UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS

LUIZA SARMENTO DIVINO

ÍNDICES LEXICAIS DE ANÁLISE PARA A CARACTERIZAÇÃO DOS NÍVEIS INTERMEDIÁRIO E AVANÇADO SUPERIOR NO EXAME CELPE-BRAS: UMA PESQUISA GUIADA POR CORPUS

Porto Alegre
2021

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS

LUIZA SARMENTO DIVINO

ÍNDICES LEXICAIS DE ANÁLISE PARA A CARACTERIZAÇÃO DOS NÍVEIS INTERMEDIÁRIO E AVANÇADO SUPERIOR NO EXAME CELPE-BRAS: UMA PESQUISA GUIADA POR CORPUS

Trabalho de conclusão de curso apresentado como requisito parcial para o grau de Licenciada em Letras pela Universidade Federal do Rio Grande do Sul.
Orientadora: Profª. Drª. Juliana Roquele Schoffen

Porto Alegre
2021

À minha avó Jurita, por todo o amor e cuidado.

AGRADECIMENTOS

À minha mãe, Simone, que é minha maior inspiração. Obrigada por toda a calma e compreensão, por todo o carinho e afeto. Obrigada pelo companheirismo. Obrigada por sempre me incentivar a fazer o que eu acredito e por me mostrar que as coisas vão dar certo.

Ao meu irmão, Vitor, aos meus avós, Luci e Juca, e à minha bisavó, Gertrudes, por sempre torcerem por mim e por me darem força. Vocês são o meu porto seguro.

Ao meu pai, João, por me entender e me tranquilizar. Obrigada pela amizade e por sempre me apoiar nas minhas escolhas.

À minha família de Campinas, com quem compartilho felicidades, tristezas, vitórias e derrotas. Obrigada por sempre estarem próximos, mesmo a quilômetros de distância.

À minha orientadora, Juliana Schoffen, a quem tenho grande admiração desde que ingressei no curso de Letras. Obrigada por todos os ensinamentos e pela oportunidade de trabalhar contigo ao longo desses (quase) cinco anos. Tu és um exemplo para mim.

Aos membros do grupo Avalia, por terem me proporcionado a sensação de fazer parte de algo maior. Obrigada por todas as sextas-feiras de discussão e por todas as reflexões, que foram de grande contribuição para meu desenvolvimento acadêmico e pessoal. Obrigada pela parceria e por despertarem em mim a vontade de dominar o mundo.

À Alessandra, à Carla, à Isadora e à Julia, pelas horas disponibilizadas. Obrigada pela contribuição para que este trabalho se concretizasse.

Aos professores do Instituto de Letras, pelo trabalho eficiente. Obrigada por todos os conhecimentos compartilhados ao longo da graduação, por todo o carinho e compreensão.

À Karen Pupp Spinassé e ao Cléo Vilson Altenhofen, pela oportunidade de fazer com que eu descobrisse uma das minhas paixões: o ensino de alemão. Obrigada pela inspiração e incentivo.

À Ana Luiza Freitas, por todas as discussões e reflexões que me proporcionou este ano, fundamentais para que eu realizasse este trabalho. Obrigada por toda a atenção e dedicação.

À banca examinadora deste trabalho, Kaiane Mendel e Marine Matte. À Kaiane, pelos anos de trabalho conjunto, pelas trocas, e por me apresentar novas possibilidades. À Marine, por ter me auxiliado nos meus primeiros contatos com a Linguística de Corpus, pela paciência e disposição. A ambas, por serem grandes inspirações para mim e, especialmente, por terem aceitado o convite para compor a banca do meu Trabalho de Conclusão de Curso.

Às minhas amizades, por caminharem ao meu lado e por serem casa e conforto sempre que preciso. Obrigada pela cumplicidade e suporte.

Às amizades que fiz no curso de Letras, pelas horas de estudo e lazer. Obrigada por tornarem estes anos os melhores.

À Universidade Federal do Rio Grande do Sul, por todas as oportunidades.

RESUMO

Este trabalho tem por objetivo elencar índices lexicais de análise relevantes para a caracterização dos níveis Intermediário e Avançado Superior na Tarefa IV da Parte Escrita da edição de 2015-2 do Celpe-Bras, um exame que busca avaliar as práticas linguísticas em Português como Língua Adicional em contextos variados, não prevendo um falante nativo como modelo. Os níveis avaliados no Exame são definidos de acordo com necessidades de usos futuros da língua (BRASIL, 2020), e a proficiência não é medida por meio de questões de conhecimento específico de vocabulário ou gramática, mas pela capacidade do examinando de usar a língua portuguesa em situações semelhantes às que demais falantes a utilizam para se comunicar (SCHOFFEN, 2009). Para este trabalho, foram utilizados dois corpora de estudo, avaliados com as notas 2 (Intermediário) e 5 (Avançado Superior), com o intuito de realizar análises quantitativas, em que os construtos linguísticos se manifestam a partir da própria análise, sendo considerada uma pesquisa guiada por corpus (BIBER, 2009). Para efeitos de comparação, buscando itens específicos dos corpora de estudo, utilizou-se um corpus de português brasileiro geral, o Corpus Brasileiro. Com a utilização do Sketch Engine, um conjunto de ferramentas de análise de textos para uso online (KILGARRIFF et al., 2004), foi possível chegar a resultados referentes a uma maior extensão, maior número de palavras e sentenças por texto no corpus de textos nota 5, em relação ao de textos nota 2. Além disso, chegou-se a resultados referentes a uso de léxico que aponta para uma maior adequação ao gênero do discurso solicitado pela tarefa por examinandos no nível mais avançado, articulando recursos linguísticos importantes, que não são dados no material da tarefa, que inclui texto de insumo e enunciado. Estudos com corpora como estes têm o potencial de aumentar a transparência e a consistência na avaliação de proficiência (CALLIES; GÖTZ, 2015). Desse modo, este trabalho contribui para o aumento de pesquisas sobre o Exame Celpe-Bras e de informações relacionadas ao seu construto, o que tem potencial de afetar a vida de professores, estudantes e avaliadores do Exame (SCHOFFEN et al., 2017; NAGASAWA, 2018; SIRIANNI et al., 2019).

Palavras-chave: Exame Celpe Bras; Níveis de proficiência; Português como Língua Adicional; Pesquisa em avaliação guiada por corpus; Riqueza Lexical.

ABSTRACT

This paper aims to approach lexical analysis indices relevant to the characterization of the Intermediate and Upper Advanced levels of 2015-2 edition's Written Part of Celpe-Bras, a large scale language proficiency exam that seeks to assess linguistic practices in Portuguese as an Additional Language in varied contexts. The levels assessed in the Exam are defined according to the needs of future uses of the language (BRASIL, 2020), and proficiency is not measured by questions of specific knowledge of vocabulary or grammar, but by the examinee's ability to use the Portuguese language. in situations similar to those that other speakers use it to communicate (SCHOFFEN, 2009). For this study, two study corpora were used: (1) texts evaluated with grade 2 (Intermediate) and (2) texts evaluated with 5 (Higher Advanced), in order to carry out quantitative analysis, in which the linguistic constructs manifest themselves from the analysis itself, being considered a corpus-drive research (BIBER, 2009). For comparison purposes, looking for specific items from the study corpora, a corpus of general Brazilian Portuguese, the Brazilian Corpus, was used. With the use of Sketch Engine, a set of text analysis tools for online use (KILGARRIFF et al, 2004), it was possible to reach results referring to greater length, greater number of words and more sentences per text in the corpus composed of grade 5 texts, in relation to texts graded 2. In addition, results point to a more adequacy to the discourse genre requested by the task by examinees at the most advanced level, articulating important linguistic resources, not present in the assignment material, which includes input text and statement. Studies with corpora such as these have the potential to increase transparency and consistency in proficiency assessment (CALLIES; GÖTZ, 2015). Thus, this study contributes to expand the types of research conducted on the Celpe-Bras Exam adding information related to its construct, which has the potential to affect the lives of teachers, students and examiners of the Exam (SCHOFFEN et al., 2017; NAGASAWA , 2018; SIRIANNI et al., 2019).

Keywords: Celpe Bras Exam; Proficiency levels; Portuguese as an Additional Language; Corpus-driven assessment research; Lexical sophistication.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Iniciei meu trabalho como bolsista de iniciação científica com enfoque nos diferentes níveis de proficiência do Exame Celpe-Bras. Ingressei, em 2017, como bolsista voluntária, no projeto intitulado *Exame Celpe-Bras: análise do acervo de provas já aplicadas, manuais, legislação e estudos realizados*, orientado e coordenado pela Profª Drª Juliana Roquele Schoffen. Inicialmente, a minha participação como bolsista se deu pela contagem do número de examinandos por nível de proficiência ao longo de todas as edições já aplicadas do Celpe-Bras até aquele momento. A partir de então, comecei a me familiarizar com o Exame, bem como com metodologias de ensino e avaliação baseadas em tarefas, compreendendo o conceito de proficiência como "a capacidade do aprendiz de usar adequadamente a língua para desempenhar ações no mundo" (BRASIL, 2020), ou seja, como práticas de linguagem associadas ao seu contexto de uso, propósito e interlocutores envolvidos. Objetivando traçar e identificar perfis em cada uma das tarefas da Parte Escrita ao longo dos 20 anos de aplicação do Exame, foi exposto um panorama dos níveis de certificação em todas as edições até aquele momento. Os resultados do trabalho foram apresentados em eventos e disponibilizados no Acervo Celpe-Bras[1], não apenas com o intuito de contribuir com a preservação da história do Exame, como também para democratizar o acesso aos dados (SCHOFFEN et al., 2017). Estudos relatam que a disponibilização pública tanto de materiais quanto de pesquisas relacionadas ao Celpe-Bras afetam a vida de professores, estudantes e avaliadores (SCHOFFEN et al., 2017; NAGASAWA, 2018; SIRIANNI et al., 2019), beneficiados pelo aumento de informações relacionadas ao construto do Exame.

Para além disto, como membro do grupo Avalia - Avaliação de Uso de Linguagem[2], participei, ao longo da graduação, de diferentes projetos relacionados ao Exame Celpe-Bras e a avaliações de larga escala. Atuo, também, desde o início do segundo semestre de 2021, como professora bolsista de Português como Língua Adicional (PLA) no Programa Idiomas sem Fronteiras da Universidade Federal de Ciências da Saúde de Porto Alegre, onde ministro cursos de português acadêmico, cultura brasileira e, atualmente, um curso preparatório para o Exame

[1] Banco de dados disponível online que reúne as provas já aplicadas, documentos públicos, legislação e estudos referentes ao Celpe-Bras, resultado do projeto de pesquisa "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 Juliana Roquele Schoffen na UFRGS. Disponível em: http://www.ufrgs.br/acervocelpebras.

[2] Grupo de pesquisa que atua no Instituto de Letras da Universidade Federal do Rio Grande do Sul, registrado no CNPq (www.dgp.cnpq.br/dgp/espelhogrupo/8480692535262657), coordenado pela Professora Dra. Juliana Schoffen. O grupo se debruça sobre estudos referentes à avaliação de línguas.

Celpe-Bras. Por meio dessas experiências, tive a oportunidade de me aprofundar em leituras e práticas concernentes ao ensino, à avaliação e à pesquisa sobre o Exame.

Ao longo dos últimos anos, membros do grupo Avalia se dedicaram a olhar para os textos produzidos pelos examinandos em diferentes tarefas, iniciando estudos e análises sobre aspectos variados dos níveis de proficiência (SIRIANNI, 2016; KUNRATH, 2019; MENDEL; OLIVEIRA, 2020; HANAUER, 2020; 2021; DIVINO et al., 2021; SILVEIRA, 2021). A partir disso, surgiu o interesse em analisar, sob a perspectiva da Linguística de Corpus, os textos avaliados com diferentes notas em uma tarefa específica do Exame, a fim de aprimorar a descrição dos seus níveis de proficiência. Para Schoffen (2009) a relação entre o conteúdo que o teste pretende avaliar e o conteúdo que ele, de fato, avalia é um dos aspectos importantes para determinar sua validade, chamada de validade de construto (McNAMARA, 2000). O construto pode ser definido como teorias, hipóteses ou modelos que sejam capazes de dar conta de fenômenos empiricamente observáveis (SCHLATTER et al., 2005). McNamara (2000) afirma que a validação de um exame envolve pensar suas intenções, analisando também as evidências empíricas, ou seja, fatos concretos, emergentes de dados deste exame.

Este trabalho dá seguimento a esses estudos previamente realizados, buscando destacar e reconhecer padrões de uso da língua recorrentes nos níveis Intermediário e Avançado Superior para viabilizar uma descrição destes níveis a partir, também, de aspectos linguísticos. O trabalho busca responder à pergunta: **Quais índices lexicais de análise são relevantes para a caracterização de diferentes níveis de proficiência na Tarefa IV da edição de 2015-2 do Celpe-Bras?** A partir desta pergunta, buscou-se chegar a um maior detalhamento de quais são os recursos lexicais característicos dos textos avaliados com notas 2 e 5. Uma descrição mais detalhada dos níveis de proficiência permite que se tenha, também, mais informações a respeito do que usuários de PLA[3] são capazes de fazer em cada nível, o que pode ser útil tanto para professores de PLA quanto para avaliadores do Exame.

