela parte da visão de língua à qual o pesquisador se filia.

Como previamente apresentado, o foco de pesquisa pode variar muito de projeto para projeto, então é comum que, além da terminologia, o *design* do corpus também varie (Tono, 2003). Em uma subcategoria própria desses corpora, com foco voltado para a área de avaliação de proficiência linguística, encontram-se os corpora compilados a partir da produção de examinandos em contexto avaliativo (Gablasova, 2021; Götz, 2022). Gablasova (2021) aponta para a utilidade desses corpora para a pesquisa de aspectos diretamente aplicáveis à avaliação de proficiência, como a investigação de usos típicos da língua em diferentes níveis de proficiência (Cushing, 2022), refinando os descritores desses níveis (Götz, 2022), tendo o potencial de aumentar a transparência, a validade e a confiabilidade na avaliação (Callies; Götz, 2015).

Para classificar o corpus deste trabalho com base na categorização apresentada por Viana (2011) e na discussão terminológica realizada acima, faz-se relevante ressaltar que o Celpe-Bras “não tem como base, para a definição do nível mais alto de proficiência, o falante nativo idealizado” (Brasil, 2020, p. 27) e é realizado por pessoas interessadas em comprovar sua proficiência em português brasileiro, não necessariamente alunos de português, mas

falantes estrangeiros da língua no geral. Por esse motivo, optou-se, nesta pesquisa, pela categorização do corpus de estudo como especializado, escrito, sincrônico e contemporâneo, de língua adicional, com o adendo de ter sido compilado a partir da produção de examinandos em contexto de avaliação de proficiência. É importante ressaltar que, apesar da escolha terminológica por “corpus de língua adicional”, são também utilizados para embasar as análises realizadas neste trabalho referenciais teóricos comumente associados aos outros termos apresentados, sobretudo ao termo “corpus de aprendizes”, cujas descrições são muito mais robustas e pesquisas relacionadas são muito mais consolidadas do que no caso dos demais.

2.1.5 Corpora de línguas adicionais e avaliação de proficiência

Desde a virada do milênio, percebe-se o aumento do interesse no uso de LC para informar o desenvolvimento e a validação de exames de língua (Cushing, 2017; 2022). De acordo com Barker (2010), havia pouca discussão sobre a utilização de corpora no campo de estudos sobre avaliação até a década de 1990, até que Alderson (1996) chamou, pela primeira vez, a atenção dos pesquisadores para possibilidades de aplicações de pesquisas com corpora nesse âmbito (Cushing, 2017). No princípio, os estudos da área versavam exclusivamente sobre o uso de corpora de falantes nativos, e foi apenas depois de comprovados os benefícios de sua utilização para avaliações de línguas adicionais que corpora de línguas adicionais começaram, finalmente, a ser levados em consideração nas pesquisas (Barker, 2010). Barker (2010) afirma que corpora de línguas adicionais podem informar e trazer contribuições a vários estágios do ciclo de vida de um exame, dentre os quais Cushing (2022) lista quatro: **(1) definição do construto; (2) validação do instrumento avaliativo por meio de uma abordagem baseada em argumentos; (3) descrição linguística dos níveis de proficiência; e (4) avaliação automatizada e ferramentas de *feedback*.**

Fundamentalmente, a LC é considerada uma disciplina que consiste em dados de frequência sobre como algum fenômeno linguístico é distribuído em um corpus (Gries, 2010) e como diferentes fenômenos linguísticos são combinados uns com os outros (Cushing, 2022). Cushing (2017) afirma que ter familiaridade com as vantagens e as limitações dos possíveis usos de ferramentas desenvolvidas para análise de corpora é indispensável para pesquisadores da área de avaliação, que devem insistir em demonstrar a confiabilidade e a **validação do instrumento avaliativo por meio de uma abordagem baseada em argumentos**. Os

pesquisadores devem, segundo a autora, utilizar as análises de corpora como um suporte para a **definição do construto**, triangulando-as com demais dados relacionados ao exame pesquisado. Um importante avanço no tocante à **definição do construto**, possível graças à capacidade da realização de análises comparativas do uso da língua (Cushing, 2017), é a conclusão de que gramática e léxico não podem ser considerados como subconstrutos distintos da proficiência linguística, que está relacionada à habilidade de combinação efetiva de palavras (Paquot, 2019). Ou seja, gramática e léxico são indissociáveis e operam conjuntamente na distinção entre níveis de proficiência (Cushing, 2022).

Em relação à **descrição linguística dos níveis de proficiência**, a utilização de corpora de línguas adicionais permite a comparação do desempenho linguístico em diferentes tarefas sob uma perspectiva longitudinal (Barker, 2010), bem como em diferentes contextos, gêneros ou até mesmo entre diferentes níveis de proficiência (Cushing, 2017). Callies e Götz (2015) argumentam sobre a importância da operacionalização dos níveis de proficiência baseada no uso da língua, ou seja, centrada nas produções de alunos e/ou examinandos, e orientada pelos dados obtidos a partir das análises, questionando crenças previamente mantidas sobre a estrutura, as funções e o uso da língua (Cushing, 2017). Se os corpora escolhidos e amostrados são representativos do que se deseja observar, as análises de comparação dos recursos linguísticos em diferentes grupos são uma fonte valiosa na validação dos exames, afirma Cushing (2017).

Essencialmente, os dados do corpus revelam o que as pessoas realmente fazem com a língua na vida real, em contextos de aprendizado e avaliação. Portanto, os corpora podem ser usados para informar nossa compreensão do discurso [...] em vários domínios, bem como sugerir aspectos adequados da língua a serem testados ou evitados e, além disso, mostrar o que diferencia falantes de diferentes níveis de proficiência entre si [...] ²⁵ (Barker, 2010, p. 638).

O quarto tópico trazido por Cushing (2022) é a **avaliação automatizada e ferramentas de *feedback***, que não é um consenso entre avaliadores e pesquisadores, afirma Barker (2010). Há mais de uma década, há pesquisadores desenvolvendo maneiras de avaliar produções escritas e orais de maneira automatizada com o apoio de *softwares*, geralmente por meio de comparação do texto com um conjunto padronizado de exemplos de desempenho em determinados níveis (Barker, 2010). Barker é um dos pesquisadores que reconhecia o papel da

²⁵ No original: *Essentially, corpus data reveal what people actually do with language in real life, learning and testing contexts. Corpora therefore can be used to inform our understanding of learner and expert discourse in various domains, as well as suggesting suitable aspects of language to test or to avoid and, additionally, to show what differentiates learners of different proficiency levels from each other and from an expert user of the language* (Barker, 2010, p. 638).

tecnologia na avaliação de língua, mas designava à tecnologia um papel complementar à avaliação humana, não como uma possibilidade de substituição à avaliação manual. Um exemplo concreto desse tipo de avaliação é o TOEFL iBT que, além de oferecer questões-simulado avaliadas por Inteligência Artificial (IA) em seu site²⁶, tem “todas as respostas [...] revisadas por IA e humanos para obter o mais alto nível de precisão” (EDUCATIONAL TESTING SERVICE, s.d.) em sua avaliação oficial.

Além dos quatro apontamentos de Cushing (2022), Barker (2010) faz considerações a respeito do **(5) *design do exame***. O autor afirma ser possível utilizar corpora de línguas adicionais para acessar a influência de variáveis distintas, como variáveis demográficas, variáveis relacionadas ao formato do exame e variáveis associadas ao ambiente de aprendizagem (Barker, 2010), a depender da quantidade de informações que se tem sobre o examinando além da sua produção escrita/oral, por exemplo: idade, contexto linguístico, país de origem, contexto de aprendizado da língua do exame etc. Para além de fatores externos ao exame, as próprias produções servem para “mostrar quais padrões são comuns/menos frequentes em determinados níveis” (Barker, 2010, p. 640) de proficiência, possibilitando que esses padrões sejam incluídos nas tarefas do exame, sendo usados também para apoiar ou refutar especulações dos elaboradores sobre o desempenho dos examinandos.

Nagasawa (2023) alerta sobre a necessidade de se tomar cuidado com as interpretações a respeito das análises de dados provenientes de pesquisas realizadas com corpora advindos de exames de proficiência, sendo importante que os pesquisadores considerem seu conhecimento acerca do construto do exame pesquisado. A autora sugere que, não tão somente a análise, mas a própria escolha dos fenômenos linguísticos a serem analisados deve ser pautada nos usos da língua explicitados como relevantes de acordo com o construto do exame em questão para que seja significativa. Há, neste sentido, uma necessidade de, em avaliações holísticas, como é a do Celpe-Bras, analisar medidas linguísticas superficiais de maneira cautelosa, sempre ponderando a relação entre as medidas analisadas e o cumprimento da interação comunicativa (Nagasawa, 2023).

²⁶ <https://www.ets.org/toefl/test-takers/ibt/prepare/practice-tests.html#accordion-31f3d4f46f-item-cd4759b094>. Acesso em: 09 de janeiro de 2024.

2.1.6 Corpora de PLA e contribuições para o ensino e a avaliação

Nos últimos anos, houve um aumento significativo no uso de corpora de línguas adicionais em pesquisas voltadas não somente para o ensino (Sommer-Farias *et al.*, 2023), como também para avaliações de proficiência (Wisniewski, 2017). Houve, segundo Cushing (2020), um fortalecimento, na última década, da relação entre LC e Avaliação de línguas. Dados provenientes de corpora de línguas adicionais podem revelar características importantes sobre níveis de proficiência (Mendes *et al.*, 2016). Estudos sobre o Quadro Europeu Comum de Referência (CEFR), por exemplo, buscam uma classificação confiável e transparente do idioma dos examinandos e potenciais examinandos (estudantes) para relacionar ao quadro. Resultados advindos destes estudos são úteis não apenas para que se tenha um maior entendimento do significado das classificações de proficiência em um nível empírico (Callies; Díez-Bedmar; Zaytseva, 2014; Barker; Salamoura; Saville; 2015; Callies; Götz, 2015), mas também para que se possa aumentar a transparência e a coerência de testes de proficiência relacionados ao CEFR (Wisniewski, 2017).

Relacionado ao CEFR, já foram realizados alguns estudos que apresentam uma visão geral das características de critérios na sintaxe e no léxico (Hawkins, Filipović, 2012), que dão enfoque a marcadores epistêmicos (Gablasova *et al.*, 2017), que sugerem uma tendência do nível B2 de ser indicativo de uma série de desenvolvimentos na linguagem (Thewissen, 2013; Díez-Bedmar, 2015; Chen; Baker, 2016), que olham para o tamanho do texto como possível característica que distingue as produções com classificação mais alta e mais baixa (Prodeau; Lopez; Véronique, 2012; Vyatkina, 2012; Gyllstad *et al.*, 2014), que buscam pacotes lexicais padrão entre os níveis de proficiência (Callies; Díez-Bedmar; Zaytseva, 2014), entre outros. Muitos destes estudos foram possíveis graças à existência de corpora disponíveis para pesquisa, como o Cambridge Learner Corpus (CLC) (Saville; Hawkey, 2010; Harrison, Barker, 2015), o Norwegian ASK corpus (Carlsen, 2013), o Italian RLD project (Spinelli; Parizzi, 2010).

No caso do português, também há alguns projetos de compilação de corpora de textos escritos por alunos ou examinandos, que serão descritos a seguir. O Corpus de Aquisição de L2 (CAL2) (CLUL, 2012), compilado pelo grupo de pesquisa no Centro da Universidade NOVA de Lisboa (CLUNL) é um corpus organizado por língua materna e por nível de proficiência, havendo três níveis: iniciante, intermediário e avançado. A parte escrita do CAL2 é composta por 1.607 textos produzidos por 1.184 informantes, totalizando 314.817

palavras, e a compilação do corpus fez parte de um projeto cujo interesse era averiguar propriedades gramaticais na aquisição do português. O Corpus de Português Língua Estrangeira/Língua Segunda (COPLE2) (Mendes *et al.*, 2016), compilado por um grupo de pesquisa no Centro de Linguística da Faculdade de Letras da Universidade de Lisboa (CLUL/FLUL), contém 966 textos produzidos em contextos avaliativos, sendo o primeiro deles os testes aplicados em sala de aula no Instituto de Cultura e Língua Portuguesa (ICLP), e o segundo os exames realizados no Centro de Avaliação de Português Língua Estrangeira (CAPLE). Os textos estão organizados em 5 níveis de proficiência relacionados ao CEFR (A1, A2, B1, B2 e C1), contando 14 línguas maternas.

Há, também, o projeto de Recolha de dados de Aprendizagem do Português Língua Estrangeira (Centro de Linguística da Universidade de Lisboa, s.d.), compilado em parceria entre o Instituto Camões e o Centro de Linguística da Universidade de Lisboa com objetivo de ter impacto positivo na formação de professores e na preparação de material didático. O corpus é composto por, aproximadamente, 70.500 palavras, e conta com 470 textos escritos por 397 alunos de PLA entre os níveis A1 e C1, totalizando 28 diferentes línguas maternas. Cada texto tem, em média, 150 palavras. Juntamente com os textos, os pesquisadores coletaram dados pessoais (como idade, nacionalidade etc.) e sociolinguísticos (como língua materna, língua de escolarização, tempo estudando português etc). O Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) (Martins *et al.*, 2019) também foi compilado no Centro de Estudos de Linguística Geral e Aplicada da Universidade de Coimbra (CELGA) a partir de textos de alunos de diferentes níveis (A1-C1) em cursos de PLA da Faculdade de Letras da Universidade de Coimbra (FLUC), totalizando 458 sujeitos, falantes de 39 línguas maternas diferentes. Os dados deste corpus também abrangem a idade e gênero dos informantes. O intuito da criação deste corpus era o de dar subsídios para jovens pesquisadores interessados na área de aquisição e aprendizagem de PLA.

O projeto Multilingual Academic Corpus of Assignments - Writing and Speech (Macaws) (Staples *et al.*, 2019), assim como alguns dos projetos apresentados acima, foi compilado a partir de tarefas de produção dadas em sala de aula. Este projeto ainda está em andamento e é atualizado a cada semestre com mais textos. O corpus conta com 536,168 palavras e foi compilado com o intuito de servir para o uso de professores e alunos. O objetivo da compilação deste corpus é permitir que se tenha acesso a dados empíricos da língua em contexto, auxiliando alunos a desenvolverem sua proficiência linguística em cursos de PLA (Sommer-Farias *et al.*, 2023).

O Corpus do Celpe-Bras (CorCel) (Schoffen *et al.*, em preparação) é um projeto em andamento desenvolvido pelo Grupo Avalia, da Universidade Federal do Rio Grande do Sul, que visa a compilação de um corpus de 15.315 textos produzidos por examinandos nas edições de 2015/1, 2016/1, 2016/2 e 2017/1 do Celpe-Bras (Schoffen *et al.*, 2024). Cada edição contém 4 tarefas, e, para cada tarefa, há seis subcorpora de textos avaliados com as notas 0, 1, 2, 3, 4 e 5 por ambos os avaliadores. Com o objetivo de que cada subcorpus acomodasse uma quantidade similar de textos, foram selecionados, aleatoriamente, 200 textos²⁷ de cada nota em cada tarefa de cada edição, e, quando alguma das notas não apresentava uma quantidade igual ou maior que 200, foram incluídos todos os textos avaliados com aquela nota, fazendo com que, nesses casos, o referido subcorpus seja composto por um número menor de textos. Os textos foram disponibilizados pelo Inep em forma de cópia digitalizada do manuscrito, já foram digitados e estão passando neste momento pelo processo de revisão da digitação. Pretende-se, também, que os textos sejam anotados. Juntamente com os textos de cada edição, foram cedidas planilhas de Excel contendo as notas que cada texto recebeu de cada avaliador em cada tarefa, assim como a nota final da PE. As produções escritas foram cedidas sem nenhuma informação extra-textual capaz de identificar o examinando e sem qualquer informação sobre os avaliadores envolvidos nas avaliações. Fazem parte dos dados disponibilizados, também um documento com a Resposta Esperada em cada uma delas. O fato de não se ter informações a respeito dos examinandos impossibilita que se utilize o termo “corpus de aprendizes” para classificar o conjunto de textos compilados a partir desse material, já que, apesar de não terem o português como primeira língua de socialização, não é possível saber se esses examinandos são estudantes de português ou não, pois o fato de não serem falantes nativos não os torna automaticamente aprendizes, apenas falantes de uma LA.

O interesse na descrição dos níveis de proficiência do exame Celpe-Bras com base em produções escritas de alunos ou examinandos não é novidade na área de estudos sobre avaliação, como já testemunham os estudos de Sidi, 2002; Schoffen, 2009; Gomes, 2009; Evers, 2013; Dell’Isola; Tosatti, 2015; Sirianni, 2016; 2020; Kunrath, 2019; Mendel, 2019; Divino, 2021; Tosatti, 2021; Vicentini, 2022; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023). Pode-se dizer que as pesquisas nesse nicho são divididas em dois grandes grupos: um dos grupos é caracterizado pela adoção de metodologias qualitativas de análise (Sidi, 2002; Schoffen, 2009; Gomes, 2009; Dell’Isola; Tosatti, 2015; Sirianni, 2016; 2020;

²⁷ Para a seleção aleatória dos textos, foi utilizado o site: <https://www.invertexto.com/lista-aleatoria>

Mendel, 2019; Tosatti, 2021; Vicentini, 2022), enquanto o outro se utiliza de metodologias quantitativas (Evers, 2013; Kunrath, 2019; Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023).

Os estudos qualitativos focalizaram suas análises, majoritariamente, na maneira como os elementos constitutivos do enunciado são considerados no momento da avaliação (Sirianni, 2016), triangulando essas informações com os parâmetros de avaliação e o desempenho nas produções escritas em resposta às tarefas para, dessa forma, verificar aspectos determinantes da complexidade (Gomes, 2009). Alguns dos trabalhos se debruçam sobre a recuperação de informações do texto de insumo, ou seja, sobre a integração das habilidades de compreensão e produção escrita (Schoffen, 2009; Mendel, 2019; Vicentini, 2022). Há, também, foco na coesão textual (Sidi, 2002; Sirianni, 2020), com o objetivo de descrever mais detalhadamente os pontos de corte entre os níveis de proficiência (Schoffen, 2009; Sirianni, 2020) ou realizar uma caracterização mais minuciosa dos diferentes perfis dentro de um mesmo nível (Sidi, 2002). Os trabalhos também olham para a qualidade da produção com foco na estruturação do gênero, observando a variedade dos níveis de formalidade (Dell'Isola; Tosatti, 2015) e da construção da interlocução (Tosatti, 2021), e na noção de autoria pressuposta nos parâmetros de avaliação (Mendel, 2019).

As pesquisas quantitativas objetivaram, por meio de ferramentas de análise da LC, detectar padrões lexicais (Evers, 2013; Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023) e coesivos (Evers, 2013; Kunrath, 2019; Sostruznik, 2023), bem como informações referentes à extensão dos textos e riqueza lexical (Evers, 2013; Divino, 2021; Hanauer, 2023) que pudessem melhor descrever os níveis de proficiência. Assim como nos estudos qualitativos, os estudos quantitativos também olharam para a recuperação de informações do texto de insumo (Kunrath, 2019; Divino, 2021; Hanauer, 2023). Os *softwares* utilizados nesses estudos foram o *Coh-Matrix-Port* (Evers, 2023; Kunrath, 2019) e o *SkE* (Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023).

De modo geral, os trabalhos trouxeram discussões mais aprofundadas acerca dos níveis de proficiência, propondo novos parâmetros de avaliação (Schoffen, 2009), transformações dos parâmetros gerais do exame para parâmetros mais específicos (Sidi, 2002; Sirianni, 2016; Vicentini, 2022), e até mesmo uma progressão de níveis baseada em características distintas que emergiram das análises (Kunrath, 2019). Gomes (2009) também propõe orientações para que se construa tarefas que possam conferir maior validade ao exame, afirmando que, quanto mais estiverem explicitados os propósitos e os papéis dos

interlocutores, maiores as chances de haver coerência dos parâmetros de avaliação em relação ao desempenho dos examinandos.

Dell'Isola e Tosatti (2015) orientam os examinandos a considerarem a diversidade de situações comunicativas, articulando os elementos do enunciado de forma a levar em conta o propósito e os movimentos retóricos do gênero proposto. Schoffen (2009), Kunrath (2019), Mendel (2019), Divino (2021) e Hanauer (2023) argumentam sobre a importância da recuperação do conteúdo informacional, concluindo que a demonstração da compreensão se dá pela escrita (Mendel, 2019). Vicentini (2022) propõe uma reflexão acerca da possibilidade de diferenciação do uso do conteúdo informacional em tarefas que integram compreensão oral e produção escrita das tarefas que integram leitura e escrita, sob hipótese de o uso dessas informações não ser o mesmo nos dois tipos de tarefas. Além da recuperação das informações do texto de insumo ser feita mais adequadamente nos níveis mais proficientes, o repertório próprio do examinando também é mobilizado nesses níveis (Mendel, 2019), articulando recursos linguísticos que não são dados no material de insumo (Divino, 2021; Hanauer, 2023). Isso diz respeito, também, à adequação ao gênero e à articulação da relação de interlocução propostos, realizados mais satisfatoriamente à medida que o nível de proficiência aumenta (Sirianni, 2016; Mendel, 2019; Divino, 2021; Hanauer, 2023; Schoffen; Divino, 2023). A mesma progressão é vista no que diz respeito ao número de conjunções (Sostruznik, 2023) e de recursos linguístico-discursivo utilizados para dar consistência e coesão ao texto (Kunrath, 2019).

Os resultados dos estudos contribuem para a discussão sobre as diferenças entre os níveis de proficiência e sobre pontos de corte entre esses níveis (Sirianni, 2020; Evers, 2013) e visam ter implicações para o ensino e a aprendizagem de PLA (Sidi, 2002; Kunrath; 2019, Divino, 2021; Vicentini, 2022), colaborando para a qualificação dos professores (Sirianni, 2016) devido ao seu alto potencial de incentivar o desenvolvimento de atividades pedagógicas de leitura e escrita (Mendel, 2019; Hanauer, 2023; Sostruznik, 2023), especialmente em contexto de ensino preparatório ao exame (Kunrath, 2019; Mendel, 2019). Nesse sentido, Schoffen e Divino (2023) ressaltam a importância do ensino baseado em estudos de corpus, incentivando o planejamento pedagógico ancorado em dados empíricos. Por demonstrarem a operacionalização dos parâmetros de avaliação da PE, os resultados desses estudos podem ter, também, implicações para a elaboração e avaliação do exame (Hanauer, 2023), fornecendo subsídios para o aprimoramento do Celpe-Bras (Mendel, 2019). As análises realizadas em um dos trabalhos também levanta a discussão a respeito do público para o qual o exame é exigido,

em específico aos indivíduos dos PALOP (Tosatti, 2021) que, de acordo com a autora, tem suas variações do português estigmatizadas e não deveriam ter de prestar o exame. As pesquisas qualitativas e quantitativas são complementares umas às outras e contribuem para que se tenha mais informações a respeito do exame (Divino, 2021; Sostruznik, 2023; Schoffen; Divino; 2023).

Dos trabalhos previamente citados, quatro deles analisam subcorpora do Corpus do Celpe-Bras (CorCel), corpus analisado neste trabalho, com uso do *SkE* (Divino, 2021; Hanauer, 2023; Sostruznik, 2023; Schoffen; Divino; 2023). O corpus, ainda em preparação, pretende possibilitar análises que ofereçam maior conhecimento sobre os níveis de proficiência do Celpe-Bras, permitindo que se faça pesquisas que se valham de comparações entre as produções escritas de diferentes notas em uma mesma tarefa, ou até mesmo de uma mesma nota em diferentes tarefas. É importante que sejam fomentadas pesquisas neste sentido, pois têm o poder de contribuir tanto para que professores tenham mais acesso a características definidoras dos gêneros do discurso avaliados pelo exame e possam aprimorar seus materiais didáticos com cada vez mais dados empíricos, quanto para avaliadores, dando a eles mais subsídios para compreender mais profundamente os níveis de proficiência descritos nos parâmetros de avaliação.

2.2 Resumo do capítulo

Neste capítulo, foram apresentados os pressupostos que servem de pilar para a execução desta pesquisa, importantes para situar este trabalho na área interdisciplinar entre Avaliação e Linguística de Corpus. Para tal, foram elucidados fundamentos da Linguística de Corpus, iniciando com algumas maneiras com que se pode utilizar a LC em pesquisas. Nesta mesma linha, foram apresentadas as duas principais abordagens de pesquisa, bem como algumas técnicas de manuseio, com exemplos de busca no *SkE*. Depois, foram apresentados os tipos de corpora, fomentando a discussão acerca do termo “corpora de aprendizes” e argumentando sobre o uso de “corpus de língua adicional” para descrever o corpus compilado para esta pesquisa. A penúltima seção deu conta de estabelecer a relação entre corpora de línguas adicionais e a avaliação de proficiência, seguido de contribuições que estes corpora podem trazer para o ensino e a avaliação, na última seção. Por fim, foram apresentados alguns corpora de PLA, bem como foram trazidos trabalhos, qualitativos e quantitativos, que se

debruçaram sobre os níveis de proficiência do Celpe-Bras, servindo de pano de fundo para este trabalho.

O capítulo 1 e 2 compõem o arcabouço teórico que sustenta esta dissertação. A seguir, serão detalhados os procedimentos metodológicos.

3. MEDIDAS IMEDIATAS: Procedimentos metodológicos

Uma pesquisa se inicia com a escolha de um objeto de pesquisa, um fragmento da realidade que se deseja investigar, e com uma pergunta de pesquisa, algo sobre esse fragmento de realidade sobre o qual se gostaria de saber (Stefanowitsch, 2020). Esta pesquisa visa à caracterização dos níveis de proficiência escrita a partir de dados quantitativos provenientes de um corpus de produções escritas em resposta à Tarefa 4 da edição de 2015/2 do exame Celpe-Bras.

Neste capítulo, serão explicitados, inicialmente, as perguntas e objetivos que nortearam esta pesquisa, seguidos do detalhamento dos estágios de preparação do *corpus* analisado. Por fim, serão descritos os procedimentos para as análises, divididos em diferentes etapas.