Este trabalho foi organizado em seis capítulos, incluindo a Introdução. O segundo capítulo dá conta da apresentação do Exame Celpe-Bras e suas características, com foco na Parte Escrita, objeto de estudo deste trabalho. No terceiro capítulo, são articuladas as concepções teóricas do trabalho, trazendo a noção de proficiência e o construto do Exame, bem como apresentando a área da Linguística de Corpus, incluindo algumas pesquisas em avaliação

[3] Como apontam Schlatter e Garcez (2009) o termo língua adicional (em vez de língua estrangeira ou segunda língua) enfatiza o convite para que educandos (e educadores) usem essas formas de expressão para participar na sua própria sociedade, entendendo-as como línguas de comunicação transnacional, estando, muitas vezes, a serviço da interlocução. Portanto, a distinção entre nativo/estrangeiro ou primeira/segunda língua não se faz relevante.

realizadas a partir desta perspectiva. O capítulo quatro aborda os objetivos do trabalho, os procedimentos metodológicos adotados, as especificações da Tarefa IV da Edição de 2015-2, o corpus e as ferramentas de análise utilizados neste estudo. No quinto capítulo, são apresentadas as análises realizadas, que trazem características específicas de cada corpora. No sexto capítulo, os resultados do trabalho são discutidos e são apresentadas as limitações do estudo, bem como sugestões e, por fim, as considerações finais.

2. O EXAME CELPE-BRAS

O Certificado de Proficiência em Língua Portuguesa para Estrangeiros, Celpe-Bras, foi desenvolvido pela Comissão para a Elaboração do Exame de Proficiência de Português para Estrangeiros, constituída pelo Ministério da Educação (MEC) em 1993. O Exame foi concebido para exercer efeito positivo ou redirecionador no ensino do Português como Língua Adicional (PLA) (SCARAMUCCI, 2004) e tem potencial para desempenhar um importante papel na vida de examinandos e professores de PLA (SCHLATTER et al, 2009). A elaboração do Exame foi pensada a partir da necessidade de criar um instrumento único para permitir que estudantes comprovassem sua proficiência e pudessem ingressar em cursos de graduação no Brasil, sobretudo aqueles inscritos no Programa de Estudantes-Convênio de Graduação (PEC-G), um programa do MEC/Ministério das Relações Exteriores (MRE), que busca oferecer, em universidades brasileiras, educação superior a cidadãos dos países em desenvolvimento que possuem acordos educacionais e culturais com o Brasil. A partir disso, buscou-se pensar um exame voltado para a utilização da língua em contextos da vida universitária e do cotidiano, testando a capacidade dos examinandos de ler, escrever, ouvir e falar nestas esferas de atuação da língua, em que

> – a proficiência no uso da língua portuguesa fosse analisada por meio do desempenho dos candidatos em tarefas o mais próximo possível de usos autênticos da língua;
> – as tarefas propusessem a compreensão de textos escritos e orais e a produção escrita e oral a partir desses textos;
> – os critérios de avaliação fossem holísticos e baseados nas condições de recepção e produção propostas nas próprias tarefas;
> – o resultado da avaliação fosse expresso em descritores de desempenho do examinando;
> – os parâmetros de correção tivessem como base os próprios objetivos das tarefas e os recursos discursivos exigidos para sua realização (SCHLATTER, 2014).

Para tanto, foi elaborado um exame cujas tarefas buscam aproximar-se de situações autênticas de uso da língua, pensando que uma tarefa é autêntica quando ela testa ambos

conhecimento linguístico e conhecimento específico acerca do conteúdo proposto (DOUGLAS, 2000). Esta abordagem permite que o examinando seja posto em contato com situações de uso do idioma que provavelmente encontrará fora do próprio teste (BACHMAN; PALMER, 1996).

A primeira aplicação do Exame foi em 1998, tendo tido, desde então, quase sempre duas aplicações anuais. A partir do segundo semestre de 2009, o Exame passou a ser atribuição do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). O Celpe-Bras teve, ao longo dos anos, uma crescente procura por pessoas interessadas em comprovar sua proficiência, tendo contabilizado, somando as duas edições de 2019, 11.266 inscritos (NAGASAWA, 2021).

Figura 1: Número de examinandos do Exame Celpe-Bras

Adaptado de SCHLATTER et al (2009) e DAMAZO (2012) por Ellen Yurika Nagasawa. FONTE: MEC e INEP. Atualizado em 31 março 2021.

Testes de proficiência em língua têm grande influência na vida dos indivíduos envolvidos (SHOHAMY, 2001; 2006), e os usos dos resultados destes testes podem ter efeitos positivos e negativos na vida de examinandos, sendo fatores determinantes para definir seu futuro. De acordo com Shohamy (2001), obter bons resultados em um teste pode garantir que o examinando tenha acesso a uma educação de qualidade, a bolsas de estudo, à profissão desejada, e à possibilidade de migrar para outro país e iniciar uma vida nova. Em contrapartida, obter resultados ruins pode bloquear as chances que este examinando teria de ingressar no

ensino superior, pode levá-lo a um trabalho indesejado, e pode diminuir suas chances de sair do país em que está (SHOHAMY, 2001).

O impacto de um exame como o Celpe-Bras abarca inúmeras consequências sociais, como a influência no ensino, na aprendizagem, no currículo, na preparação de candidatos (SCHLATTER et al., 2005), na elaboração de materiais didáticos e nas atitudes das pessoas envolvidas (SCARAMUCCI, 2011), tornando-se um exame de alta relevância (SCHLATTER et al., 2009). Schlatter et al. (2005) ressalta ainda, entre os impactos de um teste, as consequências em relação à interpretação dos resultados, à criação de manuais e de cursos preparatórios.

2.1 CARACTERÍSTICAS DO EXAME

O Celpe-Bras é realizado por pessoas interessadas em certificar sua proficiência em língua portuguesa, sendo exigido por universidades para o ingresso em cursos de graduação e programas de pós-graduação no Brasil, para a validação de diplomas estrangeiros e para a inscrição profissional em algumas entidades de classe.

O Exame busca avaliar as práticas linguísticas em contextos variados, e o examinando considerado proficiente é aquele que faz uso da língua de maneira condizente com a situação comunicativa proposta pelas tarefas. Segundo Bakhtin (2003), o emprego da língua se dá em forma de enunciados (orais e escritos), que refletem condições específicas, determinadas pelas diferentes esferas de atividade humana. Embora cada enunciado seja individual e singular, cada esfera de utilização da língua possui tipos relativamente estáveis de enunciado, denominados gêneros do discurso (BAKHTIN, 2003). À vista disso, cada uma das tarefas da Parte Escrita do Celpe-Bras prevê uma produção que se adeque ao gênero do discurso proposto, inserida em determinada esfera de comunicação humana, cumprindo com o propósito solicitado e configurando a relação de interlocução apropriada (SCHOFFEN, 2021). Tais especificações são explicitadas no enunciado de cada tarefa e definem o contexto comunicativo, levando em conta aspectos que contribuem para a produção de sentido.

Tarefa, para Bachman e Palmer (1996), é definida como uma "atividade que envolve indivíduos no uso da linguagem com propósito de atingir uma meta ou um objetivo particular em uma situação particular"[4]. Para os autores, a definição de tarefa, neste contexto, está

[4] No original: "(...) an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation."

relacionada tanto à atividade específica, quanto à situação em que ela ocorre. “Na Parte Escrita do Celpe-Bras, cada tarefa específica, a ser cumprida a partir de seu enunciado, está circunscrita a um evento comunicativo, em que são explicitadas as condições de produção para o participante ajustar seu texto aos propósitos dessa tarefa” (BRASIL, 2020, p. 31).

De acordo com o Documento Base do Exame Celpe-Bras (BRASIL, 2020), são, portanto, considerados para a avaliação os seguintes aspectos:

- Enunciador;
- Interlocutor;
- Propósito;
- Informações;
- Organização do texto;
- Recursos linguísticos (gramática e vocabulário).

Os elementos linguísticos que constituem o texto, como adequação lexical e adequação gramatical, são avaliados conforme a relevância de sua utilização, à medida que são empregados adequadamente para o estabelecimento da relação de interlocução dentro do gênero discursivo solicitado.

2.2 A PARTE ESCRITA

A Parte Escrita, foco deste trabalho, tem duração de 3 horas e é constituída por quatro tarefas. As tarefas I e II contêm textos de insumo multimodais, apresentando, respectivamente, um vídeo e um áudio. As tarefas III e IV têm como insumo textos escritos.

Figura 2: 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

O Celpe-Bras faz uso, na Parte Escrita, de uma avaliação holística, levando em conta o texto em sua totalidade (McNAMARA, 2000). De acordo com Schoffen (2021), a opção por este tipo de avaliação, em que todos os aspectos listados na seção 2.1 são avaliados conjuntamente, está de acordo com o construto do Exame, uma vez que contribuem de forma simultânea para o cumprimento da tarefa, rejeitando a ideia de avaliar os recursos linguísticos de maneira isolada e independente. Os parâmetros específicos de avaliação da Parte Escrita (Anexo 1) variam de acordo com a tarefa, uma vez que as configurações dos textos produzidos não se mantêm sempre as mesmas (BRASIL, 2020). Estes parâmetros são descritores que vão da nota 0 até a nota 5, relacionando-se com a adequação na configuração da relação de interlocução, na recontextualização de informações necessárias para o cumprimento do propósito, na construção coerente e coesa do texto e na utilização de recursos linguísticos.

Cada texto é avaliado por, no mínimo, dois avaliadores, que atribuem uma nota de 0 a 5. A nota final de cada uma das tarefas é calculada, através de um sistema e sem o conhecimento dos avaliadores, a partir da média aritmética das notas atribuídas pelos dois avaliadores. Caso haja discrepância (diferença maior que 1 ponto entre as duas notas), um terceiro avaliador avalia o texto, e a nota final da tarefa é, portanto, a média entre a nota deste terceiro avaliador e a nota mais próxima atribuída (BRASIL, 2020).

Segundo Callies e Götz (2015), os resultados de exames de proficiência são interpretados de acordo com inferências a respeito do que os examinandos devem ser capazes de fazer em cada um dos níveis de proficiência (BACHMAN; PALMER, 1996). As características e os critérios desses níveis são, conforme Hawkins e Filipović (2012), as propriedades que os avaliadores procuram, consciente ou inconscientemente, ao avaliar o desempenho dos examinandos. A grade de avaliação da Parte Escrita do Celpe-Bras apresenta os descritores gerais de cada um dos níveis do Exame, no entanto, cada uma das tarefas da Parte Escrita solicita que o examinando mobilize diferentes recursos a depender do gênero, do propósito e da relação de interlocução pré-estabelecidos, para que a tarefa seja cumprida de maneira satisfatória. Tendo em vista essa característica, as expectativas a respeito das habilidades linguístico-discursivas variam de acordo com cada uma das tarefas, portanto, assim como as expectativas, as interpretações da habilidade linguística do examinando devem, também, variar (DOUGLAS, 2020).

3. CONCEPÇÕES TEÓRICAS

Nesta seção, é apresentada a fundamentação teórica que embasou este estudo. Inicialmente é abordada a noção de proficiência, com enfoque no Exame Celpe-Bras. Em seguida, é apresentada a Linguística de Corpus e seus procedimentos de análise e, por fim, a terceira subseção dá conta de um panorama dos estudos sobre PLA com a abordagem da Linguística de Corpus.

3.1 A NOÇÃO DE PROFICIÊNCIA

Para Scaramucci (2000), proficiência é um compromisso com o construto teórico do exame, ou seja, consiste nas especificações definidas com base em uma análise de necessidades do público-alvo. Reconhecer o contexto de uso dos participantes é importante para a validade da avaliação (DOUGLAS, 2000). A noção de proficiência no Exame Celpe-Bras não tem como modelo o falante nativo idealizado, "proficiência, conforme o construto do Exame, é sempre relativa, isto é, apresenta níveis definidos de acordo com as necessidades de uso futuro da língua" (BRASIL, 2020). O Celpe-Bras, portanto, avalia vários níveis de proficiência com um mesmo instrumento, fundamentado em uma visão de uso da lingua(gem) baseado em propósitos sociais.

Através de um único instrumento de avaliação, o Celpe-Bras certifica quatro diferentes níveis de proficiência: o nível Intermediário, o nível Intermediário Superior, o nível Avançado e o nível Avançado Superior. Os descritores gerais dos níveis apresentam especificações distintas para a Parte Escrita e para a Parte Oral (Anexo 2). Para cada um dos níveis, há informações referentes à adequação dos examinandos no cumprimento do propósito, na produção de diferentes gêneros do discurso e no estabelecimento da relação de interlocução. Além disso, há especificações a respeito da recontextualização de informações relevantes, que se relaciona com a compreensão e a interpretação do material de insumo. O resultado final depende da Parte Escrita e da Parte Oral, e, caso o desempenho do examinando seja diferente nas duas partes, prevalecerá o menor resultado (BRASIL, 2020).

A proficiência é avaliada no Exame "não através da medição de conhecimento gramatical ou de conhecimento específico de vocabulário, mas através da capacidade de agir no mundo em situações similares às reais, possíveis de acontecer com pessoas que utilizam a língua portuguesa para se comunicar" (SCHOFFEN, 2009). Douglas (2000) afirma que, com tarefas vinculadas a situações reais de uso da língua, aumenta-se a probabilidade de que o

examinando realize a tarefa da mesma maneira que faria na situação-alvo real. O que diferencia, no Exame Celpe-Bras, um nível de proficiência do outro é a qualidade do desempenho do examinando nas tarefas propostas (BRASIL, 2020).