3.1 Perguntas de pesquisa

Para fornecer subsídios para o refinamento da descrição dos níveis de proficiência do Celpe-Bras, esta pesquisa de cunho quantitativo e qualitativo tem o intuito de responder à seguinte pergunta norteadora: ***Como as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência em tarefas de leitura e escrita no exame Celpe-Bras?***

Para responder a esta pergunta, foram elaboradas as seguintes questões específicas:

1. O que a extensão dos textos dos subcorpora informa sobre os níveis de proficiência?
 - a. Qual a quantidade de *types* e de *tokens* em cada subcorpus?
 - b. Qual o *Type-Token-Ratio (TTR)* em cada subcorpus?
 - c. Qual a média de *types* e de *tokens* por texto em cada subcorpus?
 - d. Quantos *tokens* tem o menor texto em cada subcorpus?
 - e. Quantos *tokens* tem o maior texto em cada subcorpus?
 - f. Qual a quantidade de sentenças em cada subcorpus?
 - g. Qual a média de sentenças por texto em cada subcorpus?
2. Qual a relação do léxico usado em cada subcorpus com o material de insumo da tarefa?
 - a. Quantos *n-grams* de 6 itens lexicais podem ser encontrados em cada subcorpus e com que *Relative DOCF*?

- b. Qual a relação dos *n-grams* encontrados em cada subcorpus com o material de insumo da tarefa?
 - c. Quantas das primeiras 100 *keywords* de cada subcorpus em comparação com um corpus geral de português não estão contidas no material de insumo da tarefa?
 - d. Quantas palavras das *wordlists* são utilizadas em pelo menos 5% dos textos de cada subcorpus?
 - i. Quantas destas palavras estão no material de insumo?
 - ii. Quantas destas palavras não estão no material de insumo?
3. Quais semelhanças e diferenças na utilização de termos relacionados ao gênero do discurso solicitado na tarefa podem ser encontradas nos subcorpora?
- a. Com que frequência (co-)ocorrem termos de saudação em cada subcorpus?
 - b. Com que frequência (co-)ocorrem termos de endereçamento em cada subcorpus?
 - c. Com que frequência (co-)ocorrem termos de despedida em cada subcorpus?
 - d. Com que frequência (co-)ocorrem termos de assinatura em cada subcorpus?
 - e. Quais termos co-ocorrem com os pronomes possessivos no singular e no plural em cada subcorpus?

3.2 Objetivos do trabalho

O objetivo geral deste trabalho é investigar de que maneira as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência na Tarefa 4, que envolve habilidades de leitura e escrita, da edição 2015/2 do exame Celpe-Bras. Os objetivos específicos se dividem em três grupos. O primeiro deles diz respeito à extensão dos textos, em que se buscou informações sobre: a) a quantidade de *types* e *tokens* em cada subcorpus; b) o *TTR* em cada subcorpus; c) a média de *types* e *tokens* por texto em cada subcorpus; d) a quantidade de *tokens* no menor texto de cada subcorpus; e) a quantidade de *tokens* no maior texto de cada subcorpus; f) a quantidade de sentenças em cada subcorpus; e g) a média de sentenças por texto em cada subcorpus.

O segundo grupo de objetivos é voltado para a integração das habilidades de leitura e escrita, ou seja, para a relação do léxico utilizado em cada subcorpus com o material de insumo da tarefa. Para isto, objetivou-se elencar: a) a quantidade e *Relative DOCF* com que

n-grams de ao menos 6 itens podem ser encontrados em cada subcorpus; b) a relação dos *n-grams* com o material de insumo; c) a quantidade de palavras entre as primeiras 100 *keywords* de cada subcorpus que pode ser encontrada no material de insumo; c) a quantidade de palavras das *wordlists* que estão em pelo menos 5% dos textos de cada subcorpus e, dentre essas, i) quais estão no material de insumo e ii) quais não estão.

O terceiro grupo se refere à incidência de termos relacionados à construção da relação de interlocução dentro do gênero do discurso solicitado na tarefa, buscando quais desses termos são utilizados com maior recorrência em cada subcorpus e, mais especificamente: a) a frequência com que (co-)ocorrem termos de saudação em cada subcorpus; b) a frequência com (co-)ocorrem termos de endereçamento em cada subcorpus; c) a frequência com que (co-)ocorrem termos de despedida em cada subcorpus; d) a frequência com que (co-)ocorrem termos de assinatura em cada subcorpus; e e) quais termos co-ocorrem com os pronomes possessivos no singular e no plural em cada subcorpus.

3.3 CorCel: o corpus da pesquisa

O corpus da pesquisa é composto por textos produzidos por examinandos em resposta à Tarefa 4 da edição de 2015/2 do exame Celpe-Bras, e faz parte do CorCel (Schoffen *et al.*, em preparação). Ao subcorpus utilizado neste estudo, foi atribuído o nome de Corcel20152t4.

O critério de seleção da tarefa para análise neste trabalho foi motivado pela intenção de dar seguimento aos resultados de Divino (2021) que, por meio de uma pesquisa guiada por corpus, chegou a alguns índices de análise que poderiam ser mais explorados posteriormente. Além disso, outros membros do Grupo Avalia já haviam trabalhado previamente com a mesma tarefa (Sirianni, 2016; Mendel, 2019), de forma que esta pesquisa pode corroborar alguns dos resultados dos trabalhos anteriores, baseados em análises qualitativas, com a utilização de métodos estatísticos. Já o critério para a seleção dos textos foi a condição de terem, como nota final, um valor inteiro (5, 4, 3, 2, 1 e 0). Como explicado na seção **1.2.3.1 Processo de avaliação da Parte Escrita**, a avaliação de cada tarefa da PE é composta pela média aritmética da nota atribuída por dois avaliadores distintos. Havendo diferença de mais de um ponto entre as notas, há, ainda, uma terceira avaliação, e a nota final da tarefa é a média aritmética das duas notas mais próximas. Sendo assim, é possível que a nota final da tarefa seja um número decimal (4,5, 3,5, 2,5, 1,5 e 0,5), cujos textos não foram incluídos no recorte deste estudo. No total, o corpus deste trabalho é composto por 2.293 textos. A edição de

2015/2 contou com 4.443 examinandos, isto significa dizer que este trabalho dá conta dos textos de, aproximadamente, 51,6% dos examinandos da edição.

Apesar da anotação do corpus utilizado neste trabalho estar em andamento, a versão aqui analisada não conta com anotações estruturais ou linguísticas, de modo que os textos que o compõem estão “crus”, exatamente como foram escritos originalmente. A única modificação realizada se deu com intuito de garantir a anonimização total dos textos, processo que será detalhado na seção que trata da compilação do corpus.

O subcorpus composto pelo texto de insumo e pelo enunciado da tarefa recebeu o nome de CorCel20152t4mi, em que “mi” corresponde a “material de insumo”. Juntando o texto de insumo com o enunciado, o subcorpus é composto por 472 palavras (*tokens*). Em algumas das etapas de análise, em que o objetivo consistia em averiguar se palavras isoladas encontradas nos corpora dos textos produzidos pelos examinandos constavam ou não no material de insumo, descritas em detalhe na seção **3.6.2 Dados linguísticos**, o CorCel20152t4mi foi formatado de maneira que todas as palavras ficaram em lista, com um termo abaixo do outro. Além disso, foram excluídas da lista todas as palavras que se repetiam no corpus, de modo que restaram 270 palavras diferentes (*types*).

3.4 Apresentação da tarefa

Como foi apresentado no exemplo de tarefa integrada na seção **1.2.2.1.1 Integração de habilidades**, a tarefa analisada neste trabalho, intitulada **Azulejos Valiosos**, propôs que se escrevesse uma **carta aberta à prefeitura municipal**. O autor deveria escrever como um **morador de Belém inconformado com a situação dos casarões históricos da cidade** com o propósito de **explicar a situação dos casarões históricos de Belém e argumentar sobre as necessidades de se tomarem medidas imediatas para solucionar a situação**. O conteúdo informacional a ser recuperado pelo examinando, de acordo com a Resposta Esperada, disponibilizado pelo Inep, era o seguinte:

- 1) Capital paraense já foi uma das cidades brasileiras com maior variedade de azulejos valiosos, boa parte importada da Europa, na virada do século XIX para o XX. Da década de 1970 para cá, mais de 50% dos azulejos se perderam.
- 2) Casarões estão sendo alvos de depredação e de vandalismo, como o caso do Palacete Vitor Maria da Silva.
- 3) O Departamento do Patrimônio Histórico, Artístico e Cultural (Dphac) iniciou o processo de tombamento do casarão.

- 4) Desde fevereiro, quatro casarões tiveram azulejos do século XIX furtados. Há quem suspeite de encomenda de roubos.
- 5) Tudo indica que há um mercado de azulejos na cidade. Os exemplares fora das áreas tombadas não têm qualquer proteção (Inep, 2015).

Para cumprir a tarefa, o examinando deveria mobilizar estas informações específicas do texto de insumo para que pudesse dar conta dos critérios apresentados no enunciado.

3.5 Compilação do corpus

Os textos foram disponibilizados pelo Inep em forma de cópia digitalizada, precisando passar pelos processos de digitação e posterior revisão da digitação para que fossem lidos pelo *software* de análise. Os textos de nota 2 e 5 foram utilizados em Divino (2021), de modo que, para este trabalho, foi realizada a digitação das demais notas e a revisão de todo o corpus, a fim de garantir a fidedignidade ao manuscrito original. Para assegurar a padronização entre os textos, o Grupo Avalia elaborou um tutorial para as digitações e revisões (Anexo C), (Schoffen *et al*, 2024). Uma medida utilizada para aumentar as chances de que os textos estivessem fidedignos aos originais foi delegar as tarefas de digitação e revisão para pessoas diferentes, de modo que cada texto passou por, pelo menos, duas pessoas. Um passo importante do tutorial diz respeito à criação do documento em que será digitado cada texto e à sua nomeação, que deve seguir a ordem ano, edição, tarefa, participante e nota, sem espaço entre as informações, como no exemplo **20152t4p903n3**. Isto indica que o texto em questão foi produzido no ano de 2015, segunda edição, em resposta à Tarefa 4, pelo participante de número 903 e obteve a nota 3. Os textos digitados são indicados com números. Quando o texto digitado passa pelo processo de revisão, que inclui a comparação da digitação com a imagem original, adiciona-se um “r” ao final, ficando **20152t4p903n3r**.

No tutorial, foram incluídos alguns cuidados a serem tomados no momento da digitação e revisão, de modo a tornar os critérios acessíveis a quem fosse realizar esta tarefa. Os critérios foram atualizados à medida que foram surgindo dúvidas relacionadas aos textos. O primeiro critério indica que não devem ser corrigidos erros gramaticais ou ortográficos, mantendo o texto exatamente como foi escrito²⁸. Por se tratar de um manuscrito, a letra do examinando pode ocasionar certa dificuldade na leitura, portanto o segundo critério se refere à

²⁸ Na seção 5. **PROCESSOS DE TOMBAMENTO: Discussão dos resultados** serão discutidas as limitações decorrentes da pesquisa em um corpus não normalizado.

importância de considerar o que faz sentido quando a leitura estivesse difícil, analisando o restante do texto e levando em consideração o contexto, sendo a leitura do enunciado e do texto de insumo da tarefa o primeiro passo antes do início do trabalho com os textos dos examinandos. Embora se trate de um corpus para ser analisado com ferramentas de LC, a formatação dos textos (divisão de parágrafos, alinhamento, quebra de linha, etc) é, também, um elemento relevante para as análises, de modo que um dos critérios do tutorial inclui a indicação de manter a formatação original do texto, exceto quando há informações de ordem diferente na mesma linha, como, por exemplo, local e data alinhados à direita, e remetente e destinatário alinhados à esquerda. Nesse caso, a indicação é manter o alinhamento, deixando, todavia, cada informação em uma linha. O tutorial também conta com instruções a respeito de rasuras, que podem aparecer entre parênteses, riscadas com uma linha na horizontal, ou totalmente rabiscadas, e não devem ser consideradas para a digitação.

Quanto à anonimização, apesar de os textos terem sido cedidos sem qualquer informação sobre o examinando que o escreveu, optou-se por padronizar qualquer nome próprio utilizado ao longo do texto, substituindo-os por “Bibiana Cambará”, personagem da trilogia “O Tempo e o Vento” de Érico Veríssimo²⁹. Este procedimento foi realizado independente de o nome estar no início ou no final do texto, ou do nome usado ser o de alguma pessoa famosa, como, por exemplo, “Luís Inácio Lula da Silva” ou “Ana Maria Braga”, visto que, em virtude de não se ter nenhuma informação sobre a identidade dos examinandos, não é possível saber se o nome utilizado é verídico ou não. Para impedir que essa alteração no nome modificasse o número de palavras no corpus, foi previsto, no tutorial de digitação e revisão, um total de três nomes (Bibiana Arminda Flora) e três sobrenomes (Cambará Terra Amaral), todos oriundos da trilogia “O Tempo e o Vento”. Ou seja, caso houvesse no texto apenas um nome e um sobrenome, a desidentificação seria efetuada com “Bibiana Cambará”, caso houvesse dois nomes e dois sobrenomes, “Bibiana Arminda Cambará Terra”, havendo três nomes e três sobrenomes, “Bibiana Arminda Flora Cambará Terra Amaral”, etc. Embora haja uma variação de nomes, a ordem de preferência é sempre respeitada para facilitar a contagem. Além disso, há também instruções para apelidos e siglas, buscando manter a digitação o mais aproximada possível do texto original. Além de evitar que qualquer nome que o examinando tenha escrito permaneça no documento por questões éticas de respeito ao sigilo da identidade dos autores, a padronização permite que se contabilize

²⁹ A escolha é uma forma de homenagear, na pessoa de Bibiana Cambará, todas as importantes personagens femininas da literatura brasileira.

informações importantes para a análise, ao permitir que seja verificada a quantidade de textos assinados.