3.2 LINGUÍSTICA DE CORPUS

Milton (2010) sugere que, à medida que o nível de proficiência aumenta, amplia-se também o conhecimento sobre vocabulário, bem como sobre a sofisticação de seu uso. Distante de ser um elemento meramente incidental para a aprendizagem da língua adicional, atualmente defende-se que o vocabulário pode ser crucial para o desenvolvimento do desempenho da linguagem (MILTON, 2013). Uma das formas de medir o vocabulário e a riqueza lexical entre os participantes de diferentes conjuntos de textos, neste caso, o corpus de textos nota 2 e o corpus de textos nota 5, é o cálculo *Type-Token-Ratio* (TTR)[5], que diz respeito à quantidade de palavras individuais diferentes em um texto (PAQUOT, 2017).

Para que a questão linguística seja compreendida, é necessário que o estudo seja baseado em um "alicerce empírico, no qual os resultados advêm da observação de dados reais" (VIANA, 2011). A Linguística de Corpus, uma área de estudos voltada para a identificação desses aspectos, baseia-se na busca por apoiar investigações empíricas de variação e uso da linguagem (BIBER, 2009), ocupando-se da coleta e da exploração de corpora (BERBER SARDINHA, 2000). Viana (2008) define um corpus como sendo uma compilação de textos realizada de forma criteriosa em formato eletrônico, cujo objetivo é o de representar uma determinada língua ou algum aspecto pontual, possibilitando uma análise linguística. Como afirma Gablasova (2020), uma das principais funções dos corpora é fornecer informações sobre a frequência de (co)ocorrência de traços linguísticos e, por meio desta perspectiva, é possível perceber "que a linguagem é usada de modo padronizado" (BERBER SARDINHA, 2011).

Muitos dos trabalhos de corpora são iniciados com uma metodologia quantitativa, como a identificação de escolha e de utilização lexical em produções textuais. A partir disso, em alguns casos, utiliza-se também uma abordagem qualitativa, em que os resultados são contextualizados, o que possibilita não somente o reconhecimento de padrões de uso de linguagem entre textos de um mesmo corpus, como também um entendimento do que estes padrões significam e do que se pode fazer com estes resultados. Portanto, através da comparação entre os resultados quantitativos de dois (ou mais) corpora, pode-se chegar a

[5] Cálculo utilizado para se chegar à diversidade lexical de um texto, apresentado com detalhes na seção 4.4.

generalizações gramaticais em cada um dos corpora, entendidas como o acúmulo de padrões de ocorrências individuais de palavras e frases (SINCLAIR, 1991). Gablasova (2020) ainda destaca que estas informações se fazem muito úteis no que diz respeito à pesquisa sobre o uso de uma língua adicional, pois evidenciam características linguísticas típicas e particulares de domínio da língua por grupos específicos de falantes. A autora ainda afirma que os corpora de referência devem ser semelhantes em todos os aspectos principais e diferir apenas no que diz respeito à variável cujo efeito deseja-se observar que, no caso desta proposta de estudo, são os diferentes elementos lexicais que definem os diferentes níveis de proficiência.

3.2.1 TIPOS DE CORPORA

Neste trabalho, foi realizada a comparação entre corpora, por isso torna-se importante elucidar os conceitos de **corpus de estudo** e **corpus de referência**. O primeiro deles se refere ao corpus cuja linguagem dos textos nele contido se deseja observar e descrever. O segundo é um corpus 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. É através de comparações e contrastes que se consegue delimitar o que é específico em um corpus de estudo.

Os subcorpora utilizados neste trabalho são resultado de um exame de proficiência em língua adicional, o que os configuraria, segundo a literatura da área, como **corpora de aprendizes**. Granger (2009) define estes tipos de corpora como "coleções eletrônicas de textos produzidos por aprendizes de segunda língua e/ou língua estrangeira, organizados de acordo com critérios específicos", sendo o critério, no caso deste trabalho, a nota obtida pelos examinandos (nota 2 e nota 5). Os dados deste tipo de corpus são geralmente produzidos a partir de tarefas abertas, em que os usuários escolhem suas próprias palavras (CALLIES; GÖTZ, 2015). Esta terminologia (corpora de aprendizes) parte do pressuposto de que há aprendizes e há o padrão falante-nativo (SARMENTO, 2008) e vai, portanto, contra o referencial teórico do Celpe-Bras, visto que o Exame não tem como parâmetro um falante nativo idealizado, que seria o modelo em relação ao qual os falantes não-nativos seriam "aprendizes", e sim um falante capaz de usar a língua de maneira adequada para determinada situação comunicativa, independentemente de qual foi sua primeira língua de socialização. A depender dos fins específicos para que cada examinando vai utilizar seu resultado final, é necessário atingir determinado nível de proficiência, que varia do Intermediário ao Avançado Superior, ou seja, cada examinando busca um nível final diferente de acordo com suas necessidades e objetivos. Nesse sentido, faz-se necessário o apontamento de algumas questões

que podem contribuir para futuras discussões a respeito de outras maneiras para denominar corpora como estes:

1. Até que nível de proficiência um falante é considerado aprendiz?
2. Um falante não-nativo sempre será alocado nesta categoria? Quais perspectivas teóricas estão subjacentes a essa decisão?
3. Falantes nativos não podem ser considerados aprendizes dessa língua?
4. Quais parâmetros definem os limites entre falantes considerados aprendizes e demais falantes de uma língua?

Tais questionamentos não serão respondidos neste trabalho, porém, se fazem relevantes para demonstrar que o uso dessa terminologia não é um consenso. Por se tratar de uma terminologia que vai de encontro às concepções teóricas do Exame, os corpora de estudo utilizados neste trabalho não foram denominados desta forma (mas o termo será utilizado para fazer referência aos estudos que denominam seus corpora de estudos assim).

3.2.2 PESQUISAS EM AVALIAÇÃO DE PLA À LUZ DA LINGUÍSTICA DE CORPUS

Pesquisas sobre níveis de proficiência se fazem importantes no que diz respeito à validade de um exame, pois se propõem a fazer descrições mais robustas acerca desses níveis. Validade, aqui, é entendida como a "relação entre as especificações de conteúdo que o teste pretende avaliar e o conteúdo que ele efetivamente avalia" (SCHLATTER et al., 2005). Estas pesquisas contribuem para a validade à proporção que se tem mais informações a respeito do que está sendo avaliado, e pode-se saber se o exame avalia o que se propõe a avaliar. Pesquisas realizadas com corpora de aprendizes[6] têm, conforme Callies e Götz (2015), o potencial de aumentar a transparência e a consistência na avaliação de proficiência, permitindo validar e promover a maneira como a proficiência em língua é avaliada. O uso de informações e resultados obtidos através de análises de corpora de aprendizes permite que também avaliadores de exames de proficiência tenham descritores mais refinados sobre os níveis, centrados em textos autênticos. Isso permite, também, que elaboradores de exames de proficiência adotem abordagens baseadas em dados empíricos para pensarem a avaliação (CALLIES; GÖTZ, 2015).

[6] *Learner Corpora*

Através de uma busca no Google Scholar[7] por trabalhos na área de Linguística de Corpus com foco em PLA, com os termos "corpus ple", "corpus aquisição ple", "corpus aquisição pla", "corpus aprendizes português", "corpus português língua estrangeira" e "learner corpus portuguese", foi percebido que os estudos utilizando corpora de aprendizes de português ganharam força na última década, tendo aumentado o número de trabalhos ao longo dos anos (por exemplo SOUZA, 2012; BENTO, 2013; STICHINI, 2014; MARTINS, 2015; SHEN, 2017). Estes dados sugerem que este seja considerado um campo de pesquisa relativamente novo.

Mesmo em estágio inicial, estudos com corpora de aprendizes já apontam para um grande impacto na área de avaliação de proficiência em língua adicional (CALLIES; GÖTZ, 2015), nos estudos de aquisição de língua adicional e no seu ensino e aprendizagem (MENDES, et al. 2016). Apesar disso, ainda não se tem tantos estudos capazes de potencializar este impacto, pois, segundo Mendes et al. (2016), não há, com exceção da língua inglesa, recursos suficientes para viabilizar tais estudos. Um exemplo disso são os programas existentes para análise de corpora, que apresentam, em muitos casos, limitações nas suas ferramentas quando utilizados corpora de outras línguas que não a língua inglesa Além disso, não há, também, uma quantidade tão grande de corpora de aprendizes compilados e disponibilizados para pesquisa. Neste sentido, trabalhos que se propõem a complicar e analisar corpora de aprendizes buscam preencher esta lacuna.

Para o caso da língua portuguesa, há algumas iniciativas de compilação de corpora de aprendizes como: o projeto **Recolha de dados de Aprendizagem do Português Língua Estrangeira**[8], que resulta de uma parceria entre o Instituto Camões e o Centro de Linguística da Universidade de Lisboa (FLUL), constituído por 470 produções escritas, produzidas por 397 informantes, falantes de 28 diferentes línguas maternas; o **Corpus de Aquisição de L2 (CAL2)**[9], da NOVA University de Lisboa (CLUNL), apresentando 1607 textos escritos, além de entrevistas, produzidos por adultos e crianças; e o **Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2)**[10], compilado na Universidade de Coimbra (CELGA), constituído por 629 textos, produzidos por 458 sujeitos, falantes de 39 línguas maternas.

Um dos estudos apoiados no PAEPL2 (MATTE; GOULART, 2021) se propôs a analisar os níveis de proficiência em português como língua adicional partindo dos pacotes

[7] https://scholar.google.com.br/
[8] http://www.clul.ulisboa.pt/recurso/recolha-de-dados-de-ple
[9] http://cal2.clunl.fcsh.unl.pt/index.html
[10] http://teitok2.iltec.pt/peapl2/#http://teitok.iltec.pt/peapl2/

lexicais[11] apresentados em cada nível, bem como suas funções comunicativas. O trabalho foi realizado com 517 textos distribuídos entre os níveis iniciante e intermediário. As conclusões do trabalho apontam para semelhanças e diferenças entre a utilização de pacotes lexicais nos corpora entre os dois níveis, ressaltando a distinção entre o uso de vocabulário e conhecimento de estruturas gramaticais. Além disso, as autoras apontam para a necessidade de aplicação destes resultados, baseados em dados empíricos, no ensino, redirecionando práticas docentes, bem como a produção de materiais.

Além dos corpora acima apresentados, há também o **Corpus de Português Língua Estrangeira/Língua Segunda (COPLE2)**[12], um corpus de aprendizes de português da Universidade de Lisboa (FLUL), contendo 966 produções escritas coletadas em duas situações distintas, sendo uma delas resultados de provas de português aplicadas em contexto de aula no Instituto de Cultura e Língua Portuguesa (ICLP), e outra a partir de produções resultantes do exame de proficiência aplicado no Centro de Avaliação de Português Língua Estrangeira (CAPLE)[13], representando 14 línguas maternas distintas (MENDES et al., 2016). Os textos selecionados para a compilação deste corpus estão inseridos em cinco diferentes níveis de proficiência de acordo com o Quadro Europeu Comum de Referência para as Línguas (CEFR), indo do nível A1 ao nível C1.

A partir do acesso ao COPLE2, inúmeros trabalhos no âmbito da descrição de aspectos específicos de utilização linguística por aprendizes de português foram publicados (por exemplo MENDES et al., 2016; CASTELO et al., 2016; TALHADAS, 2016; ANTUNES, 2017; DEL RIO, 2019; ESTRELA; AMARO et al., 2020), em sua maioria, em centros de estudos portugueses. Em artigo destinado à apresentação e à divulgação de resultados iniciais a partir da análise do corpus (MENDES et al., 2016), os autores afirmam que os dados provenientes deste corpus fornecem contribuições diretas para aplicações e recursos didáticos para a aprendizagem de português e, por se tratar de um corpus compilado a partir de textos produzidos em contexto de avaliação de proficiência, podem revelar características relevantes sobre os níveis de língua adicional do Quadro Europeu Comum de Referência para as Línguas (CEFR).

A especificação da descrição dos níveis de proficiência do Celpe-Bras tem sido, ao longo dos últimos anos, tema de diversos estudos (SIDI, 2002; SCHOFFEN, 2003; 2009; FORTES, 2009; EVERS, 2013; SIRIANNI, 2016; 2020; QUEIROZ, 2017; KUNRATH, 2019;

---

[11] Sequência de um número de itens lexicais em uma determinada amostra de texto.
[12] http://teitok.clul.ul.pt/cople2/index.php?action=files
[13] Certificado de proficiência oficial de português europeu

MENDEL, 2019; SOUZA NETO, 2019; OLIVEIRA, 2020; HANAUER, 2020; 2021; DIVINO et al., 2021; SILVEIRA, 2021). Muitos desses trabalhos relacionam os diferentes níveis de proficiência com a recuperação, mais ou menos consistente, das informações do texto de insumo, bem como com a coesão e coerência textual. Alguns deles se referem a aspectos relacionados ao uso de recursos linguístico-discursivos para a construção desta coesão, determinante para a diferenciação entre os níveis. Esta pesquisa se propõe a complementar estes estudos já realizados, porém, com enfoque no léxico, visto que não há, ainda, trabalhos sobre níveis de proficiência do Celpe-Bras com este enfoque.

4. PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, serão apresentados os objetivos deste trabalho, bem como os procedimentos adotados para atingir tais objetivos. Inicialmente, são apresentados a pergunta de pesquisa e seus desdobramentos, também, em forma de perguntas. Na sequência, são apresentados os corpora utilizados nesta pesquisa, o enunciado da tarefa, o texto de insumo, o programa utilizado para a análise dos dados e, por fim, os procedimentos para as análises.