Depois de digitados e revisados, as pastas de cada corpus, salvas em Google Docs, foram baixadas em formato zip para o computador da pesquisadora para, dessa forma, poderem ser carregadas no *SkE* (Kilgarriff *et al.*, 2004), *software* utilizado para manuseio e análise inicial dos corpora. A licença paga do *software* foi custeada pelo Programa de Pós-Graduação em Letras (PPGLET) e cedida ao grupo de pesquisa para ser utilizada neste e em outros projetos. A seguir, serão apresentadas as ferramentas do *SkE* utilizadas, bem como o passo-a-passo para as análises desenvolvidas.

3.6 Procedimentos para as análises

As análises foram divididas em duas etapas. A primeira etapa incluiu a pergunta 1, ou seja, deteve-se às análises relacionadas à extensão dos corpora. A segunda etapa incluiu as perguntas do 2 e 3, de modo a averiguar questões relacionadas ao gênero e à recuperação de informações do material de insumo. Para descrever os fenômenos linguísticos de cada subcorpora, foram utilizadas as ferramentas *N-grams*, *Keywords*, *Wordlist* e *Concordance* do *SkE*. A seguir, serão descritas as etapas apresentadas.

3.6.1 Análise da extensão dos corpora

Como mencionado, buscou-se, inicialmente, descrever o tamanho dos corpora. As informações selecionadas para realizar a descrição foram: **número de textos em cada corpus, número total de *types* em cada corpus, número total de *tokens* em cada corpus, texto com o menor número de *tokens* em cada corpus, texto com o maior número de *tokens* em cada corpus, média de *tokens* por texto, *Type-Token-Ratio (TTR)* de cada corpus, número total de sentenças em cada corpus, média de sentenças por texto em cada corpus**. Nesta seção, será explicado como cada uma das informações foi acessada.

Conforme explicado na seção 3.3 **CorCel: o corpus da pesquisa**, o **número de textos de cada corpus** representa a totalidade de textos que receberam cada nota final cheia na tarefa, ou seja, 5, 4, 3, 2, 1 e 0. Para se ter acesso a esta quantidade, foram filtrados, na tabela cedida pelo Inep, os textos que obtiveram estas notas. Esses textos, então, passaram pelo processo descrito na seção 3.5 **Compilação do corpus** para que pudessem ser compilados no

SkE. Quando finalizado o *upload* dos textos para o *software*, no estágio anterior à compilação no programa, os textos aparecem em lista, de modo que se vê, na tela, o nome de cada um dos arquivos, bem como a quantidade de *tokens* em cada um deles. Dessa forma, consegue-se chegar ao **texto com o menor número de *tokens*** e ao **texto com o maior número de *tokens***. *Tokens*³⁰, segundo Biber *et al.* (2002), são entendidos como o número total de palavras, ou seja, tem-se acesso ao texto com a maior e com a menor quantidade de palavras no corpus. Estas informações são relevantes para ilustrar a amplitude no que diz respeito ao tamanho dos textos.

Depois de compilados os corpora, o painel do *SkE* apresenta dados numéricos de cada corpus, a partir dos quais foram coletadas as informações referentes ao **número total de *types***³¹ **em cada corpus**, sendo *types* as diferentes formas de palavras (Biber *et al.*, 2002), ao **número total de *tokens* em cada corpus** e ao **número total de sentenças em cada corpus**. Com estas informações, foram calculadas a **média de *tokens*** e a **média de sentenças por texto**, dividindo, respectivamente, a quantidade total de *tokens* e a quantidade total de sentenças pelo número total de textos em cada um dos corpora. Esses dados também possibilitaram que fosse calculado o *TTR*, uma medida simples usada para comparar a diversidade lexical entre mais de um corpora. Um *TTR* mais alto indica que um corpus tem maior diversidade lexical do que outro. O *TTR* é calculado ao dividir a quantidade de *types* pela quantidade de *tokens* e multiplicando por 100, como pode ser visualizado na Figura 15.

Figura 15 - Fórmula do cálculo *Type-Token-Ratio*

$$\text{word} \div \text{tokens} \times 100 = \text{TTR}$$

Fonte: Sketch Engine.

Dessa forma, tem-se o resultado do *TTR* em porcentagem (Sketch Engine, s.d.). Como exemplo, o *site* do *SkE* apresenta a frase “nos encontramos cara a cara³²”, que contém 4 *types* e 5 *tokens*. Aplicando a fórmula a esta frase, temos, como resultado, um *TTR* de 80%.

³⁰ Informações sobre o total de palavras podem ser acessadas na quantidade de *Words* (com letra maiúscula), informada no painel de informações de cada corpus.

³¹ Informações sobre o número de palavras diferentes podem ser acessadas na quantidade de *words* (com letra minúscula), informada no painel de informações de cada corpus.

³² No original: *We met face to face*.

3.6.2 Análise do léxico utilizado nos corpora

A descrição dos fenômenos linguísticos característicos de cada subcorpus teve como objetivo central indicar como é realizada a recuperação de informações em cada subcorpus, bem como revelar como são articulados elementos constitutivos do gênero carta aberta. A frequência de interesse para as análises não foi sempre a frequência bruta, isto é, o número de vezes que um elemento apareceu no corpus como um todo. O foco foi direcionado à dispersão, que diz respeito à quantidade de textos em que os elementos aparecem em cada corpus. Nas ferramentas em que é possível alterar a ordem em que os elementos aparecem automaticamente, a ordem de relevância foi feita com base na frequência por documento (*DOCF*), que informa em quantos textos do corpus o termo aparece, e na frequência relativa por documento (*Relative DOCF*), que informa o percentual de textos em que o termo aparece. Os resultados obtidos com cada ferramenta para cada subcorpus foram posteriormente comparados, possibilitando que se visualizasse semelhanças e diferenças tanto de escolha lexical, quanto de frequência de uso. A seguir, serão apresentadas as ferramentas *N-grams*, *Keywords*, *Wordlist* e *Concordance*, utilizadas para as análises comparativas entre os corpora:

1. *N-Grams*

Nesta ferramenta, foi selecionada a configuração de buscar por palavra (*word*) como atributo nas configurações, levando em conta a maneira exata como foi escrita, em vez de *lemma*, que inclui as diferentes flexões de gênero e número. Foi escolhida a opção que permite que letras maiúsculas e minúsculas não sejam diferenciadas (A=a). A frequência mínima padrão da ferramenta é de duas ocorrências, que foi mantida. A extensão escolhida para análise foi de 6 elementos lexicais. Para esta busca, foram incluídas não-palavras³³, de modo a permitir que sinais como “%” fossem incluídos. É importante nesta etapa deixar claro que sinais de pontuação, como “.” ou “,” contam como elementos isolados, sendo incluídos nas contagens dos *n-grams*. Os *n-grams* foram, então, comparados com o material de insumo, para averiguar se fragmentos do texto de insumo e do enunciado foram recuperados da exata maneira que estavam. Em cada subcorpus, os *n-grams* foram agrupados de acordo com os trechos do material de insumo ao qual faziam referência. Os resultados de cada corpora foram posteriormente comparados entre si.

2. *Keywords*

³³ Em inglês: “nonwords”.

Como mencionado na seção **2.1.3 Técnicas de manuseio**, as *keywords* não são, necessariamente, as palavras mais frequentes em um corpus quando comparado com outro, mas sim aquelas que são significativamente frequentes ou raras. Neste trabalho, o foco são as palavras significativamente frequentes nos corpora de estudo. Foi utilizada esta ferramenta para que se conseguisse contrastar dois corpora, um corpus de estudo e um como corpus de referência. Os corpora de estudo foram os corpora n5, n4, n3, n2 e n1. Como corpus de referência, foi utilizado o Portuguese Trends³⁴, um corpus de português disponível para uso no *SkE*. O corpus Portuguese Trends apresenta 374.409.743 *tokens* e 1.461.503 *types* e é composto, majoritariamente, por textos jornalísticos de sites da internet sobre temáticas diversas, como esportes, educação, cultura etc. Os textos são, em sua maioria, coletados de sites brasileiros e o corpus é atualizado diariamente com novos textos. Além disso, o Portuguese Trends é disponibilizado com fácil acesso no *SkE*. Esta comparação é interessante, pois os termos relacionados ao contexto de produção proposto pela tarefa ficam em evidência, tornando possível visualizar o que é considerado mais relevante em cada subcorpus.

Os comandos utilizados para a extração das *keywords* incluíram: não diferenciar entre letra maiúscula e minúscula ($A=a$); ter pelo menos um elemento alfanumérico; ter frequência mínima de uma ocorrência, padrão nesta ferramenta; buscar por *word* como atributo nas configurações. Para facilitar a comparação, foram elaborados quadros com as informações encontradas nas buscas. As primeiras 100 *keywords* foram analisadas na ordem em que apareceram no *SkE*. Entre as palavras encontradas, foram feitas análises separadas das palavras que estão no material de insumo e das que não estão. Para isto, foi feita a comparação das listas com o material de insumo. Esta comparação foi feita automaticamente no Google Sheets, possibilitando que se pudesse encontrar com facilidade os termos em comum com o material de insumo. Os resultados de todos os corpora foram comparados entre si.

3. *Wordlist*

Outro recurso utilizado foi a ferramenta *Wordlist*, que elenca as palavras mais frequentes em um corpus. Neste caso, o comando também incluiu a não diferenciação entre letra maiúscula e minúscula e a busca por *word* em vez de *lemma* ou de alguma classe gramatical específica (substantivos, adjetivos, verbos etc). Além disso, não foram estabelecidas frequências mínimas e/ou máximas de ocorrência. Depois de geradas as listas de

³⁴ Esta é a única etapa em que o corpus Portuguese Trends é utilizado.

palavras, foram contabilizadas quantas delas estavam no material de insumo. Para esta comparação, foram incluídas as palavras que aparecem em, pelo menos, 5% dos textos em cada subcorpus. O mesmo foi realizado com as palavras que aparecem em, pelo menos, 10%, 20% e 50% dos textos.

4. *Concordance*

A última ferramenta utilizada teve como objetivo focalizar em elementos constitutivos do gênero carta aberta, como saudações, endereçamento, assinatura, despedida e pronomes possessivos. Todos esses elementos surgiram a partir das análises feitas anteriormente, sobretudo com as ferramentas *Keywords* e *Wordlist*, só então se tornaram interessantes para serem analisados em contexto. De modo a encontrar palavras que co-ocorrem com os termos buscados, foi utilizada a ferramenta *collocations*, buscando termos classificados como primeiros e segundos colocados à direita ou à esquerda do termo pesquisado. As posições dos colocados foram pesquisadas individualmente, por exemplo: primeiro somente os termos que apareceram como primeiro colocado à direita, depois os que aparecem como segundo colocado à direita etc. A frequência padrão das *collocations* é de 5 ocorrências do colocado no corpus e 3 ocorrências da colocação (co-ocorrência do colocado com o termo pesquisado, denominado “núcleo”), que foi mantida.

Como atributo de busca, foi utilizado majoritariamente *word*, mas em alguns casos foi utilizado *gender_lemma*, pois em *word* não há como evitar que se faça a diferenciação entre maiúsculas e minúsculas (A=a). Além da possibilidade de não diferenciar maiúsculas e minúsculas, a opção pelo atributo *gender_lemma* se dá pelo fato de que apenas este atributo tem a capacidade exibir a terminologia na forma correta em idiomas que observam a concordância de gênero entre adjetivos e substantivos, como é o português, de modo que o lema padrão (*lemma*), muito utilizado em análises de línguas que não observam essa concordância, produziria uma combinação gramaticalmente inaceitável de formas de palavras em português (Sketch Engine, s.d.). A forma lematizada não foi buscada em todos os casos, pois para algumas palavras ela apresenta um comportamento atípico. Ao mesmo tempo que escolher a opção *gender_lemma* faz com que as palavras “prefeitura” e “Prefeitura” sejam contabilizadas como uma só, ela transforma o adjetivo “prezado” no lema “prezar”, como se “prezado” estivesse no particípio. Isto tem relação com o lematizador de língua portuguesa utilizado no Sketch Engine. Com outro lematizador, é possível que os resultados fossem

diferentes. A pesquisa padrão foi feita com *word* e, em apresentando um comportamento atípico, foi feita com *gender_lemma*.

A medida de associação escolhida foi *logDice* que, como apresentado na seção **3.1.3 Técnicas de manuseio**, permite que se faça a comparação entre corpora de tamanhos diferentes. Um valor próximo a 14 significa que os termos co-ocorrem mutuamente quase com exclusividade, de modo que valores que se aproximam do 14 (valor máximo) indicam que há uma associação forte entre os termos.