4.1 OBJETIVOS DO TRABALHO

Com este trabalho, busca-se começar a entender que tipos de pesquisa, à luz da Linguística de Corpus, podem revelar diferenças significativas entre os níveis de proficiência na Parte Escrita do Exame Celpe-Bras. O objetivo central deste trabalho é responder à pergunta de pesquisa: : **Quais índices lexicais de análise são relevantes para a caracterização de diferentes níveis de proficiência na Tarefa IV da edição de 2015-2 do Celpe-Bras?** A partir desta pergunta, buscou-se responder aos seguintes questionamentos:

- Qual a extensão de cada texto?
    - Qual a média de palavras por texto?
    - Qual a média de sentenças por texto?
    - Qual a média de palavras por sentença?
- O que se pode dizer a respeito da diversidade lexical de cada corpora?
- Quais palavras-chave são diferentes entre os textos nota 5 e nota 2?
- Quais são as palavras-chave, presentes no texto de insumo e no enunciado, mais frequentes entre os corpora?

- Quais são as palavras-chave, que não ocorrem nem no texto de insumo, nem no enunciado, mais frequentes entre os corpora?
- Quais são os pacotes lexicais chave, que não ocorrem nem no texto de insumo, nem no enunciado, mais frequentes entre os corpora?

As relações com o texto de insumo e o enunciado se fazem relevantes, visto que trabalhos anteriores apontam para uma relação entre a progressão de níveis e a cópia do material de insumo. Sirianni (2016) afirma que foram constatadas cópias diretas do texto de insumo no nível Intermediário. Para examinandos menos proficientes, o material de insumo poderia suprir eventuais lacunas linguísticas por terem acesso ao texto durante a realização da prova, enquanto examinandos mais proficientes seriam menos dependentes do material de insumo, utilizando outras estratégias de recontextualização das informações que não a cópia direta (MENDEL, 2019).

Os corpora utilizados neste trabalho e os procedimentos adotados para responder às perguntas acima são descritos nas subseções que seguem.

4.2 CORPORA DA PESQUISA

Nesta etapa, serão apresentados os corpora analisados neste estudo. Para as análises, foram utilizados dois corpora compostos por textos produzidos em resposta a uma tarefa da Parte Escrita do Celpe-Bras e avaliados como Intermediário, e outra parte, como Avançado Superior. Além disso, serão apresentados também o enunciado da tarefa e o texto de insumo.

4.2.1 TEXTOS PRODUZIDOS POR EXAMINANDOS

Para este trabalho, foram analisados textos produzidos por examinandos na edição de 2015-2 do Exame Celpe-Bras. O corpus é composto por 237 textos nota 5 e 628 textos nota 2, que foram disponibilizados em forma de cópia digitalizada pelo Inep[14].sendo considerados aqui todos os textos que obtiveram estas notas como média final da tarefa, independentemente das notas atribuídas por cada corretor ou de discrepâncias entre as notas atribuídas.

[14] O grupo Avalia solicitou, em 2017, acesso ao banco de dados do Celpe-Bras mantido pelo Inep, a fim de obter amostras de produções escritas necessárias aos interesses de pesquisa do grupo. O contato com o Inep e os pormenores da solicitação dos dados se deram através do Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC). O pedido incluiu o cadastro no portal e envio da documentação e dos termos de sigilo e compromisso das pesquisadoras envolvidas.

Tabela 1: Número de textos em cada corpora

	Textos	
	Nota 5	**Nota 2**
Número de textos	237	628
	Total: 865	

Fonte: Elaborado pela autora

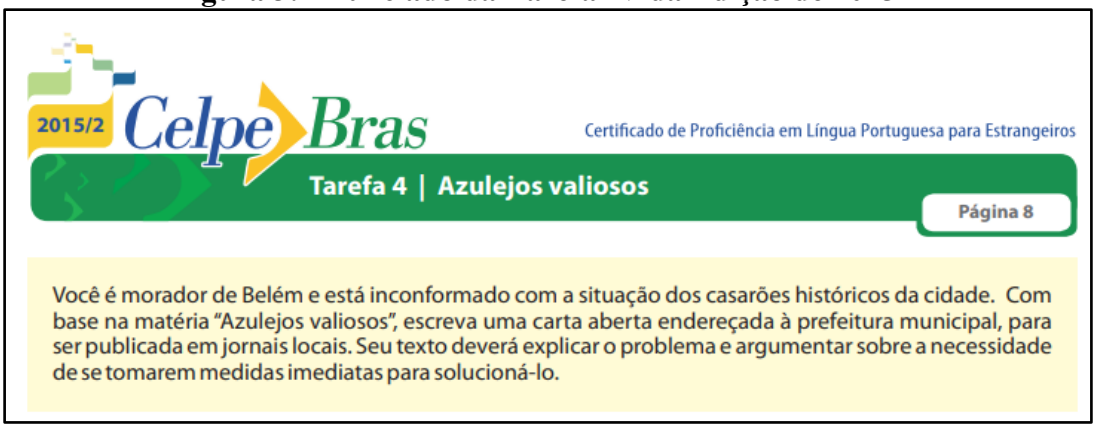
4.2.2 MATERIAIS DE INSUMO DA TAREFA ANALISADA

Também foram considerados como corpus neste trabalho os materiais de insumo da tarefa IV da edição 2015-2 do Celpe-Bras. Nesta subseção, serão apresentados o enunciado e o texto de insumo da tarefa "Azulejos Valiosos", aplicada na edição de 2015-2.

4.2.2.1 O ENUNCIADO DA TAREFA

Para o cumprimento da Tarefa IV da edição de 2015-2, o examinando deveria escrever uma carta aberta para a prefeitura municipal, que seria publicada em jornais locais, se colocando na posição de um morador de Belém inconformado com a situação dos casarões históricos da cidade. Em seu texto, o examinando deveria explicar qual era a situação e argumentar sobre a necessidade de se tomarem medidas imediatas para solucionar o problema. Abaixo, o enunciado da Tarefa IV:

Figura 3: Enunciado da Tarefa IV da Edição de 2015-2



Fonte: Caderno de Questões Celpe-Bras, edição 2015-2. Disponível no Acervo Celpe-Bras (http://www.ufrgs.br/acervocelpebras/arquivos/Provas/2015-2).

4.2.2.2. O TEXTO DE INSUMO

Além do enunciado, o examinando dispõe de um texto de insumo, que serve de ponto de partida para a produção do examinando. O texto de insumo da Tarefa IV desta edição (Anexo 3) foi uma reportagem publicada no Jornal Em Dia, em 2012, adaptado para o Exame. No texto é apresentada a situação em que se encontram quatro casarões históricos da cidade de Belém, no Pará, alvos de furtos e depredações, que danificaram painéis de valiosos azulejos trazidos da Europa há décadas. Há informações, sob a visão de especialistas, sobre o demorado processo de tombamento e a suspeita de encomenda de roubos ou de tentativa de desqualificação da propriedade para que se possa fazer o que quiser com o patrimônio.

4.3 O *SKETCH ENGINE*

Neste trabalho, foram analisados textos de diferentes níveis de proficiência através do *Sketch Engine*[15], um conjunto de ferramentas para análise de corpora, projetado para uso online (KILGARRIFF et al., 2004). Seus algoritmos são utilizados para analisar textos autênticos de bilhões de palavras, apresentando resultados instantâneos sobre a linguagem daquele corpus. Dentre as diversas funções do programa, os recursos utilizados para este trabalho foram (KILGARRIFF et al., 2014):

- **Pacotes lexicais chave[16]:** sequência contígua de um número de itens lexicais a partir de uma determinada amostra de texto ou fala, sendo eles mais relevantes em um corpus de estudo em relação a um corpus de referência;
- **Palavras-chave**[17]: lista de palavras mais relevantes presentes em um corpus em relação a um corpus de referência.

O *Sketch Engine* utiliza um método de matemática simples para identificar as palavras-chave de um corpus em comparação com outro. Este método inclui uma variável que permite ao usuário focar palavras de frequência mais alta ou mais baixa (SKETCH ENGINE, s.d.).

---

[15] Disonível em: https://www.sketchengine.eu/
[16] No original: *Multi-Word-Terms*
[17] No original: *Keywords*

A pontuação automaticamente extraída pelo programa é usada para identificar palavras-chave e também pacotes lexicais chave. Assim, é possível identificar itens que aparecem com mais frequência no corpus de estudo do que no corpus de referência. O método de matemática simples usa frequências relativas (por milhão) e, portanto, permite contrastar corpora de tamanhos diferentes (SKETCH ENGINE, s.d.).

4.4 PROCEDIMENTOS PARA AS ANÁLISES

Este é um trabalho quantitativo, em que se buscou identificar diferenças estatisticamente significativas que pudessem ser relevantes na diferenciação entre níveis. A metodologia utilizada na realização deste trabalho é a de pesquisa guiada por corpus[18]. Diferentemente da pesquisa baseada em corpus[19], a utilizada neste estudo não busca analisar padrões sistemáticos de variação e uso a partir de características linguísticas predefinidas, derivadas da teoria linguística, mas é mais indutiva, de forma que os próprios construtos linguísticos emergem a partir da análise do corpus (BIBER, 2009).

Inicialmente, foi realizada a compilação dos corpora. Para esse fim, foram realizadas a digitação e a revisão dos textos disponibilizados em forma de cópia digitalizada pelo Inep. Os 865 textos foram produzidos por examinandos e avaliados em dois diferentes níveis de proficiência. Foram digitados e revisados 237 textos nota 5 e 628 textos nota 2, que representam a totalidade dos textos avaliados em cada uma dessas notas na edição em questão. As análises quantitativas foram realizadas a partir do cálculo TTR, da extração de palavras e pacotes lexicais chave com a utilização do *Sketch Engine*[20], e dos resultados do teste de significância estatística *Log-Likelihood* (LL)[21].

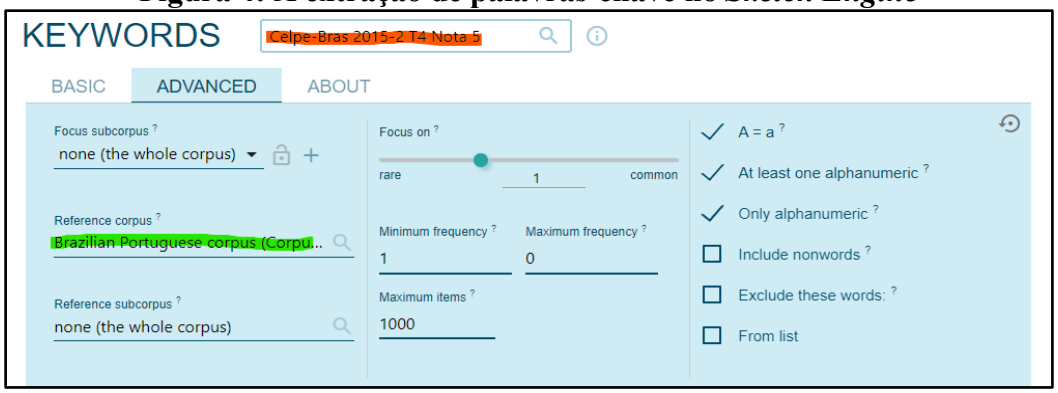
O TTR apresenta dados referentes à riqueza lexical dos textos, relacionado à diversidade de palavras. Para chegar a este percentual, é preciso multiplicar por 100 a quantidade de *types*, entendido como o número de diferentes formas de palavras, e dividir pela quantidade de *tokens*, entendido como o número total de palavras no corpus (BIBER et al., 2002). Quanto mais *types* houver em comparação com o número de *tokens*, ou seja, quanto maior o valor percentual, maior será a riqueza lexical do corpus.

---

[18] *Corpus driven research*
[19] *Corpus based research*
[20] O Sketch Engine é um pacote de ferramentas online de análise de textos que trabalha com grandes amostras de linguagem, para identificar o que é típico e frequente em um corpus.
[21] Disponível em: http://ucrel.lancs.ac.uk/llwizard.html

A extração de palavras-chave com o *Sketch Engine* acontece de maneira automática com a ferramenta *Keywords*. É preciso selecionar o corpus de estudo, do qual se deseja obter as palavras-chave, e escolher um corpus de referência, que serve para contrastar com o corpus de estudo, podendo extrair dele o que é peculiar e específico. A busca por palavras-chave permite que se identifique quais palavras são únicas ou características de um corpus de estudo em relação a um corpus de referência, por isso, se faz relevante a comparação de um corpus de estudo com outro corpus da língua alvo. A lista de palavras-chave indica quais palavras são mais frequentes no corpus de estudo, ou seja, quais palavras são mais relevantes em um corpus de estudo em relação a um corpus de referência. Estas palavras são, portanto, relevantes à medida que são mais utilizadas no corpus de estudo do que no corpus de referência. Dessa forma, é possível identificar especificidades do corpus de estudo. Para apontar a proporção de uso das palavras e pacotes lexicais mais frequentes no corpus de nota 2 e no corpus de nota 5, foi feita a comparação com um corpus de português brasileiro, intitulado Corpus Brasileiro[22][23], utilizado como corpus de referência em uma etapa da extração. O Corpus Brasileiro foi disponibilizado no *Sketch Engine*. Este corpus foi escolhido para ser utilizado neste trabalho por ser considerado o "maior acervo da língua portuguesa brasileira existente" (CORPUS BRASILEIRO, s.d.). Abaixo, uma imagem apresentando a ferramenta para extração de palavras-chave, em que o corpus de estudo está marcado em vermelho e o corpus de referência está marcado em verde.

Figura 4: A extração de palavras-chave no *Sketch Engine*



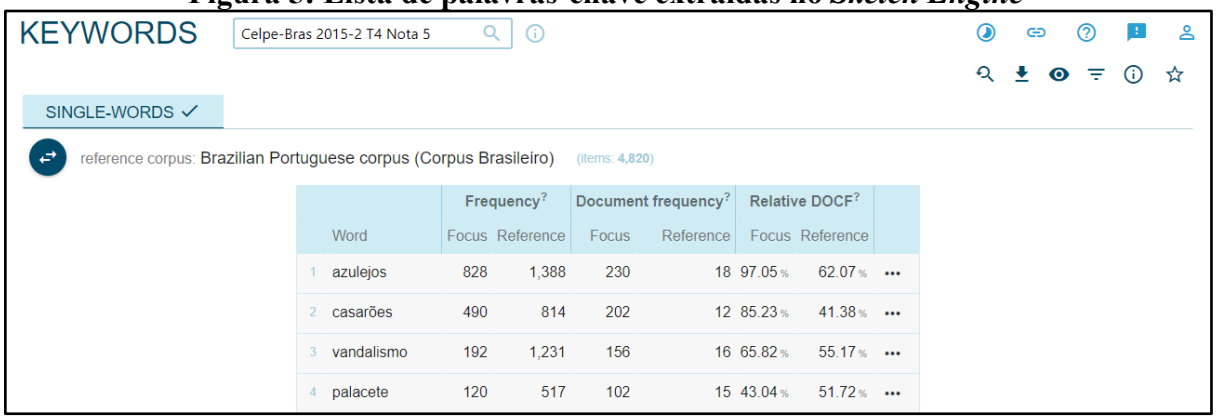
Fonte: Captura de tela. Disponível em: <https://auth.sketchengine.eu/>

[22] O Corpus Brasileiro foi compilado entre maio de 2008 e abril de 2010 pelo grupo Grupo de Estudos de Linguística de Corpus (GELC), composto pelos pesquisadores Tony Berber Sardinha, José Lopes Moreira Filho e Eli Alambert, sediado no Centro de Pesquisas, Recursos e Informação de Linguagem (CEPRIL), Programa de Pós-Graduação em Linguística Aplicada (LAEL) da PUCSP, com apoio da FAPESP. Considerado um corpus geral, ele é composto por um bilhão de palavras e seus textos estão distribuídos entre diversos gêneros discursivos de circulação pública (SKETCH ENGINE, s.d.).

[23] Projeto disponível em: http://corpusbrasileiro.pucsp.br/cb/Inicial.html

O programa, então, gera as listas das palavras chave mais relevantes no corpus de estudo, como apresentado na figura abaixo.

Figura 5: Lista de palavras-chave extraídas no *Sketch Engine*

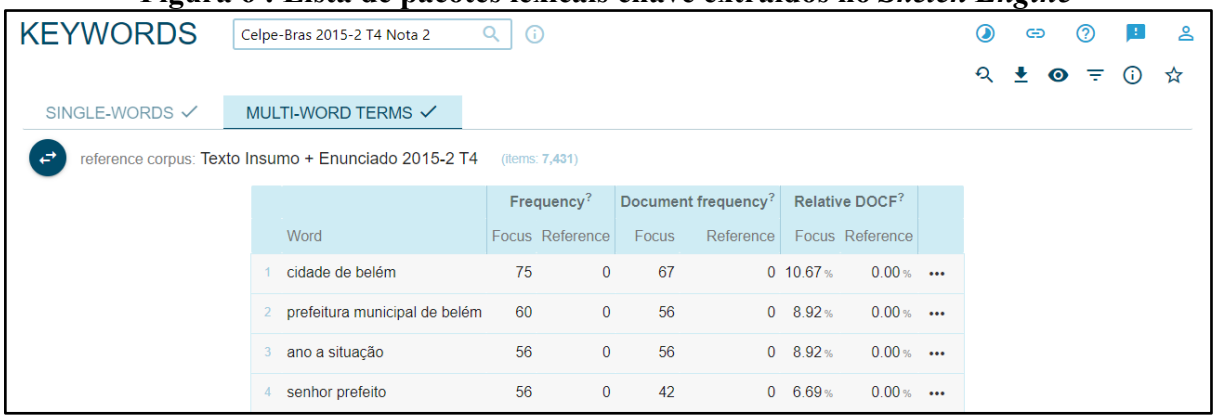


	Word	Frequency Focus	Frequency Reference	Document frequency Focus	Document frequency Reference	Relative DOCF Focus	Relative DOCF Reference
1	azulejos	828	1,388	230	18	97.05 %	62.07 %
2	casarões	490	814	202	12	85.23 %	41.38 %
3	vandalismo	192	1,231	156	16	65.82 %	55.17 %
4	palacete	120	517	102	15	43.04 %	51.72 %

Fonte: Captura de tela. Disponível em: <https://auth.sketchengine.eu/>

Além da extração de palavras-chave, foi realizada, também, a extração de pacotes lexicais chave, entendido como uma sequência de um número de itens lexicais a partir de uma determinada amostra de texto. Assim como para a extração de palavras-chave, a de pacotes lexicais chave ocorreu para identificar que pacotes lexicais são únicos e característicos de um corpus de estudo em relação a um corpus de referência. Nesta etapa, o corpus de referência foi composto pelo texto de insumo e pelo enunciado da tarefa, e os corpora de estudo foram compostos pelo corpus de textos nota 5 e o corpus de textos nota 2. O foco foi encontrar pacotes lexicais utilizados nos corpora de estudo que não estão contidos no material disponibilizado aos examinandos.

Figura 6 : Lista de pacotes lexicais chave extraídos no *Sketch Engine*



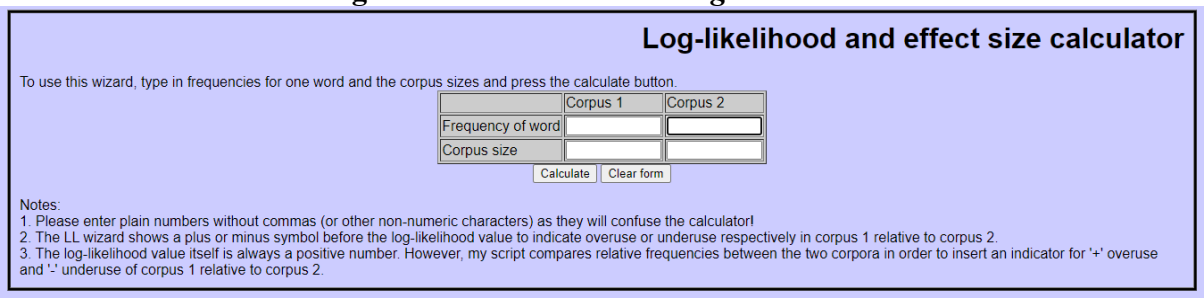
	Word	Frequency Focus	Frequency Reference	Document frequency Focus	Document frequency Reference	Relative DOCF Focus	Relative DOCF Reference
1	cidade de belém	75	0	67	0	10.67 %	0.00 %
2	prefeitura municipal de belém	60	0	56	0	8.92 %	0.00 %
3	ano a situação	56	0	56	0	8.92 %	0.00 %
4	senhor prefeito	56	0	42	0	6.69 %	0.00 %

Fonte: Captura de tela. Disponível em: <https://auth.sketchengine.eu/>

Neste caso, deve-se selecionar a opção “Multi-Word-Terms”, para a extração de uma lista de pacotes lexicais chave, em vez de “Single-Words”, que gera a lista de palavras-chave. Na figura 6, o corpus de estudo utilizado foi o de textos nota 5 e o corpus de referência foi composto pelo enunciado e pelo texto de insumo. Dessa forma, obteve-se uma lista com os pacotes lexicais específicos do corpus de textos nota 5 que não apareciam nem no texto de insumo, nem no enunciado da tarefa.

Quanto ao teste estatístico LL, este é considerado o mais adequado (RAYSON, 2002) para a comparação da frequência de palavras ou pacotes lexicais entre dois corpora. Dessa forma, chega-se a resultados que permitem afirmar se as diferenças na frequência de uso de determinadas palavras ou pacotes lexicais é aleatória ou não. Se a diferença é estatisticamente significativa, ela não é considerada aleatória. Após a aplicação do LL, quando o resultado for de 6,63 (positivo ou negativo) ou maior, significa que há uma probabilidade menor que 1% de a diferença entre os dois corpora ter acontecido aleatoriamente. Assim, o pesquisador pode estar 99% certo de que o resultado é significativo (RAYSON, 2002). A aplicação de um teste estatístico (ou de significância) serve para verificar se os dados amostrados fornecem evidência suficiente para que se possa aceitar como verdadeira a hipótese de pesquisa, precavendo-se, com certa segurança, de que as diferenças observadas nos dados não são meramente casuais. Estes testes servem para minimizar as chances de a diferença ser devida ao acaso. A diferença é considerada estatisticamente significativa se um teste indicar que é muito improvável que tenha ocorrido por acaso. Esta ferramenta pode ser encontrada online[24], e, para se obter o resultado, deve-se preencher a tabela apresentada na imagem abaixo.

Figura 7: Calculadora do *Log-Likelihood*



Fonte: Captura de tela. Disponível em: <http://ucrel.lancs.ac.uk/llwizard.html>

Para preencher a tabela corretamente, deve-se inserir o número referente à quantidade total de palavras em cada um dos corpora (na imagem, Corpus 1 e Corpus 2) no local indicado

[24] Disponível em: http://ucrel.lancs.ac.uk/llwizard.html

para o tamanho do corpus (Corpus Size). Depois disso, deve-se escolher o termo que se deseja descobrir sobre a diferença estatística de frequência de uso e inserir o número referente à quantidade de vezes que aparece em cada corpus no local indicado para a frequência da palavra (Frequency of word). Para exemplificar, utilizou-se o termo ***azulejos***, que aparece 828 vezes no corpus de textos nota 5, e 2169 vezes no corpus de textos nota 2.

Figura 8: Utilização da calculadora do *Log-Likelihood*

	Corpus 1	Corpus 2
Frequency of word	828	2169
Corpus size	53463	111630

Fonte:

Depois de preenchida a tabela, é necessário apenas clicar no botão indicado (Calculate), que os resultados aparecem.

Figura 9: Apresentação dos resultados do *Log-Likelihood*

Item	O1	%1	O2	%2	LL	%DIFF	Bayes	ELL	RRisk	LogRatio	OddsRatio
Word	828	1.55	2169	1.94	-31.84	-20.29	19.82	0.00003	0.80	-0.33	0.79

Fonte: Captura de tela.

Além da frequência de uso, o LL também leva em conta a quantidade de ocorrências de uma palavra em ambos os corpora. Por exemplo, se uma palavra for utilizada 3 vezes no corpus de nota 5 e 6 vezes no corpus de textos nota 2, o resultado do LL será menor do que ocorrerá se uma palavra for utilizada 30.000 no corpus de textos nota 5 e 60.000 no corpus de texto nota 2, mesmo que, em ambos os casos, um seja o dobro do outro. Neste trabalho, quando o resultado do LL apresenta valor negativo, isto significa que a palavra ou o pacote lexical é significativamente mais frequente no corpus de nota 2. Quando for positivo, significa que a palavra ou o pacote lexical é significativamente mais frequente no corpus de nota 5. A Figura 9 apresenta, como resultado, o valor de -31,84, o que indica que a diferença de frequência da palavra ***azulejos*** se mostra estatisticamente relevante, sendo mais utilizada no corpus de textos nota 2, ou seja, a diferença de frequência da palavra ***azulejos*** não deve ser ao acaso, não é aleatória. Para uma melhor visualização, os resultados nas figuras de comparação estão sublinhados em azul, quando a diferença for maior no corpus de textos nota 5, e amarelo, quando a diferença for maior no corpus de textos nota 2.

5. ANÁLISE DOS RESULTADOS

Para descrever cada um dos subcorpora, são apresentadas informações a respeito de sua extensão, como a quantidade de *types*, a quantidade de *tokens*, e a quantidade total de sentenças. Abaixo, uma tabela com informações de ambos os subcorpora. Para que estes valores fossem normalizados, foi realizada uma regra de 3, igualando os dois corpora, para que o número de textos fossem equivalentes por motivos de comparação. Para que os números fossem igualados, optou-se pela normalização a partir de 100 textos por corpus.

Tabela 2: Informações gerais sobre os corpora de estudo

	Textos			
	Nota 5		**Nota 2**	
	Números totais	**Valores normalizados**	**Números totais**	**Valores normalizados**
Número de Textos	237	100	628	100
Types	5.845	2.466,2	8.772	1.396,8
Tokens	53.463	22.558,2	111.630	17.775,4
Número de Sentenças	2.571	1.084,8	5.590	890,1

Fonte: Elaborado pela autora

Os resultados apontam para o fato de que os textos nota 5 são, em média, mais extensos que os textos nota 2. Este resultado sugere que, quanto maior o nível de proficiência, mais extensos são os textos.

5.1. RIQUEZA LEXICAL

Nesta etapa, serão apresentadas análises que dizem respeito à riqueza lexical em ambos os corpora de estudo. A partir de diferentes testes estatísticos e comparações entre os corpora, chegou-se a resultados que apontam para diferenças relevantes neste aspecto. No início de cada subseção, será apresentada uma tabela com o corpus de referência e o corpus de estudo utilizado para a etapa em questão.