3.7 Resumo do capítulo

Este capítulo deu conta da descrição dos procedimentos metodológicos aplicados às análises. Para iniciar o capítulo, foram apresentadas as perguntas e os objetivos da pesquisa. Depois, foi descrita a composição do corpus da pesquisa, com a apresentação da tarefa que gerou os textos que o compõem e o detalhamento da sua compilação. Por fim, foram elucidadas as etapas das análises.

No próximo capítulo, serão apresentados os resultados do trabalho.

6. MERCADO DE AZULEJOS: Considerações finais e limitações do trabalho

A partir do meu contato com o exame Celpe-Bras, me constituí pesquisadora, professora e avaliadora, e é da união entre estes três olhares que partiram as análises aqui apresentadas. Antes de concluir, gostaria de juntar os fragmentos que compuseram este trabalho para finalizar a montagem do quebra-cabeça. Além disso, aponto algumas limitações e sugiro algumas medidas imediatas para dar seguimento às pesquisas.

Com o intuito de responder à pergunta “**Como as análises lexicais realizadas com ferramentas de Linguística de Corpus podem contribuir para a descrição dos níveis de proficiência em tarefas de leitura e escrita no exame Celpe-Bras?**”, iniciei o segundo capítulo apresentando o exame Celpe-Bras, contando sobre a história de sua criação e trazendo suas especificações, pautadas nos oito componentes descritos por Douglas (2000) a respeito das principais características que deve ter um exame de língua. Foram apresentadas, também, as concepções teóricas do exame, aprofundando a noção de proficiência no Celpe-Bras, compreendida a partir de uma escala composta por níveis de proficiência (Scaramucci, 2000) e pautada na capacidade de um indivíduo de participar em diferentes práticas em que a língua é utilizada (Brasil, 2020), materializadas em gêneros do discurso (Bakhtin, 2003). Depois, no mesmo capítulo, foi descrita a estrutura do exame, dando conta da Parte Escrita e da Parte Oral, incluindo seus processos de avaliação e resultado final. Especificamente sobre a Parte Escrita, foi abordada a integração de habilidades, que se dá, no exame, pela recontextualização de informações do material de insumo (Mendel, 2019) a partir de especificações particulares dos enunciados das tarefas (Mendel; Schoffen, 2017), de modo que os propósitos de leitura são diretamente pautados pelos propósitos de escrita (Scaramucci, 2016).

Para encerrar o segundo capítulo, foram discutidos os conceitos de confiabilidade, validade, autenticidade, impacto e praticidade (Douglas, 2000), como parte do plano para avaliar a qualidade de uma boa avaliação. Especialmente relacionadas à confiabilidade e à validade, estão duas contribuições que a LC pode trazer a estágios do ciclo de vida de um exame, apresentadas no terceiro capítulo: a validação do instrumento avaliativo por meio de uma abordagem baseada em argumentos e a descrição dos níveis de proficiência (Cushing, 2022). Neste sentido, este trabalho se propõe a contribuir para a validade e a confiabilidade do exame ao buscar um refinamento na descrição de seus níveis de proficiência, tendo o intuito de elucidar como os critérios de avaliação foram interpretados pelos avaliadores daquela

edição. A divulgação destas informações também contribui para a validade do exame, visto que a confiabilidade pode ser considerada, de acordo com Yan e Fan (2022), uma condição de validade, embora Scaramucci afirme que pode haver certa tensão entre ambos os termos. Na divulgação destas informações há uma tentativa de diminuir esta tensão, pois torna públicos resultados que podem auxiliar no refinamento dos critérios de avaliação, possibilitando que avaliadores tenham mais conhecimento sobre o que é avaliado. Dado o alto impacto de um exame como o Celpe-Bras na vida de milhares de pessoas, o acesso à informação é muito importante, de modo que este trabalho se propõe a contribuir, também, com o resgate da história do exame iniciada em 2014 com a publicização do Acervo Celpe-Bras (Schoffen *et al.*, 2017).

No terceiro capítulo, foram introduzidos alguns conceitos relacionados à Linguística de Corpus. Neste capítulo, foram apresentadas algumas possibilidades de uso de um corpus, técnicas de manuseio e abordagens de pesquisa. Motivadas por discussões realizadas com membros do Grupo Avalia, trouxemos algumas questões relacionadas ao termo *corpora de aprendizes*, por acreditarmos haver a necessidade de marcar a diferenciação do corpus utilizado neste trabalho daqueles nomeados como corpora de aprendizes. Provavelmente há muitos estudantes de PLA que prestam o Celpe-Bras, mas esta não é uma afirmação que pode ser feita sobre todos os examinandos. Atribuir o rótulo de *aprendiz* a qualquer estrangeiro é, ao nosso ver, concordar com a ideia de pureza linguística, que está atrelada à imagem de um falante nativo idealizado. Para nos distanciarmos de um viés que pode, em certa medida, deslegitimar variações não-nativas do português, como aponta Jenkins (2005), aderimos ao termo *corpus de língua adicional*, também com intenção de dar seguimento a este debate área da pesquisa informada por corpus.

Relacionado ao terceiro capítulo, é apresentada a primeira limitação deste trabalho. Como mencionado na seção **2.1.5 Corpora de línguas adicionais e avaliação de proficiência**, este trabalho foi realizado com um corpus não anotado, o que restringe a pesquisa em certa medida. Os subcorpora do CorCel compostos de produções escritas em resposta às tarefas do Celpe-Bras estão, atualmente, passando pelo processo de anotação, procedimento que, além de incluir anotações para o acesso automático a elementos típicos do gênero discursivo, como saudações, endereçamentos e despedidas, tem o objetivo de regularizar as inadequações ortográficas para que uma mesma palavra grafada de diferentes maneiras seja automaticamente reconhecida como a mesma palavra, sem que as versões originais sejam apagadas. Fazer a normalização das inadequações ortográficas é um processo

comum em textos de corpora que apresentam inadequações (del Rio *et al.*, 2016) e permite que os dados sejam analisados de forma mais precisa (Aluísio; Almeida, 2006). Com os corpora devidamente marcados, espera-se acessar: a quantidade de *types* considerando a mesma palavra escrita de diferentes maneiras como uma só, possibilitando que o *TTR* seja recalculado para ver se há uma diferença muito grande em relação aos resultados aqui apresentados; mais *n-grams* que possam evidenciar a recuperação de trechos do material de insumo; outras *keywords* que possam dar indícios de especificidades de cada corpus; quantidades diferentes de palavras da *wordlist* que tenham relação com o material de insumo (ou não); e mais elementos característicos do gênero carta aberta, pela possibilidade de detecção automática.

A segunda limitação do trabalho diz respeito às ferramentas do *SkE* escolhidas para análise. Como apresentado na seção **2.1.3 Técnicas de manuseio**, existem outras ferramentas que não foram exploradas nesta pesquisa, como *Word Sketch*, que apontaria semelhanças e diferenças entre o comportamento gramatical de termos específicos, e *Thesaurus*, que indicaria quais termos estão no mesmo campo semântico de termos específicos em cada corpus. Além de trabalhar com outras ferramentas, as próprias ferramentas utilizadas neste trabalho (*N-grams*, *Keywords*, *Wordlist* e *Concordance*) sugerem outras possibilidades de análise que não foram abordadas. Com *n-grams*, pode-se olhar para sequências de extensão menores a partir das quais se pode encontrar resultados diferentes, como cópia de trechos mais curtos, por exemplo. Entre as *keywords* e as palavras nas *wordlists*, pode-se aprofundar mais em itens que estão no texto de insumo e itens que não estão, ou dar enfoque a classes gramaticais específicas, para, depois, levar outros termos para a pesquisa em *Concordance*. Outras análises podem levar a resultados que, até então, não foram previstos.

A terceira limitação do trabalho se relaciona com a segunda, porém, diz respeito ao *software* utilizado. Para além do *SkE*, há uma infinidade de programas de análise de corpus que podem ser utilizados para analisar corpora, como o AntConc, LancsBox e Coh-Metrix, além de pacotes de análise de dados, como RStudio, SAS etc. Alguns *softwares* apresentam a possibilidade de realizar pesquisas semelhantes, muito embora possuam especificidades que os diferem dos demais, o que significa que podem indicar caminhos muito diferentes dos aqui apresentados. É possível utilizar apenas um *software*, como realizado neste trabalho, ou mais de um, tendo um leque maior de possibilidades de análise, indo além das limitações apresentadas por cada *software*.

Analisar somente as diferenças de frequência pode não ser suficiente para elucidar o que diferencia uma nota da outra, e com isto, apresentamos a quarta limitação do trabalho. Muitas vezes, resultados alcançados sem análise estatística podem parecer significativos, porém, matematicamente, não são tão representativos das semelhanças e diferenças que se deseja analisar. De modo a averiguar se as diferenças apresentadas são estatisticamente relevantes, seria necessário que todos os dados provenientes deste estudo passassem por testes estatísticos, não somente aqueles analisados com a ferramenta *Concordance*. Assim, poderíamos ter a certeza de que os resultados são realmente significativos.

A quinta limitação do trabalho diz respeito aos resultados obtidos considerarem apenas fragmentos isolados dos textos e não cada texto em sua integralidade. Embora tenham sido incorporadas a este trabalho conclusões que são fruto dos trabalhos de Sirianni (2016) e Mendel (2019), que analisaram as produções completas em resposta a esta mesma tarefa, seria importante que os resultados apresentados neste trabalho fossem observados também levando em conta o texto como um todo e não apenas fragmentos isolados. Isto faria com que se pudesse tirar conclusões sobre como os fenômenos linguísticos aqui descritos estão interligados em um contexto maior, operando conjuntamente para a construção do texto. Além disso, seria interessante que isto fosse feito em uma grande quantidade de textos. Como dito anteriormente, os textos de uma mesma nota não são todos iguais, mas apresentam características em comum, de modo que os achados deste trabalho podem indicar diferentes maneiras com que alguns itens padronizados foram articulados nos textos para compor estas notas.

Gostaria, por fim, de dar algumas outras sugestões possíveis de estudos que podem ser realizados a partir dos resultados deste trabalho. Os números indicaram que há textos avaliados em todas as notas, com atenção especial à nota 5, que não utilizaram termos que podem ser considerados essenciais para explicar o problema dado no enunciado, como é o caso de “azulejos” e “casarões”, presentes no material de insumo. Analisar estes textos de maneira qualitativa pode permitir que se tenha uma compreensão mais profunda sobre as possibilidades de recuperação do conteúdo informacional do texto de insumo utilizadas pelos examinandos, auxiliando professores de cursos preparatórios a terem mais conhecimento sobre as possibilidades de reestruturação do conteúdo lido nas produções escritas. Também se pode ir para o sentido contrário, analisando qualitativamente aqueles textos que apresentam cópia de trechos longos do material de insumo, com o objetivo de descrever como a cópia se integra ao restante do texto. Um estudo neste sentido pode ter o potencial de explicar

empiricamente porque a sensação dos avaliadores, durante o processo de correção, de estar lendo trechos copiados, é maior quando se lê textos avaliados com notas mais baixas do que com notas mais altas.

Ainda relacionado às cópias, seria interessante fazer um trabalho que extraísse *n-grams* longos de produções escritas em resposta às Tarefas 1 e 2 do Celpe-Bras, que têm, respectivamente, um vídeo e um áudio como insumos. Posteriormente, poderia-se fazer a comparação dos resultados com as conclusões apresentadas neste trabalho. Fundamentalmente, as Tarefas 1 e 2 se diferenciam da tarefa aqui analisada pela modalidade do insumo, que implica um tipo diferente de compreensão (oral). Quanto aos propósitos, enquanto a Tarefa 4 solicita, majoritariamente, que o examinando se posicione perante uma situação, as Tarefas 1 e 2 têm como propósito mais recorrente o de divulgar (Schoffen *et al.*, 2018). Estas duas características contribuem para que a maneira com que o conteúdo informacional a ser recuperado seja distinta. O palpite é que, por estes motivos, haja proporcionalmente mais cópia em textos de notas 5 e 4 em resposta às Tarefas 1 e 2 do que há em textos que têm como material de insumo um texto escrito.