5.1.1 EXTENSÃO DOS TEXTOS E *TYPE-TOKEN RATIO*

Tabela 3: Corpora utilizados na Etapa 1

Corpus de estudo	Celpe-Bras 2015-2 T4 Nota 2
Corpus de referência	Celpe-Bras 2015-2 T4 Nota 5

Fonte: Elaborado pela autora

Para chegar a dados referentes à extensão dos textos, foi calculada a média de palavras por texto, dividindo o número de *tokens* pelo número de textos de cada corpus, e a média de sentenças por texto, dividindo o número total de sentenças pelo número de textos nos dois diferentes níveis. Além disso, chegou-se também a resultados referentes à riqueza lexical, que tende, também, a ser maior nos textos de nota 5 do que nos textos de nota 2, como é possível verificar pelo cálculo de TTR.

Tabela 4: Informações gerais sobre a extensão e riqueza lexical dos corpora

	Textos	
	Nota 5	**Nota 2**
Média Palavras por Texto	225,6	117,7
Média Sentenças por Texto	10,8	8,8
Média de Palavras por Sentença	20,9	13,4
TTR	*10,9%*	*7,8%*

Fonte: Elaborado pela autora

Na tabela acima, podemos verificar que os textos de nota 5 têm uma média de 225,6 palavras e 10,8 sentenças, enquanto os textos nota 2 apresentam uma média de 117,7 palavras e 8,8 sentenças. Além disso, os corpora apresentam uma média de 20,9 (nota 5) e 13,4 (nota 2) palavras por sentença. Quanto ao TTR, os textos de nota 5 apresentam uma maior variação de vocabulário do que os textos nota 2, sendo essa variação, respectivamente, de 10,9% e 7,8%.

Foi percebido também que quanto menor o nível de proficiência, maior a incidência de inadequações ortográficas. Esse resultado pode ser verificado por meio da análise da lista de palavras-chave *(Keywords)* do corpus Celpe-Bras 2015-2 T4 Nota 2, utilizando como corpus

de referência o corpus Celpe-Bras 2015-2 T4 Nota 5. Entre as 50 primeiras palavras-chave do corpus dos textos de nota 2 (Anexo 4), 24 são palavras que apresentam alguma inadequação ortográfica, como adição ou supressão de acentos ou sinais gráficos, adição ou supressão de alguma letra, ou apenas a escrita diferente. A cada vez que uma palavra é escrita de maneira diferente, ela é lida pelo programa como uma nova palavra, aumentando, consequentemente, o número de *types*. Um corpus que contém um alto número de inadequações formais pode afetar o TTR, uma vez que todas as formas grafadas da mesma palavra são contadas como palavras diferentes (GRANGER, 2021). Como o Celpe-Bras não é um exame que visa a grafia correta das palavras como ponto chave para a avaliação, mas sim uma série de fatores relacionados à comunicação mais ou menos adequada, neste trabalho se optou por manter as palavras escritas como estão.

Como apontado anteriormente, a lista de palavras-chave apresenta as palavras mais relevantes em um corpus de estudo em relação a um corpus de referência. A Figura 10 apresenta a frequência *(Frequency)* com que um termo aparece no corpus de estudo *(Focus)* e no corpus de referência *(Reference)*, o número de textos *(Document frequency)* em que este termo aparece em cada um dos corpora, e o valor percentual de frequência *(Relative DOCF)*, indicando a porcentagem de textos em que o termo aparece. Abaixo, a imagem com as 10 primeiras palavras-chave dos textos nota 2 (marcados em amarelo) em relação aos textos nota 5.

Figura 10: Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_ Celpe-Bras 2015-2 T4 Nota 5

	Frecuency		Document frequency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
ciudade	46	0	26	0	4.14013	0
seguridade	23	0	19	0	3.02548	0
debe	19	0	15	0	2.38854	0
montando	16	0	13	0	2.07006	0
casaroes	16	0	12	0	1.91083	0
superindentende	16	0	16	0	2.54777	0
vitimas	15	0	13	0	2.07006	0
actos	15	0	13	0	2.07006	0
pido	15	0	15	0	2.38854	0
augusto	14	0	14	0	2.2293	0

Fonte: Elaborado no *Sketch Engine* (grifo da autora)

Na figura acima, vê-se, lado a lado, a frequência relacionada ao número de ocorrências desta palavra no corpus, o número de textos em que cada uma das palavras apareceu no corpus

e o percentual deste número de textos. Todas as palavras apresentadas foram utilizadas apenas no corpus de texto nota 2, não aparecendo em nenhum texto nota 5.

Sobre o corpus de textos nota 5 (representado na Figura 10 por *Reference*), percebe-se que todas as palavras indicadas apresentam frequência, frequência por texto e frequência relativa equivalente a 0. Quanto ao corpus de estudo, neste caso, o de textos nota 2 (representado na tabela por *Focus*), os valores variam ao longo da lista. Para verificar se a frequência de uso de palavras com inadequações ortográficas no corpus de textos nota 2 é estatisticamente significativa, isto é, não acontece por acaso, foi calculado o LL. Na figura abaixo, vê-se a mesma lista de palavras, porém, com a frequência referente ao corpus de textos nota 5 na esquerda, e ao corpus de textos nota 2 na direita. Abaixo, são apresentados os resultados do LL, lembrando que resultados iguais ou maiores do que 6,63, possuem uma chance menor do que 1% de serem considerados aleatórios.

Figura 11: Log-Likelihood comparação Celpe-Bras 2015-2 T4 Nota 5 e Celpe-Bras 2015-2 T4 Nota 2 / Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_ Celpe-Bras 2015-2 T4 Nota 5

Nota 5			Nota 2	
Item	**Frecuency**	**LL**	**Item**	**Frecuency**
ciudade	0	- 36.00	ciudade	46
seguridade	0	- 18.00	seguridade	23
debe	0	- 14.87	debe	19
montando	0	- 12.52	montando	16
casaroes	0	- 12.52	casaroes	16
superindentende	0	- 12.52	superindentende	16
vitimas	0	- 11.74	vitimas	15
actos	0	- 11.74	actos	15
pido	0	- 11.74	pido	15
augusto	0	- 10.96	augusto	14

Fonte: Elaborado pela autora

Os valores mostram que as diferenças estatísticas são relevantes. Das 10 palavras-chave acima, 8 apresentam inadequações ortográficas. As palavras ***ciudade***, ***debe***, ***actos*** e ***pido*** indicam influência da língua espanhola, pois estão escritas em espanhol. A palavra ***superindentende*** está erroneamente grafada assim no texto de insumo, o que explica sua cópia desta forma. Como não há, no programa *Sketch Engine*, uma ferramenta de identificação dessas inadequações que possibilite agrupar palavras escritas de maneiras distintas, por exemplo,

cidade e ***ciudade***, como uma mesma palavra, o TTR é alterado. Os dados apontam, portanto, que isso pode ter influenciado no cálculo do TTR. Mesmo com as inadequações ortográficas, que aumentam a quantidade de *types* presentes no corpus, os textos nota 5 apresentam um TTR maior do que o dos textos nota 2. Se as palavras grafadas incorretamente não fossem contadas como *types* distintos, a diferença de TTR entre os textos nota 5 e os textos nota 2 seria ainda maior.

5.1.2 DIFERENÇAS LEXICAIS EM RELAÇÃO AO MATERIAL DE INSUMO

Tabela 5: Corpora utilizados na Etapa 2

Corpora de estudo	Celpe-Bras 2015-2 T4 Nota 2
	Celpe-Bras 2015-2 T4 Nota 5
Corpus de referência	Corpus Brasileiro

Fonte: Elaborado pela autora

O mesmo teste realizado para chegar às palavras-chave relevantes no corpus de textos nota 2 em relação ao corpus de textos nota 5 foi realizado para chegar às palavras relevantes de ambos os corpora em relação a um corpus geral de português brasileiro, o Corpus Brasileiro. Considerando que produções escritas em resposta a tarefas que avaliam leitura e escrita são, em parte, afetadas pelo material de insumo (MENDEL, 2019), surgiu a hipótese de que a maioria das palavras-chave encontradas com este teste estivessem no texto de insumo e no enunciado, por se tratarem de palavras específicas desta tarefa. Tal hipótese foi comprovada pelos dados. Para verificar a hipótese levantada, todas as primeiras 50 palavras das listas geradas para ambos os corpora foram manualmente buscadas no texto de insumo e no enunciado. Entre as primeiras 50 palavras-chave da lista gerada utilizando o corpus de textos nota 5 como corpus de estudo (Anexo 5), 34 aparecem ou no texto de insumo, ou no enunciado, ou em ambos, representando 68% das palavras deste recorte da lista. Isto significa que o percentual de palavras-chave diferentes das encontradas no texto de insumo ou no enunciado, entre as primeiras 50 palavras-chave da lista, é de 32%. Nesta lista, assim como na anterior, são apresentadas a frequência de vezes em que a palavra apareceu no corpus de estudo e no corpus de referência, a quantidade de documentos em que esta palavra apareceu e o percentual relacionado à quantidade de documentos em que a palavra apareceu. Abaixo, a lista com as

primeiras 10 palavras-chave do corpus de textos com nota 5, sendo elas as mais relevantes em relação ao corpus de referência, que, neste caso, é o Corpus Brasileiro.

Figura 12: Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	828	1388	230	18	97.04641	62.06897
casarões	490	814	202	12	85.23207	41.37931
vandalismo	192	1231	156	16	65.82278	55.17241
palacete	120	517	102	15	43.03797	51.72414
atenciosamente	117	772	117	14	49.36709	48.27586
roubos	220	5053	145	21	61.18143	72.41379
belem	41	138	22	9	9.2827	31.03448
dphac	29	0	27	0	11.39241	0
coloriam	28	23	28	7	11.81435	24.13793
prezados	42	611	42	13	17.72152	44.82759

Fonte: Elaborado pela autora

O teste foi realizado novamente, desta vez, utilizando o corpus de textos nota 2 como corpus de estudo e o Corpus Brasileiro como corpus de referência. As palavras-chave que não constam nem no texto de insumo, nem no enunciado, na amostra das 50 primeiras palavras-chave do corpus composto pelos textos nota 2 (Anexo 6) representam 8% desta lista, sendo apenas 4 palavras, ou seja, o percentual de palavras-chave que estão no texto de insumo, ou no enunciado, ou em ambos nesta amostra é de 92%. Estes resultados vão ao encontro do que afirmam Sirianni (2016) e Mendel (2019), pois apontam para uma maior frequência de palavras copiadas do texto de insumo, ou no enunciado, por parte de examinandos menos proficientes em comparação a examinandos mais proficientes. Abaixo, um recorte das 10 primeiras palavras desta amostra.

Figura 13: Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	2169	1388	607	18	96.65605	62.06897
casarões	1012	814	482	12	76.75159	41.37931
palacete	321	517	236	15	37.57962	51.72414
vandalismo	446	1231	365	16	58.12102	55.17241
belem	225	138	140	9	22.29299	31.03448
atenciosamente	210	772	210	14	33.43949	48.27586
patrimonio	141	385	109	8	17.35669	27.58621
valiosos	313	2288	226	20	35.98726	68.96552
roubos	554	5053	360	21	57.32484	72.41379
coloriam	103	23	101	7	16.0828	24.13793

Fonte: Elaborado pela autora

Os resultados referentes ao TTR e ao percentual de palavras iguais ou diferentes do enunciado e do texto de insumo mostram que os textos de nota 5 apresentam uma riqueza lexical consideravelmente maior do que os textos de nota 2.

5.1.3 SEMELHANÇAS LEXICAIS EM RELAÇÃO AO MATERIAL DE INSUMO

Tabela 6: Corpora utilizados na Etapa 3

Corpora de estudo	Celpe-Bras 2015-2 T4 Nota 2
	Celpe-Bras 2015-2 T4 Nota 5
Corpus de referência	Corpus Brasileiro

Fonte: Elaborado pela autora

Para uma nova comparação, a lista de palavras-chave é apresentada levando em consideração a frequência de uso por documento no corpus nota 5. Para isso, se utilizou como parâmetro o percentual de frequência por documento no corpus de estudo, apresentado na penúltima coluna da Figura 14. Nesta figura, a ordem apresentada ainda não está de acordo com o percentual de frequência de uso, mas sim com a relevância.

Figura 14: Parâmetro para reordenar lista de palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	828	1388	230	18	97.04641	62.06897
casarões	490	814	202	12	85.23207	41.37931
vandalismo	192	1231	156	16	65.82278	55.17241

Fonte: Elaborado pela autora

Abaixo, a lista com as 20 palavras em ordem decrescente de acordo com o maior percentual de frequência por documento no corpus de textos nota 5.

Figura 15: Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Corpus Brasileiro

	Frecuency		Document frequency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
que	1634	16478129	237	29	100	100
de	2410	45589862	237	29	100	100
a	1321	28392471	237	29	100	100
e	1499	23366842	236	29	99.57806	100
da	912	14011812	235	29	99.15612	100
o	1040	21888135	233	29	98.31224	100
azulejos	828	1388	230	18	97.04641	62.06897
para	691	8523040	227	29	95.78059	100
cidade	642	440110	220	27	92.827	93.10345
belém	570	22826	215	23	90.7173	79.31034
os	690	7253803	214	29	90.29536	100
é	606	5945831	212	29	89.45148	100
do	561	15254651	211	29	89.02954	100
com	447	7404524	210	29	88.60759	100
dos	504	5391406	205	29	86.49789	100

Fonte: Elaborado no *Sketch Engine*

Para as análises desta etapa, neste trabalho, foram cortadas todas as palavras que não fossem substantivos, de forma que a lista foi refeita, apresentada na primeira coluna da Figura 16. Todos os primeiros 20 substantivos classificados com os mais altos percentuais de frequência por documento no corpus de textos nota 5 estão presentes ou no texto de insumo, ou no enunciado, ou em ambos. Para chegar a este resultado, todos os substantivos foram buscados manualmente no texto de insumo e no enunciado. A partir desta lista, apresentada na Figura 16, buscou-se, na lista de palavras-chave do corpus nota 2, as mesmas palavras para que fosse possível realizar a comparação. Optou-se por utilizar a lista dos substantivos mais

frequentes do corpus de textos nota 5 para comparar com o corpus de textos nota 2. Com base nessa lista, apenas com os substantivos, foi verificado, utilizando o LL, se as diferenças de frequência de uso no corpus de textos nota 2 em comparação com o de texto nota 5 eram estatisticamente relevantes. Abaixo, os resultados desta comparação.

Figura 16: *Log-Likelihood* comparação Celpe-Bras 2015-2 T4 Nota 5 e Celpe-Bras 2015-2 T4 Nota 2 / Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Texto de Insumo + Enunciado 2015-2 T4

Aparece no Material				
Nota 5			Nota 2	
Item	**Frecuency**	**LL**	**Item**	**Frecuency**
azulejos	828	- 31.84	azulejos	2169
cidade	642	- 1.95	cidade/ciudade	1432
belém/belem	611	+ 2.33	belém/belem	1182
casarões	490	+ 0.04	casarões	1012
prefeitura	376	+ 0.79	prefeitura	742
patrimônio/patrimonio	347	+ 254.97	patrimônio/patrimonio	174
municipal	226	- 13.17	municipal	622
situação	226	- 13.17	situação	623
roubos	220	- 5.67	roubos	554
vandalismo	192	- 1.55	vandalismo	446
histórico	176	+ 30.90	histórico	206
históricos	158	+ 1.43	históricos	293
cultural	144	+ 9.21	cultural	216
proteção	142	- 2.36	proteção	345
mercado	141	+ 6.51	mercado	223

Fonte: Elaborado pela autora

Observa-se, nesta lista, que as palavras mais frequentes no corpus de textos nota 2 em relação ao corpus de textos nota 5 são ***azulejos***, ***municipal*** e ***situação***, cujos resultados são 31,84, -13,17 e 13,17 respectivamente. As palavras ***patrimônio/patrimonio***, apresentam um resultado LL de +254,97, histórico, de +30,90, e cultural, de +9.21. Como um resultado LL de 6,63 ou mais significa que a diferença de frequência entre os corpora não é devido ao acaso, pode-se afirmar que ***patrimônio/patrimonio*** ocorre significativamente mais no corpus nota 5 do que no nota 2. Quanto ao restante das palavras, que representam mais da metade desta lista, não há diferenças significativas na frequência de uso. Portanto, olhar apenas para as palavras presentes no material de insumo como parâmetro de comparação entre níveis pode não revelar o que é peculiar de cada nível. Isso significa que deve-se olhar, também, para as palavras que não estão presentes neste material, pois, há evidências de que estas são mais utilizadas no

corpus de textos nota 5 do que no corpus de textos nota 2, representando 32% e 8% das 50 primeiras palavras-chave mais relevantes de cada corpora, respectivamente.

5.2 ADEQUAÇÃO AO GÊNERO DO DISCURSO "CARTA ABERTA"

Nesta etapa das análises, foram percebidas diferentes estratégias para a adequação ao gênero do discurso solicitado na tarefa. Chegou-se a estes resultados a partir do enfoque às palavras e pacotes lexicais que não estão presentes nem no texto de insumo, nem no enunciado, apresentados na sequência.

5.2.1 SAUDAÇÕES E DESPEDIDAS

Tabela 7: Corpora utilizados na Etapa 4

Corpora de estudo	Celpe-Bras 2015-2 T4 Nota 2
	Celpe-Bras 2015-2 T4 Nota 5
Corpus de referência	Corpus Brasileiro

Fonte: Elaborado pela autora

Quando realizadas as análises que possibilitaram identificar o percentual de palavras-chave diferentes do texto de insumo e do enunciado, notou-se que duas das palavras apresentadas na Figura 12 (replicada abaixo) possuem uma característica em comum.

Figura 12: Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	828	1388	230	18	97.04641	62.06897
casarões	490	814	202	12	85.23207	41.37931
vandalismo	192	1231	156	16	65.82278	55.17241
palacete	120	517	102	15	43.03797	51.72414
atenciosamente	117	772	117	14	49.36709	48.27586
roubos	220	5053	145	21	61.18143	72.41379
belem	41	138	22	9	9.2827	31.03448
dphac	29	0	27	0	11.39241	0
coloriam	28	23	28	7	11.81435	24.13793
prezados	42	611	42	13	17.72152	44.82759

Fonte: Elaborado pela autora

Este recorte, quando comparado ao texto de insumo e ao enunciado da tarefa, apresenta duas palavras, ***atenciosamente*** e ***prezado***, que não aparecem no enunciado ou no texto de insumo. Ambas as palavras são relevantes no que diz respeito à adequação ao gênero, utilizadas para saudação e despedida, respectivamente, visto que são tipicamente utilizadas em cartas abertas.

Quanto à lista das palavras mais relevantes no corpus de textos nota 2 em relação ao Corpus Brasileiro, apresentada na Figura 13 (replicada abaixo), nota-se que há apenas uma palavra que caracteriza o gênero carta aberta.

Figura 13: Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	2169	1388	607	18	96.65605	62.06897
casarões	1012	814	482	12	76.75159	41.37931
palacete	321	517	236	15	37.57962	51.72414
vandalismo	446	1231	365	16	58.12102	55.17241
belem	225	138	140	9	22.29299	31.03448
atenciosamente	210	772	210	14	33.43949	48.27586
patrimonio	141	385	109	8	17.35669	27.58621
valiosos	313	2288	226	20	35.98726	68.96552
roubos	554	5053	360	21	57.32484	72.41379
coloriam	103	23	101	7	16.0828	24.13793

Fonte: Elaborado pela autora

A palavra ***atenciosamente***, assim como acontece no corpus dos textos de nota 5, também está incluída entre as primeiras 10 palavras-chave do corpus de textos nota 2 em relação ao Corpus Brasileiro. Além desta palavra, relacionada à adequação ao gênero, há, entre as primeiras 50 palavras-chave, também ***prezados*** e ***prezado*** que, embora apareçam também neste corpus, são percentualmente menos utilizados do que no corpus de nota 5.

A partir da identificação de termos relevantes relacionados à adequação ao gênero em ambas as listas de palavras-chave, fez-se interessante analisar estes elementos característicos de cartas abertas em ambos os corpora. Após uma busca de termos característicos do gênero solicitado pela tarefa, chegou-se a ***prezado***, ***prezados***, ***prezada***, ***prezadas***, ***presado***, ***presados***, ***preçado***, ***preçados*** e ***aprezado*** que são, ao mesmo tempo, importantes para a identificação do gênero discursivo e para o estabelecimento da relação de interlocução, e ***atenciosamente***, ***atensiosamente*** e ***cordialmente***, tipicamente utilizados como despedida em cartas abertas.

Figura 17: Adequação ao gênero - Frequência

	Celpe-Bras 2015-2 T4 Nota 5		Celpe-Bras 2015-2 T4 Nota 2	
Item	**Quantidade**	**%**	**Quantidade**	**%**
Prezado	44	18.56%	80	12.74%
Prezados	42	17.72%	73	11.62%
Prezada	9	3.80%	25	3.98%
Prezadas	0	0%	4	0.64%
Presado	1	0.42%	1	0.16%
Presados	0	0%	1	0.16%
Preçado	0	0%	3	0.47%
Preçados	1	0.42%	0	0%
Aprezado	0	0%	1	0.16%
Atenciosamente	117	49.37%	210	33.44%
Atensiosamente	1	0.42%	1	0.16%
Cordialmente	12	5.06%	19	3.02%

Fonte: Elaborado pela autora

Neste caso, a análise deve basear-se na comparação destas palavras em relação ao número de textos e não ao número de palavras total, pois tendem a aparecer apenas uma vez em cada texto. Dessa forma, comparar a incidência dessas palavras com o número total de palavras pode mascarar os resultados, uma vez que, como já visto, os textos do corpus nota 5 possuem um maior número de palavras. Na Figura 18, estão as somas das quantidades e frequências dos termos utilizados para saudações e para despedidas.

Figura 18: Adequação ao gênero - Soma quantidades e frequências

	Celpe-Bras 2015-2 T4 Nota 5		Celpe-Bras 2015-2 T4 Nota 2	
Item	**Quantidade**	**%**	**Quantidade**	**%**
Saudação	97	44.72%	188	29.93%
Despedida	130	54.85%	230	36.62%

Fonte: Elaborado pela autora

No que diz respeito à saudação da carta aberta, tem-se 44,72% dos textos nivelados com nota 5 utilizando variações de ***prezado/a***, ao passo que, entre os textos de nota 2, tem-se apenas 29,93% de ocorrência. Quanto à despedida, que inclui ***atenciosamente***, ***atensiosamente*** e ***cordialmente***, há 54,85% dos textos nota 5 apresentando algum destes termos, e apenas 36,62% nos textos nota 2. Sobre as diferenças entre os níveis de proficiência nesta tarefa específica, pode-se aventar a hipótese de que examinandos em níveis mais avançados de

proficiência têm maior conhecimento dos elementos que caracterizam o gênero carta aberta, apresentando diferenças na frequência de uso de termos característicos do gênero em relação a examinandos de níveis menos avançados, indo de acordo com Mendel (2019).

5.2.2 PRONOMES QUE REMETEM AO COLETIVO

Tabela 8: Corpora utilizados na Etapa 5

Corpora de estudo	Celpe-Bras 2015-2 T4 Nota 2
	Celpe-Bras 2015-2 T4 Nota 5
Corpus de referência	Texto Insumo + Enunciado 2015-2 T4

Fonte: Elaborado pela autora

Pensou-se, também, em utilizar como corpus de referência o enunciado e o texto de insumo para extrair palavras-chave. A partir dessas comparações, é possível identificar quais palavras utilizadas nos textos dos examinandos não estão presentes nem no enunciado nem no texto de insumo em nenhum desses dois textos. A lista de palavras-chave, gerada automaticamente pelo Sketch Engine, foi mantida na sua ordem original e nenhum item foi descartado, pois se mostrou interessante observar todas as palavras relevantes. A partir da lista de palavras-chave extraída utilizando o corpus de textos nota 5 como corpus de estudo, buscou-se os mesmos itens na lista de palavras-chave obtida utilizando o corpus de textos nota 2 para realizar a comparação de diferenças estatisticamente relevantes em termos de frequência na utilização de palavras.

Figura 19: *Log-Likelihood* comparação Celpe-Bras 2015-2 T4 Nota 5 e Celpe-Bras 2015-2 T4 Nota 2 / Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Texto de Insumo + Enunciado 2015-2 T4

Não aparece no material

Nota 5			Nota 2	
Item	**Frecuency**	**LL**	**Item**	**Frecuency**
nossa	396	+ 23.34	nossa	602
como	264	+ 3.09	como	481
nosso	189	+ 28.88	nosso	231
história	141	+ 45.77	historia	128
nos	132	+ 4.24	nos	219
eu	130	- 15.75	eu	400
atenciosamente	117	+ 1.68	atenciosamente	210
esta	114	- 1.56	esta	273
ao	107	+ 16.67	ao	130
prefeito	105	+ 12.02	prefeito	139
todos	103	+ 8.49	todos	147
sendo	100	+ 8.45	sendo	142
isso	100	- 1.32	isso	239
senhor	94	+ 3.34	senhor	154
senhores	86	- 0.01	senhores	182

Fonte: Elaborado pela autora

A tabela acima aponta para 7 itens de classes gramaticais distintas, que são estatisticamente mais utilizadas no corpus nota 5 do que no nota 2, são elas ***nossa***, ***nosso***, ***história***, ***ao***, ***prefeito***, ***todos*** e ***sendo***. Quanto ao corpus de nível intermediário, apenas uma palavra, o pronome pessoal ***eu***, apresenta um resultado estatisticamente diferente, com valor de -15,75, que pode indicar que os examinandos têm uma escrita voltada mais para o individual, o que não é muito comum no gênero carta aberta, que, em geral, aponta para uma questão coletiva. Para destacar, dois itens específicos estatisticamente relevantes no corpus de textos nota 5, ***nosso*** +28,88 e ***nossa*** +23,34, são pronomes possessivos que tratam do coletivo, típico do gênero carta aberta.

5.2.3 DELIMITAÇÃO DO INTERLOCUTOR

Tabela 9: Corpora utilizados na Etapa 6

Corpora de estudo	Celpe-Bras 2015-2 T4 Nota 2
	Celpe-Bras 2015-2 T4 Nota 5
Corpus de referência	Texto Insumo + Enunciado 2015-2 T4

Fonte: Elaborado pela autora

Além das extração de palavras-chave de um corpus de estudo em relação a um corpus de referência, também é possível extrair pacotes lexicais. Após gerar a lista de palavras-chave, há a opção de visualizar os "Multi-Word-Terms", que apresenta uma lista pacotes lexicais chave, ou seja, uma sequências de palavras (neste caso, de 2 a 3 palavras em sequência) consideradas relevantes em um corpus de estudo em relação ao corpus de referência. Assim como foi realizado nas análises anteriores, a partir da lista de pacotes lexicais chave extraída utilizando o corpus de textos nota 5 como corpus de estudo, buscou-se os mesmos pacotes lexicais na lista de palavras-chave obtida utilizando o corpus de textos nota 2 para realizar a comparação de diferenças estatisticamente relevantes em termos de frequência na utilização desses pacotes. Esta lista apresenta 18 itens, pois foi utilizado o critério de analisar itens com, no mínimo, 10 ocorrências no corpus nota 5.