Reconhecendo as limitações deste estudo e deixando algumas sugestões para futuras pesquisas, percebe-se que a LC se mostra valiosa para indicar caminhos que podem levar a uma compreensão cada vez mais aprofundada a respeito dos níveis de proficiência. Como indicado acima, a LC não é um fim, mas um importante meio para que se reconheça mais possibilidades de pesquisa. Deixo minha contribuição para que a história do Celpe-Bras continue sendo contada, com intenção de que esta área de pesquisa siga crescendo e aumentando cada vez mais de valor.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALDERSON, J. C. Do corpora have a role in language assessment? *In*: THOMAS, J.; SHORT, M. (Eds.). **Using corpora for language research**: Studies in the honour of Geoffrey Leech. London: Longman, 1996, p. 248-259.
- ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Calidoscópico**, [S. l.], v. 4, n. 3, p. 156–178, 2021. Disponível em: <https://revistas.unisinos.br/index.php/calidoscopio/article/view/6002>. Acesso em: 5 mar. 2024.
- BACHMAN, L. F.; PALMER, A. S. **Language Testing in Practice**. Oxford: Oxford University Press, 1996.
- BAKHTIN, M. M. **Estética da Criação Verbal**. São Paulo: Martins Fontes, 2003.
- BARKER, F. How can corpora be used in language testing? *In*: O'KEEFFE, A.; MCCARTHY, M. (org). **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, p. 633-645.
- BARKER, F.; SALAMOURA, A.; SAVILLE, N. Learner corpora and language testing. *In*: GRANGER, S.; GILQUIN, G.; MEUNIER, F. (Orgs.). **The Cambridge handbook of learner corpus research**. Cambridge: Cambridge University, 2015, p. 511–534.
- BIBER, D. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. *In*: HEINE, B.; NARROG, H. (eds). **The Oxford Handbook of Linguistic Analysis**. Oxford: Oxford Academic, 2012, p. 159-192.
- BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **Guia do participante**: tarefas comentadas que compõem a edição de abril de 2013 do exame. Brasília, 2013a. Disponível em: <https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Guia-do-Participante-2013.pdf>. Acesso em: 3 de agosto de 2023.
- BRASIL. Decreto n. 7.948, de 12 de março de 2013. Dispõe sobre o programa de estudantes-convênio de graduação - PEC-G. 2013b. Disponível em: <https://legislacao.presidencia.gov.br/atos/?tipo=DEC&numero=7948&ano=2013&ato=b9aITRE5ENVpWTb33>. Acesso em: 4 de julho de 2024.
- BRASIL. **Documento-base do exame Celpe-Bras**. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020.
- BRASIL. Edital nº 8 de 1º de fevereiro de 2023. Exame Para Obtenção Do Certificado De Proficiência Em Língua Portuguesa Para Estrangeiros (Celpe-Bras) 2023. Disponível em: <https://www.in.gov.br/web/dou/-/edital-n-8-de-1-de-fevereiro-de-2023-461812319>. Acesso em: 4 de agosto de 2023.
- BREZINA, V. **Statistics in corpus linguistics**: A practical guide. Cambridge University Press, 2018.
- BULLA, G. da S.; KUHN, T. Z.. ReVEL na Escola: Português como Língua Adicional no Brasil - perfis e contextos implicados. **ReVEL**, [S.l.], vol. 18, n. 35, 2020. Disponível em:

<https://www.revel.inf.br/files/d0e7e22f8c5191e1009a4ca3dc37df2e.pdf>. Acesso em: 08 jul. 2024.

CALLIES, M.; GÖTZ, S. Learner corpora in language testing and assessment: Prospects and challenges. *In*: CALLIES, Marcus; GÖTZ, Sandra (orgs.). **Learner Corpora in Language Testing and Assessment**. Amsterdam: Benjamins, 2015.

CALLIES, M.; DÍEZ-BEDMAR, B.; ZAYTSEVA, E. Using learner corpora for testing and assessing L2 proficiency. *In*: LECLERCQ, P.; HILTON, H.; EDMONDS, A. (eds.). **Measuring L2 proficiency: Perspectives from SLA**. Multilingual Matters, 2014, p. 71-90.

CARLSEN, C. (ed.). Norsk Profil. Det europeiske rammeverket spesifisert for norsk. **Et første steg**. Oslo, Norway: Novus forlag, 2013.

CARVALHO, S. C.; SCHLATTER, M. Ações de difusão internacional da língua portuguesa. **Cadernos do IL.**, Porto Alegre, n. 42, jun. 2011. p. 260-284. Disponível em: <https://seer.ufrgs.br/index.php/cadernosdoil/article/view/26027/15240>. Acesso em: 08 jul. 2024.

CHAPELLE, C. A.; LEE, H. Conceptions of validity. *In*: FULCHER, Glenn; HARDING, Luke (orgs.). **The Routledge Handbook of Language Testing**. Londres, New York: Routledge, 2022, p. 17-31.

CHEN, Y.-h.; BAKER, P. Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. **Applied Linguistics**, Oxford, v. 37, n. 6, p. 849–880, 2016. Disponível em: https://www.researchgate.net/publication/285275355_Investigating_Criterial_Discourse_Features_across_Second_Language_Development_Lexical_Bundles_in_Rated_Learner_Essays_CEFR_B1_B2_and_C1. Acesso em: 08 jul. 2024.

CELPE-BRAS. **Parte Escrita: Caderno de Questões 2015/2**. Brasília: CEBRASPE, 2015. Disponível em: https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/PARTE-ESCRITA_CELPEBRAS_2015_II_Caderno-de-questoes.pdf. Acesso em: 11 jul. 2024.

CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DE LISBOA [Site institucional]. **Recolha de Dados de PLE**. Disponível em: <https://www.clul.ulisboa.pt/recurso/recolha-de-dados-de-ple>. Acesso em: 7 mar 2024.

CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA. **Corpus de Aquisição de L2 (CAL2)**, 2012. Disponível em: <https://cal2.clunl.fcsh.unl.pt/index.html>. Acesso em: 7 mar 2024.

CUSHING, S. T. Corpus linguistics in language testing research. **Language Testing**, [S.l.], v. 34, n. 4, p. 441-449, set. 2017. Disponível em: <https://doi.org/10.1177/0265532217713044>. Acesso em: 20 dez 2023.

CUSHING, S. T. Corpus Linguistics and Language Testing. *In*: FULCHER, Glenn; HARDING, Luke (Orgs.), **The Routledge Handbook of Language Testing**. Londres, New York: Routledge, 2022, p. 545-560.

DEL RÍO, I. *et al.* Towards error annotation in a learner corpus of Portuguese. Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016. **Linköping Electronic Conference Proceedings 130**, [S.l.], p. 8–17, 2016. Disponível em: <https://aclanthology.org/W16-6502.pdf>. Acesso em: 08 jul. 2024.

DELL'ISOLA, R. L. P. *et al.* A avaliação de proficiência em português língua estrangeira: o Exame CELPE-Bras. **Revista Brasileira de Linguística Aplicada**, v.3, n .1, p. 153-184, 2003.

DELL'ISOLA, R. L. P.; TOSATTI, N. M. Movimentos Retóricos Em Produções Escritas No Exame Celpe-Bras. **Gragoatá**, v. 20, n. 38, jun. 2015.

DELL'ISOLA, R. L. P., PORDEUS, I. R. Os Enunciados de Tarefas Integradas de Leitura e Escrita do Exame Celpe-Bras. **Inventário**, Niterói, p. 65–78, 2021. Recuperado de: <https://periodicos.ufba.br/index.php/inventario/article/view/27218>. Acesso em: 08 jul. 2024.

DÍEZ-BEDMAR, M. B. Article use and criterial features in Spanish EFL writing; A pilot study from CEFR A2 to B2 levels. *In*: CALLIES, M.; GÖTZ, S. (eds.), **Learner corpora in language testing and assessment**. Amsterdam: Benjamins, 2015, p. 163–190.

DIVINO, L. S. **Índices lexicais de análise para a caracterização dos níveis intermediário e avançado superior no Exame Celpe-Bras**. Orientadora: Juliana Roquele Schoffen. 2021. 68f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021. Disponível em: <https://lume.ufrgs.br/handle/10183/235361>. Acesso em: 09 jul. 2024.

DOUGLAS, D. **Assessing Languages for Specific Purposes**. Cambridge: Cambridge University Press, 2000.

ELLISON, F. P. **Modern Portuguese: A project of the Modern Language Association**. New York: Knopf, 1971.

EDUCATIONAL TESTING SERVICE (ETS). TOEFL iBT Practice Tests. Disponível em: <https://www.ets.org/toefl/test-takers/ibt/prepare/practice-tests.html#accordion-31f3d4f46f-item-cd4759b094>. Acesso em: 8 jan. 2024.

EVERS, A. **Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras**. Orientadora: Maria José Bocorny Finatto. 2013. 174f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: <https://lume.ufrgs.br/handle/10183/79447>. Acesso em: 09 jul. 2024.

EVINSON, J. What are the basics of analysing a corpus? *In*: O'KEEFFE, A; MCCARTHY, M. (org). **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, p. 122-135.

FERNANDES, E. **A língua portuguesa em contexto internacional: um guia para professores**. New Jersey: Boavista Press, 2020.

GOMES, M. S. **A complexidade de tarefas de leitura e produção escrita no exame Celpe-Bras**. Orientadora: Margarete Schlatter. 2009. 109f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009. Disponível em: <https://lume.ufrgs.br/handle/10183/16228>. Acesso em: 09 jul. 2024.

- MATOS, F. G. de. Ensino de português para estrangeiros: algumas contribuições brasileiras. *In: TROUCHE, L.; JÚDICE, N. (orgs.). O ensino de português para estrangeiros: Ciclo de palestras/Universidade Federal Fluminense. Niterói, RJ: EDUFF, p. 9-13, 1995.*
- GABLASOVA, D. Corpora for second language assessments. *In: WINKE, P.; BRUNFAUT, T. The Routledge handbook of second language acquisition and language testing.* Londres: Routledge, 2021, p. 45-53.
- GABLASOVA, D. *et al.* Epistemic stance in spoken L2 English: The effect of task and speaker style. **Applied Linguistics**, Oxford, n. 38, v. , p. 613-637, 2017. Disponível em: <https://doi.org/10.1093/applin/amv055>. Acesso em: 09 jul. 2024.
- GABLASOVA, D., BREZINA, V., McENERY, T. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. **Language Learning**, Michigan, v. 67, n. S1: p. 130-154, jun. 2017. Disponível em: <https://doi.org/10.1111/lang.12226>. Acesso em: 09 jul. 2023.
- GÖTZ, S. Learner Corpora to Inform Testing and Assessment. *In: WINKE, P.; BRUNFAUT, T. The Routledge handbook of corpora and English language teaching and learning.* Londres: Routledge, 2022, p. 311-326.
- GRANGER, S. Learner corpora. *In: LÜDELING, A.; KYTÖ, M. (orgs.). Corpus Linguistics: An International Handbook. Volume 1.* Berlin & New York: Walter de Gruyter, 2008, p. 259-275.
- GRANGER, S. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. *In: AIJMER, K. (org.). Corpora and Language Teaching.* Amsterdam e Philadelphia: Benjamins, 2009, p. 13-32.
- GRANGER, S. Learner corpora in foreign language education. *In: THORNE, S.; MAY, S. (orgs.) Language, Education and Technology: Encyclopedia of Language and Education.* [S.l]: Springer , 2017, p. 427–440.
- GRIES, S. T. Useful statistics for corpus linguistics. *In: SÁNCHEZ, A.; ALMELA, M. (eds.), A mosaic of corpus linguistics: selected approaches.* Frankfurt am Main: Peter Lang, 2010, p. 269-291.
- GYLLSTAD, H. *et al.* Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. **EUROSLA Yearbook**, [S.l], v. 14, n. 1, p. 1–30, ago. 2014. Disponível em: <https://doi.org/10.1075/eurosla.14.01gyl>. Acesso em: 09 jul. 2024.
- HANAUER, I. D. **Caracterização dos níveis Intermediário e Avançado Superior do exame Celpe-Bras em produções escritas de examinandos no gênero carta/e-mail: contribuições de uma análise guiada por corpus.** 2023. 84 f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.
- HARRISON, J.; BARKER, F. **English profile in practice.** Cambridge: Cambridge University Press, 2015.

HAWKINS, J. A.; FILIPOVIC, L. **Criteria Features in L2 English**: Specifying the Reference Levels of the Common European Framework. Cambridge: Cambridge University Press, 2012.

HUNSTON, S. Using a Corpus to explore patterns. *In*: O'KEEFFE, A.; MCCARTHY, M. (org), **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, p. 152-166.

JENKINS, J. ELF at the gate: the position of English as a Lingua Franca. **Humanising Language Teaching**, Oxford, v. 7, n. 2, mar. 2005. Disponível em: <http://old.hltmag.co.uk/mar05/idea.htm#C2>. Acesso em: 16 de outubro de 2023.

KANE, M. T. Articulating a validity argument. *In*: FULCHER, Glenn; HARDING, Luke (orgs.), **The Routledge Handbook of Language Testing**. Londres, New York, Routledge, 2022, p. 32-47.

KILGARRIFF, A. *et al.* The Sketch Engine. *In*: WILLIAMS, G.; VESSIER, S. (eds). **Proceedings of the XI EURALEX International Congress**. França: Université de Bretagne-Sud, 2004, p. 105–116. Disponível em: https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf. Acesso em: 30 de novembro de 2023.

KUHN, T. Z.; FERREIRA, J. P. O Corpus de Português Escrito em Periódicos - CoPEP. **DELTA**, [S. l.], v. 36, n. 2, 2020. Disponível em: <https://revistas.pucsp.br/index.php/delta/article/view/52552>. Acesso em: 29 fev. 2024.

KUNILOVSKAYA, M.; KOVIAZINA, M. Sketch engine: A toolbox for linguistic discovery. **Journal of Linguistics/Jazykovedný časopis**, Alemanha, v. 68, n. 3, p. 503-507, dez. 2017. Disponível em: https://www.researchgate.net/publication/324760982_Sketch_Engine_A_Toolbox_for_Linguistic_Discovery. Acesso em: 09 jul. 2024.

KUNRATH, S. P. **Os descritores gerais e a progressão dos níveis de proficiência do Exame Celpe-Bras**. Orientadora: Margarete Schlatter. 198f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019. Disponível em: <https://lume.ufrgs.br/handle/10183/239041>. Acesso em: 09 jul. 2024.

MARTINS, C. *et al.* **Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2)**: Subcorpus Português Língua Estrangeira. Coimbra: CELGA-ILTEC, 2019.

MAURANEN, A. Learners and users - Who do we want corpus data from? *In*: MEUNIER, F. *et al.* **A Taste for Corpora**: In Honour of Sylviane Granger. Amsterdam: John Benjamins, 2011, p. 155-171.

MCNAMARA, T. F. **Language Testing**. Oxford: Oxford University Press, 2000.

MENDES, A. *et al.* The COPLE2 Corpus: a Learner Corpus for Portuguese. **Proceedings of LREC 2016**, Portorož, Slovenia, p. 3207–3214, mai. 2016. Disponível em: <https://aclanthology.org/L16-1511/>. Acesso em: 09 jul. 2024.

MENDEL, K. **Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras**. Orientadora: Juliana Roquele Schoffen. 2019. 177f. Dissertação (Mestrado em

Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019. Disponível em: <https://lume.ufrgs.br/handle/10183/217789>. Acesso em: 09 jul. 2024.

MEUNIER, F. Corpus linguistics and second/foreign language learning: exploring multiple paths. **Revista brasileira de linguística aplicada**, Belo Horizonte, v. 11, n. 2, p. 459-477, 2011. Disponível em: <https://www.scielo.br/j/rbla/a/BLJ6xy89SLRH7KNYSzJTPjS/?lang=en>. Acesso em: 20 de outubro de 2023.

NAGASAWA, E. Y. **Postos aplicadores credenciados do exame Celpe-Bras**. Dados Celpe-Bras. Acervo Celpe-Bras: UFRGS, 2019. Disponível em: <https://www.ufrgs.br/acervocelpebras/dados-celpe-bras/>. Acesso em 13 mar. 2024.

NAGASAWA, E. Y.; HANAUER, I. D. **Número de examinandos do exame Celpe-Bras**. Acervo Celpe-Bras, 2023. Disponível em: <https://www.ufrgs.br/acervocelpebras/dados-celpe-bras/>. Acesso em: 13 mar. 2024.

NAGASAWA, E. Y. **O Conteúdo de Insumo em Tarefas que Integram Leitura e Escrita no Celpe-Bras: Uma Abordagem Informada Por Corpus**. Orientadora: Margarete Schlatter. 2023. 302f. Tese (Doutorado em Linguística Aplicada) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023. Disponível em: <https://lume.ufrgs.br/handle/10183/271615>. Acesso em: 10 jul. 2024.

O'KEEFFE, A.; MCCARTHY, M. Historical perspective: What are corpora and how have they evolved? *In*: O'KEEFFE, A.; MCCARTHY, M. (org). **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, p 3-13.

PAQUOT, M. The phraseological dimension in interlanguage complexity research. **Second language research**, Califórnia, v. 35, n. 1, p. 121-145, jan. 2019. Disponível em: <https://doi.org/10.1177/0267658317694221>. Acesso em: 10 jul. 2024.

PRODEAU, M.; LOPEZ, S.; VÉRONIQUE, D. Acquisition of French as a second language: Do developmental stages correlate with CEFR levels? **Apples**, Finlândia, v. 6, n. 1, p. 47-68, out. 2012. Disponível em: <https://apples.journal.fi/article/view/97832>. Acesso em: 10 jul. 2024.

RAMEH, C.; ABREU, M. I. **Português Contemporâneo**. Washington: Georgetown University Press, 1967.

REPPEN, R. Building a corpus: What are the key considerations? *In*: O'KEEFFE, A.; MCCARTHY, M. (org). **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, p 31-37.

ROSS, S. J. Inference and prediction in language testing. *In*: FULCHER, Glenn; HARDING, Luke (orgs.). **The Routledge Handbook of Language Testing**. Londres, New York, Routledge, 2022, p. 33-48.

RYCHLÝ, P. A Lexicographer-Friendly Association Score. *In*: SOJKA, P.; HORÁK, A. (eds.). **Proceedings of Recent Advances in Slavonic Natural Language Processing**. República Tcheca: RASLAN, 2008, p. 6-9.

SARDINHA, T. B. Linguística de Corpus: Histórico e Problemática. **DELTA.**, São Paulo, v.16, n. 2, p. 323-367, 2000. Disponível em: <https://doi.org/10.1590/S0102-44502000000200005>. Acesso em: 08 jul. 2024.

SAVILLE, N.; HAWKEY, R. The English Profile Programme: The first three years. **English Profile Journal**, Cambridge, v. 1, n. 1., set. 2010. Disponível em: <https://doi.org/10.1017/S2041536210000061>. Acesso em: 10 jul. 2024.

SCARAMUCCI, M. V. R. Proficiência em LE: considerações terminológicas e conceituais. **Trabalhos em Linguística Aplicada**, Campinas, v. 36, n. 1, p. 11-22, 2000. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/tla/article/view/8639310>. Acesso em: 10 jul. 2024.

SCARAMUCCI, M. V. R. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. **Trabalhos em Linguística Aplicada**, Campinas, v. 43, n. 2, p. 203-226, 2004. Disponível em: <https://doi.org/10.1590/S0103-18132004000200002>. Acesso em: 10 jul. 2024.

SCARAMUCCI, M. V. R. O exame Celpe-Bras em contexto hispanofalante: percepções de professores e candidatos. *In*: WIEDEMANN, L.; SCARAMUCCI, M. V. R. (orgs.). **Português para falantes de espanhol: ensino e aquisição**. Campinas: Pontes, 2008. p. 175-190.

SCARAMUCCI, M. V. R. Validade e consequências sociais das avaliações em contextos de ensino de línguas. **Lingvarvm Arena**, Porto, v. 2, p. 103-120, 2011. Disponível em: <https://ler.letras.up.pt/uploads/ficheiros/9836.pdf>. Acesso em: 10 jul. 2024.

SCARAMUCCI, M. V. R. A avaliação de habilidades integradas na Parte Escrita do Exame Celpe-Bras. *In*: ALVAREZ, Maria Luisa Ortiz; GONÇALVES, Luis. (orgs.). **O mundo do português e o português no mundo afora: especificidades, implicações e ações**. Campinas: Pontes Editores, 2016. p. 391-425.

SCARAMUCCI, M. V. R.; SANTOS JUNIOR, E.; DINIZ, L. R. A. **Matilde Virginia Ricardi Scaramucci entrevistada por Elysio Soares Santos Junior e Leandro Rodrigues Alves Diniz. Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras): histórico, questões controversas e perspectivas**. Em Aberto, Brasília, v. 32, n. 104, p. 167-181, jan./abr. 2019. Disponível em: <http://dx.doi.org/10.24109/2176-6673.emaberto.32i104.4284>. Acesso em: 10 jul. 2024.

SCHEPERS, B. M. **The use of pragmatic markers by English speaking learners of Brazilian Portuguese**. 184f. Tese (Doutorado em Linguística) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2021.

SCHLATTER, M. **Celpe-Bras: avaliação, ensino e formação de professores de português como língua adicional**. Porto Alegre: UFRGS, 2014. Disponível em: <https://www.ufrgs.br/acervocelpebras/historia/>. Acesso em: 10 de maio de 2023.

SCHLATTER, M. *et al.* Avaliação de desempenho e os conceitos de validade, confiabilidade e efeito retroativo. *In*: NASCIMENTO, V. F. *et al.* (orgs.). **A redação no contexto do vestibular 2005: a avaliação em perspectiva**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2005, p. 11-35.

SCHLATTER, M.; GARCEZ, P. Educação linguística e aprendizagem de uma língua adicional na escola. *In*: RIO GRANDE DO SUL, Secretaria de Estado da Educação, Departamento Pedagógico. **Lições do Rio Grande: linguagens, códigos e suas tecnologias**. Porto Alegre: SE/DP, 2009, p. 127-172. Disponível em: https://servicos.educacao.rs.gov.br/dados/refer_curric_vol1.pdf. Acesso em: 6 mar. 2024.

SCHLATTER, M.; BULLA, G. da S.; COSTA, E. V. da. Português como Língua Adicional: uma entrevista com Margarete Schlatter. **ReVEL**, [S.l.], vol. 18, n. 35, 2020. Disponível em: <https://lume.ufrgs.br/bitstream/handle/10183/224013/001128628.pdf>. Acesso em: 10 jul. 2024.

SCHLATTER, M.; SCARAMUCCI, M.V. R. Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). *In*: PILATI, A.; VIANA, N. (org.). **Panorama da contribuição do Brasil para a difusão do português**. Brasília: FUNAG, 2021, p. 119-129.

SCHOFFEN, J. R. **Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras**. Orientadora: Margarete Schlatter. 2009. 192f. Tese (Doutorado em Linguística Aplicada) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009. Disponível em: <https://lume.ufrgs.br/handle/10183/16900>. Acesso em: 10 jul. 2024.

SCHOFFEN, J. R. *et al.* Resgatando a História do Exame Celpe-Bras: Desenvolvimento, Disponibilização e Estudos sobre o Acervo de Provas e Documentos Públicos do Exame. **Cadernos do IL**, [S. l.], v. 1, n. 55, p. 087–113, 2017. Disponível em: <https://seer.ufrgs.br/index.php/cadernosdoil/article/view/67890>. Acesso em: 9 mar. 2024

SCHOFFEN, J. R. *et al.* **Estudo descritivo das tarefas da Parte Escrita do exame Celpe-Bras**: edições de 1998 a 2017. Porto Alegre: Instituto de Letras - UFRGS, 2018. Disponível em: <https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Schoffen-et-al-2018.pdf>. Acesso em: 10 jul. 2024.

SCHOFFEN, J.; DIVINO, L. Contribuição da Linguística de Corpus para a formação de professores de PLA: sugestões a partir da descrição dos níveis de proficiência avaliados em uma tarefa da Parte Escrita do Celpe-Bras. **Letras de Hoje**, [S. l.], v. 58, n. 1, p. e44904, 2023. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/44904>. Acesso em: 20 fev. 2024.

SCHOFFEN, J. *et al.* Compilation and tagging of a corpus with Celpe-Bras texts. **In Proceedings of the 16th International Conference on Computational Processing of Portuguese**, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics, p. 627–632, 2024. Disponível em: <https://aclanthology.org/2024.propor-1.72.pdf>. Acesso em: 12 mar. 2024.

SCHOFFEN, J. *et al.* **Corpus do Celpe-Bras (CorCel)**. Em preparação.

SEGAT, G. **Estudos sobre Confiabilidade em Exames de Proficiência: O Processo de Atribuição de Notas e a Reavaliação na Parte Escrita do Celpe-Bras**. 163f. Dissertação (Mestrado em Linguística Aplicada) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

SIRIANNI, G. R. **Descrição dos níveis de proficiência em tarefa de leitura e escrita a partir de produções textuais de alunos do curso Preparatório Celpe-Bras**. Orientadora: Juliana Roquele Schoffen. 2016. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: <https://lume.ufrgs.br/handle/10183/157769>. Acesso em: 11 jul. 2024.

SIRIANNI, G. R. **Entre a certificação e a não certificação no Celpe-Bras: um estudo sobre os níveis de proficiência na Parte Escrita do exame**. 139f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020.

SIDI, W. A. **Níveis de proficiência em leitura e escrita de falantes de espanhol no exame CELPE-Bras**. 73f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

SINCLAIR, J. **The computer, the corpus and the theory of language**. [S.l.], 1999.

SOMMER-FARIAS, B. *et al.* Teaching with learner corpus data. **FLTMAG**, [S.l.], nov. 2023. Disponível em: <https://www.doi.org/10.69732/XOKY9163>. Acesso em: 11 jul. 2024.

SOSTRUZNIK, J. L. **O Uso de Conjunções em Produções Escritas no Exame Celpe-Bras: Um Estudo Baseado em Corpus**. Orientadora: Juliana Roquele Schoffen. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023. Disponível em: <https://lume.ufrgs.br/handle/10183/270404>. Acesso em: 11 jul. 2024.

SPINELLI, B., PARIZZI, F. (Eds.). **Profilo della lingua italiana**. Florença: La Nuova Italia, 2010.

SKETCH ENGINE. **Statistics used in the Sketch Engine**. Lexical Computing Ltd. 2015. Disponível em: <https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/>. Acesso em: 11 jul. 2024.

SKETCH ENGINE. **Type-Token Ratio (TTR)**. [S.d.]. Disponível em: https://www.sketchengine.eu/my_keywords/type-token-ratio-ttr/. Acesso em: 21 jan. 2024.

STEFANOWITSCH, A. **Corpus linguistics: A guide to the methodology**. Berlim: Language Science Press, 2020.

STAPLES, S.; FERNÁNDEZ, J. Corpus linguistics approaches to L2 pragmatics research. *In*: TAGUCHI, N. (ed.). **The Routledge handbook of second language acquisition and pragmatics**. Londres: Routledge, 2019, p. 241-254.

STAPLES, S. *et al.* **Multilingual Academic Corpus of Assignments - Writing and Speech (MACAWS)**, 2019.

THEWISSEN, J. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. **The Modern Language Journal**, [S.l.], v. 97, v. 1, p. 77–101, jan. 2013. Disponível em: <https://doi.org/10.1111/j.1540-4781.2012.01422.x>. Acesso em: 11 jul. 2024.

TONO, Y. Learner corpora: design, development and applications. *In*: ARCHER, D. *et al.* (eds.). **Proceedings of the Corpus Linguistics 2003 conference**. Lancaster: Lancaster University, 2003, p. 800-809.

TOSATTI, N. M. **O desempenho de estudantes de Países Africanos de Língua Oficial Portuguesa no Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras)**. Orientador: Leandro Rodrigues Alves Diniz. 2021. 260f. Tese (Doutorado em Linguística Aplicada) – Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: <https://repositorio.ufmg.br/handle/1843/36528>. Acesso em: 11 jul. 2024.

VIANA, V. Linguística de Corpus: Conceitos, Técnicas & Análises. *In*: TAGNIN, S.; VIANA, V. (orgs.). **Corpora no Ensino de Línguas Estrangeiras**. São Paulo: Hub, 2011, p. 25-96.

VICENTINI, M. P. As tarefas de compreensão oral para produção escrita no Celpe-Bras: construto e operacionalização. **Revista Estudos Linguísticos**, São Paulo, v. 48, n. 1, p. 561-580, abr. 2019. Disponível em: <https://doi.org/10.21165/el.v48i1.2141>. Acesso em: 11 jul. 2024.

VICENTINI, M. P. **As Dimensões do Construto Compreensão Oral para Produção Escrita no Exame Celpe-Bras: Percepções, Processos, Estratégias e Desempenhos ee Examinandos**. Orientadora: Matilde Virginia Ricardi Scaramucci. 2022. 241f. Tese (Doutorado em Linguística Aplicada) – Universidade Estadual de Campinas, Campinas, 2022. Disponível em: <https://repositorio.unicamp.br/acervo/detalhe/1253389>. Acesso em: 11 jul. 2024.

VYATKINA, N. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. **The Modern Language Journal**, [S.l.], v. 96, 576–598, dez. 2012. Disponível em: <https://doi.org/10.1111/j.1540-4781.2012.01401.x>. Acesso em: 11 jul. 2024.

WISNIEWSKI, K. Empirical learner language and the levels of the Common European Framework of Reference. **Language Learning**, Michigan, v. 67, n. 1, p. 232-253, jun. 2017. Disponível em: <http://dx.doi.org/10.1111/lang.12223>. Acesso em: 11 jul. 2024.

YAN, X.; FAN, J. Reliability and dependability. *In*: FULCHER, Glenn; HARDING, Luke. **The Routledge Handbook of Language Testing**. Londres: Routledge, 2022, p. 477-494.

ANEXOS

Anexo A - Tarefa IV (Azulejos Valiosos) da edição de 2015/2



Certificado de Proficiência em Língua Portuguesa para Estrangeiros

Tarefa 4 | Azulejos valiosos

Página 8

Você é morador de Belém e está inconformado com a situação dos casarões históricos da cidade. Com base na matéria "Azulejos valiosos", escreva uma carta aberta endereçada à prefeitura municipal, para ser publicada em jornais locais. Seu texto deverá explicar o problema e argumentar sobre a necessidade de se tomarem medidas imediatas para solucioná-lo.

Azulejos valiosos

Quatro casarões do século XIX são alvo de roubos e depredações em Belém.

A capital paraense já foi considerada uma das cidades brasileiras com maior variedade de azulejos, que coloriam as fachadas e o interior de residências. Boa parte deles foi importada da Europa, principalmente na virada do século XIX para o XX, auge da produção de borracha. Da década de 1970 para cá, no entanto, mais de 50% dos azulejos se perderam. Este ano, a situação parece ter se agravado. Desde fevereiro, pelo menos quatro casarões foram alvo de vandalismo. O assunto vem se espalhando pela capital paraense, e há quem suspeite de encomenda de roubos.

Uma das construções depredadas é o Palacete Vitor Maria da Silva, batizado com o nome de seu antigo dono, inspetor de obras do estado do Pará no governo Augusto montenegro (1901-1909).

Os azulejos foram encontrados dias depois, em cacos, e estão no Laboratório de Conservação e Restauração da UFPA (Lacore): "Recebemos aqui no laboratório mais de 1.000 fragmentos de azulejos e estamos montando o quebra-cabeça para ver a que painéis pertencem. Vamos limpar e organizar o material até o fim de junho. Só depois será decidido o que pode ser restaurado ou refeito", explica Thaís Sanjad, coordenadora do Lacore.

Há cerca de um ano, o Departamento do Patrimônio Histórico, Artístico e Cultural (Dphac) iniciou o processo de tombamento do casarão. Segundo a diretora Thaís Toscano, o procedimento é demorado, por ser necessário documentar detalhes arquitetônicos e históricos da

construção. "No caso deste imóvel, os detalhes se tornam mais elaborados, dado o nível artístico dos painéis de azulejo. Mas o local já foi interditado".

A proteção do palacete parece encaminhada, mas a situação na cidade causa preocupação, já que outros três casarões tiveram azulejos do século XIX furtados. "Foram roubos pontuais muito estranhos. O Palacete Vitor Maria da Silva tem um dos interiores mais bonitos da cidade, mas por fora é muito simples, não chama atenção. As pessoas que invadiram devem ter sido encarregadas de roubar azulejos. Ou então foi uma tentativa de desqualificação da propriedade, para que se possa fazer o que quiser com o patrimônio", suspeita a arquiteta e urbanista Cláudia Nascimento. A superintendente do Iphan no Pará, Maria Dorotéia Lima, concorda: "Tudo indica que há um mercado de azulejos na cidade, até porque os exemplares fora das áreas tombadas não têm qualquer proteção, o que pretendemos fazer em breve", disse. Enquanto as investigações não forem concluídas, os poucos exemplares de azulejos que ainda restam aumentam cada vez mais de valor.



Jornal Em Dia, jun. 2012 (adaptado).

Anexo B - Parâmetros de Avaliação da Parte Escrita



PARÂMETROS DE AVALIAÇÃO DA PARTE ESCRITA DO EXAME CELPE-BRAS

5 – Configura adequadamente a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada. Recontextualiza apropriadamente e de maneira autoral as informações necessárias para cumprir o propósito interlocutivo de forma consistente. Eventuais inadequações ou equívocos não comprometem a configuração da interlocução. Produz um texto autônomo, claro e coeso, em que os recursos linguísticos acionados são apropriados para configurar a relação de interlocução no gênero solicitado, e possíveis inadequações raramente comprometem a fluidez da leitura.

4 – Configura a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada. Recontextualiza apropriadamente as informações necessárias para cumprir o propósito interlocutivo, mas possíveis equívocos ou incompletudes podem fragilizar, em momentos localizados, a consistência da interlocução. Os recursos linguísticos acionados são apropriados para configurar a relação de interlocução no gênero proposto, construindo um texto claro e coeso em que possíveis inadequações podem comprometer, em momentos localizados, a fluidez na leitura.

3 – Configura a relação de interlocução no gênero discursivo proposto na tarefa, realizando a ação solicitada, ainda que a consistência da relação de interlocução possua algumas falhas. Pode recontextualizar de forma pouco articulada e/ou equivocada ou não recontextualizar informações necessárias para cumprir o propósito dentro do contexto de produção solicitado. Os recursos linguísticos acionados são apropriados, podendo apresentar limitações ou inadequações que podem prejudicar, em alguns momentos, a configuração da interlocução no gênero proposto. Problemas de clareza e coesão podem ocasionar, em alguns momentos, dificuldades na leitura.

2 – Configura a relação de interlocução de forma pouco consistente, realizando superficialmente a ação solicitada. Pode estabelecer uma relação de interlocução próxima à solicitada, não cumprir propósito(s) menor(es) e/ou apresentar problemas na construção do gênero. Pode apresentar trechos do texto que remetem a um gênero diferente, comprometendo a relação de interlocução. A relação entre o propósito do texto e a interlocução configurada não é clara ou não é totalmente adequada. Pode não recontextualizar informações que seriam necessárias para a configuração adequada da interlocução ou não articular claramente essas informações. Equívocos de compreensão podem comprometer parcialmente o cumprimento do propósito. Os recursos linguísticos acionados são limitados e/ou inadequados, podendo prejudicar parcialmente a configuração da relação de interlocução no gênero solicitado. Problemas de clareza e coesão podem ocasionar, em diferentes momentos, dificuldades na leitura.

1 – Configura com problemas recorrentes ou não configura a relação de interlocução solicitada, realizando muito superficialmente ou não realizando a ação solicitada. Remete-se ao tema, mas pode não considerar o contexto de produção e não construir o gênero discursivo proposto ou apresentar problemas recorrentes na sua construção. Não recontextualiza informações suficientes para o cumprimento do propósito comunicativo, considerando a relação de interlocução configurada. OU Pode apresentar equívocos graves e/ou frequentes de compreensão que comprometem o cumprimento do propósito. Os recursos linguísticos acionados são muito limitados e/ou inadequados, o que prejudica substancialmente o cumprimento do propósito e a configuração da relação de interlocução, comprometendo a construção do gênero solicitado. Problemas frequentes de clareza e coesão ocasionam, em vários momentos, problemas na leitura.

0 – Não configura, ou configura de forma equivocada, a relação de interlocução, não realizando a ação solicitada. OU Trata de outro tema. OU Demonstra problemas generalizados de compreensão, impedindo o cumprimento do propósito e a configuração da relação de interlocução E/OU Limita-se a reproduzir o(s) texto(s)-base(s), sem marcas de autoria. OU Ignora completamente os texto(s)-base(s). E/OU Problemas generalizados de clareza e coesão e/ou inadequações linguísticas impedem a configuração da relação de interlocução no gênero solicitado, comprometendo a compreensão geral do texto. OU A produção é insuficiente para a avaliação.

Fonte: Documento-Base do Exame Celpe-Bras (Brasil, 2020, p. 39)

Anexo C - Tutorial de Digitações do Corpus de Textos de Examinandos do Celpe-Bras

GRUPO AVALIA

DIGITAÇÕES DO CORPUS DE TEXTOS DE EXAMINANDOS DO CELPE-BRAS

PASSO A PASSO

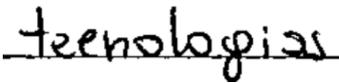
DIGITAÇÃO:

- 1 - Criar o arquivo
- 2 - Nomear o arquivo exatamente como está nomeada a respectiva imagem (AnoEdiçãoTarefaParticipanteNota) - exemplo: **20152t1p70n4**
- 3 - Digitar o texto exatamente conforme a imagem
- 4 - Atualizar planilha de controle colocando ok nos textos já digitados

REVISÃO:

- 1- Copiar o arquivo digitado e colar na pasta REVISADO
- 2 - Nomear o arquivo exatamente como está nomeada a respectiva imagem (AnoEdiçãoTarefaParticipanteNota) - exemplo: **20152t1p70n4 e ADICIONAR UM r NO FINAL: 20152t1p70n4r**
- 3 - Revisar o texto exatamente conforme a imagem
- 4 - Atualizar planilha de controle colocando ok nos textos já digitados

Cuidados a serem tomados:

- Não corrigir erros de gramática/ortografia (os “erros” devem ficar exatamente como estão escritos no texto original) - **CONSIDERAR O QUE FAZ SENTIDO QUANDO A LEITURA ESTIVER DIFICULTADA.** Atentar-se à letra do examinando, analisando outras partes do texto para tentar reconhecer o que está escrito.
- Exemplo:  → pode parecer “tecnologias”, mas analisando o restante do texto e levando em consideração o contexto, entende-se que a palavra é “tecnologias”.

- Manter a formatação do texto original (centralizado, alinhamento, parágrafos, etc...), exceto quando
 - Há várias informações de ordem diferente na mesma linha (exemplo: informações de cabeçalho de carta, como data, remetente, destinatário, assunto etc, separados apenas por espaço). Neste caso, manter o alinhamento (se está para a direita, manter na direita, se está centralizado, manter centralizado, se está para a esquerda, manter na esquerda), mas deixar cada informação em uma linha própria para que informações diferentes não sejam contabilizadas como uma frase única quando, na verdade, não é.
 - Caso haja uma frase cortada entre duas linhas (exemplo “Moradora de Belém” em uma linha e “há 30 anos” em outra, neste caso, escreve-se “Moradora de Belém” na mesma linha.
 - **NÃO É NECESSÁRIO FAZER QUEBRAS DE LINHA! Pular linha quando houver parágrafo, não no meio da frase.**
- Rasuras não são consideradas para digitação (podem aparecer entre parênteses, riscadas com uma linha ou totalmente riscadas)
- Tentar compreender/decifrar palavras não muito legíveis (em caso de muita dificuldade, pedir ajuda)
- Desidentificação: Trocar apenas os nomes de pessoas que não estão na tarefa por:

Substituir na seguinte ordem:

1º nome

1 nome - Bibiana

2 nomes - Bibiana Arminda

3 nomes - Bibiana Arminda Flora

Sobrenome

1 sobrenome - Cambará

2 sobrenomes - Cambará Terra

3 sobrenomes - Cambará Terra Amaral

- Em caso de: Lula **da** Silva → Bibiana **de** **Cambará**
- Em caso de sigla, seguir a lógica do nome. Exemplo: em caso de Bibiana Arminda Cambará Terra → B. A. C. T.
- Em caso de assinatura, substituir por [Assinatura], entre []
- Em caso de apelido: Bibi