Figura 20: *Log-Likelihood comparação* Celpe-Bras 2015-2 T4 Nota 5 e Celpe-Bras 2015-2 T4 Nota 2 / Pacotes lexicais chave Celpe-Bras 2015-2 T4 Nota 5_Texto de Insumo + Enunciado 2015-2 T4

Não aparece no material

Nota 5				Nota 2	
Item	**Frequency**		**LL**	**Item**	**Frequency**
cidade de belém	62	+	9.80	cidade de belém	75
senhor prefeito	43	+	5.25	senhor prefeito	56
prefeitura municipal de belém	40	+	2.55	prefeitura municipal de belém	60
patrimônio/patrimonio cultural	38	+	10.05	patrimônio/patrimonio cultural	38
valor histórico	23	+	4.46	valor histórico	26
prefeitura de belém	19	+	9.78	prefeitura de belém	13
moradores de belém	19	-	0.91	moradores de belém	51
roubo de azulejos	17	+	0.97	roubo de azulejos	26
atos de vandalismo	16	+	2.27	atos de vandalismo	20
séculos xix	16	-	0.02	seculo xix	35
mercado ilegal	14	+	7.82	mercado ilegal	9
moradora de belém	13	-	4.64	moradora de belém	51
roubos de azulejos	12	-	0.70	roubos de azulejos	33
medidas urgentes	12	+	10.37	medidas urgentes	0
ano a situação	12	-	7.53	ano a situação	56
prefeito municipal	10	-	2.86	prefeito municipal	37
prefeito de belém	10	+	22.54	prefeito de belém	0
casarões antigos	10	+	4.07	casarões antigos	8

Fonte: Elaborado pela autora

Os resultados do LL apontam para uma maior quantidade de itens que apresentam maior frequência no corpus de textos nota 5 do que no corpus de textos nota 2. Para destacar dois ítens específicos, os pacotes ***prefeitura de belém*** + 9,78 e ***prefeito de belém*** +22,54 sugerem uma maior adequação ao gênero carta aberta, articulando a relação de interlocução, em diálogo direto com o interlocutor solicitado pela tarefa.

6. DISCUSSÃO DOS RESULTADOS

A análise dos textos partiu da pergunta: **Quais índices lexicais de análise são relevantes para a caracterização de diferentes níveis de proficiência na Tarefa IV da edição de 2015-2 do Celpe-Bras?** Com este estudo, foi possível identificar alguns aspectos que podem ser significativos no que diz respeito à diferenciação de níveis.

Os resultados sugerem que examinandos de nível mais avançado produzem textos mais extensos, com um número maior de palavras e de sentenças por texto do que examinandos de nível intermediário. Chegou-se, também, à conclusão de que quanto menor o nível de proficiência, maior a incidência de inadequações ortográficas, que podem ser ocasionadas por

diversos fatores. Devido ao fato de que algumas delas, a citar ***ciudade***, ***debe***, ***actos*** e ***pido***, estarem em espanhol, parece haver uma forte influência da língua espanhola no corpus de textos nota 2, o que está de acordo com os dados relativos à origem dos examinandos[25]. Além disso, tais inadequações lexicais possivelmente influenciaram o cálculo do TTR, pois cada vez que uma palavra é escrita de maneira diferente, esta conta como um novo *type*. Mesmo assim, o TTR destaca que os examinandos mais avançados apresentam uma maior relação entre o número de tokens e types, 10,9 no corpus de textos nota 5 e 7,8 no corpus de textos nota 2. Os resultados indicam, assim, uma maior riqueza lexical entre os examinandos mais avançados, mas entendemos que, se não fossem consideradas como *types* as inadequações ortográficas presentes no corpus nota 2, essa diferença seria significativamente maior.

Na lista das 15 palavras-chave que estão presentes no texto de insumo e no enunciado, apenas 6 apresentam diferenças estatisticamente relevantes na frequência de uso entre o corpus nota 5 e o corpus nota 2, sendo três delas relevantes para o corpus de textos nota 2, e 3 delas relevantes para o corpus de textos nota 5. Além de ***patrimonio/patrimônio*** +254,97, que apresenta uma diferença estatisticamente relevante que destoa do restante das palavras, olhar apenas para as palavras presentes no texto de insumo e no enunciado como parâmetro de diferenciação entre níveis de proficiência pode não revelar o que é característico de cada nível. Pode ser interessante olhar para o número de palavras presentes no texto de insumo e no enunciado em contraste com as palavras que não aparecem neste material, visto que os dados apresentam uma utilização de apenas 8% de palavras diferentes do texto de insumo e do enunciado da tarefa por parte dos examinandos de nota 2, enquanto, no corpus de nota 5, este percentual chega a 32%. Isto sugere uma maior diversidade lexical no corpus de textos nota 5, ou seja, é possível afirmar que os examinandos mais avançados mobilizam um maior repertório pessoal do que examinandos menos proficientes (MENDEL, 2019). As análises realizadas com a utilização do texto de insumo e do enunciado como corpus de referência em contraste com os corpora de notas 2 e 5 evidenciam 8 termos com diferença estatisticamente relevante, sendo 7 deles mais relevantes no corpus de textos nota 5. Esse dado, novamente, pode indicar uma maior riqueza lexical dos examinandos de nível mais avançado em relação aos examinandos de nível intermediário.

Quanto à adequação ao gênero do discurso, os resultados deste trabalho concordam com Mendel (2019), que afirma que a utilização de "recursos linguísticos relevantes para a construção do gênero do discurso é bem avaliada pelo exame" (MENDEL, 2019, p. 152). O

[25] Os falantes de espanhol são a maioria entre os examinandos do Celpe-Bras.

corpus de nota 5 apresenta diferenças na frequência de uso de termos característicos do gênero em relação ao corpus de nota 2, sendo estes termos muito mais frequentes em textos de nota 5. Quanto às saudações, 44,72% dos textos nivelados com nota 5 utilizam variações de ***prezado/a***, ao passo que, entre os textos de nota 2, apenas 29,93% apresentam estes termos. Palavras como ***atenciosamente***/***atensiosamente*** e ***cordialmente*** são utilizadas em 54,85% dos textos nota 5 e, nos textos nota 2, são utilizados apenas em 36,62%. O uso desses termos, característicos do gênero carta aberta, parece indicar que há um percentual maior de examinandos, no nível avançado, que mobilizam estas estruturas características do gênero solicitado nesta tarefa.

Além dos elementos já discutidos, mobilizados pelos examinandos para a adequação ao gênero, há também, entre a lista de palavras-chave que não estão no texto de insumo ou no enunciado, outros termos que sugerem um maior grau de adequação ao gênero por parte dos examinandos nota 5. A ocorrência estatisticamente relevante do pronome pessoal ***eu*** -15,75 no corpus de textos nota 2 e dos pronomes possessivos ***nosso*** +28,88 e ***nossa*** +23,34 no corpus de textos nota 5 indica uma maior compreensão de elementos que compõem o gênero carta aberta por parte de examinandos em níveis mais avançados, que redigem seus textos apresentando uma interlocução voltada para o coletivo. Entre a lista de pacotes lexicais chave, que não estão presentes nem texto de insumo, nem no enunciado, há também ***prefeitura de belém*** + 9,78 e ***prefeito de belém*** +22,54. Estes resultados complementam as discussões realizadas sobre a adequação ao gênero carta aberta ser realizada com mais frequência no corpus de textos nota 5, uma vez que tais pacotes lexicais são importantes para a articulação da relação de interlocução dentro do gênero solicitado. Estes resultados concordam com Sirianni (2016), que afirma que examinandos de nível Intermediário realizam uma relação de interlocução menos consistente, além de apresentarem problemas na construção do gênero.

Por meio das análises baseadas em dados empíricos, chegou-se a resultados que apontam para diferenças substanciais entre os corpora de texto nota 2 e nota 5. O estudo comprova algumas hipóteses levantadas a respeito destes dois níveis de proficiência no que diz respeito à extensão dos textos, sendo esta comprovadamente maior nos textos de níveis mais avançados. O estudo também aponta para novos índices de análise, como palavras específicas utilizadas para a adequação ao gênero carta aberta, corroborando o que já havia sido apontado por estudos qualitativos anteriores, bem como para diferenças em termos de diversidade lexical, que se mostram relevantes na diferenciação entre estes dois níveis.

6.1 LIMITAÇÕES DO TRABALHO E SUGESTÕES PARA ESTUDOS FUTUROS

Este estudo possibilitou o mapeamento de alguns índices que podem ser relevantes na diferenciação entre níveis de proficiência no Exame, sendo eles: a extensão dos textos (média de palavras e sentenças por texto e média de palavras por sentença); a diversidade lexical, também relacionada a uma maior ou menor utilização de termos incluídos no material de insumo; e a adequação ao gênero, também relacionada à articulação da relação de interlocução. O trabalho, no entanto, não apresenta análises qualitativas que podem ajudar a explicar os resultados quantitativos. O presente trabalho se limitou a compreender **o que** se usa em cada nível sendo o próximo passo entender **como** se usa o que se usa. Como sugestão de continuidade de pesquisa, estão as análises qualitativas, voltadas para as linhas de concordância dos termos estatisticamente relevantes para ambos os corpora.

Este trabalho não se ateve à comparação dos pacotes lexicais em comum entre os corpus estudados e o texto de insumo e o enunciado, tratou apenas dos que diferem de ambos. Há estudos neste sentido sendo realizados por outros membros do grupo Avalia (Hanauer, 2021; Silveira, 2021) e as relações entre os resultados desses estudos e os deste trabalho podem mobilizar mais alguns índices relevantes de diferenciação entre os níveis de proficiência.

Como sugestão para futuros estudos, pode-se pensar em como trabalhar com as inadequações ortográficas: é relevante que uma mesma palavra escrita de diversas formas seja considerada como apenas uma ou não?

Além disso, pode-se investigar corpora compostos pelos outros níveis de certificação, de forma a auxiliar em uma melhor uma caracterização de cada nível. Estes índices de análise podem vir a ser testados, também, em outras tarefas desta edição ou de outras edições do Exame. Ademais, aspectos gramaticais peculiares de cada nível podem também apontar índices relevantes para suas categorizações.

Salienta-se, ainda, que se faz necessária a reavaliação da utilização da terminologia *corpora de aprendizes*, termo comumente utilizado na Linguística de Corpus, mas que pode (e deve) ser repensado a partir de visões de língua que não prevêem um falante nativo idealizado como parâmetro, mas sim um falante que se utiliza da língua para cumprir os propósitos desejados, nas suas respectivas esferas de circulação.

CONSIDERAÇÕES FINAIS

Neste trabalho, buscamos identificar aspectos lexicais que pudessem representar indicadores de diferenciação entre os níveis de proficiência na Parte Escrita do Celpe-Bras. Para tal, foram realizadas análises quantitativas de 865 textos produzidos por examinandos em dois níveis diferentes de proficiência na Tarefa IV da edição de 2015-2.

Pensando que o Celpe-Bras é um exame que prevê a avaliação da proficiência da língua em uso, o estudo se propôs a apresentar o que foi usado pelos examinandos naquela tarefa e edição específica. Os resultados permitiram que se chegasse a conclusões que parecem ir ao encontro da noção de proficiência do Exame, principalmente no que diz respeito à adequação ao gênero, contribuindo para a validade do construto do Exame.

Por fim, este estudo pode vir a ser relevante para verificar se as categorias de análise encontradas podem servir de parâmetro para a diferenciação entre mais níveis de proficiência nesta tarefa, e entre os diferentes níveis em outras tarefas da Parte Escrita. Mais dados a respeito dos níveis de proficiência avaliados em um exame de larga escala como o Celpe-Bras podem vir a ser muito úteis para professores de PLA, pois estes podem ter acesso a informações referentes a dados empíricos a respeito de características e diferenciação de níveis de proficiência.

REFERÊNCIAS BIBLIOGRÁFICAS

BACHMAN, Lyle. F.; PALMER, Adrian S. **Language testing in practice:** Designing and developing useful language tests. Oxford: Oxford University Press, 1996.

BAKHTIN, Mikhail Mikhailovich. **Estética da Criação Verbal.** São Paulo: Martins Fontes, 2003.

BENTO, Carla Isabel da Silva. **Aquisição de português língua não materna - o conjuntivo na interlíngua da falantes nativos de neerlandês**. Dissertação de Mestrado em Ensino do Português como Língua Segunda e Estrangeira. Universidade Nova de Lisboa, 2013.

BERBER SARDINHA, Tony. A.P. **Lingüística de Corpus:** histórico e Problemática. D.E.L.T.A.,Vol.16 N 2:323-367, 2000.

BERBER SARDINHA, Tony. COMO USAR A LINGUÍSTICA DE CORPUS NO ENSINO DE LÍNGUA ESTRANGEIRA. OU: POR UMA LINGUÍSTICA DE CORPUS EDUCACIONAL BRASILEIRA. In: TAGNIN, S.; VIANA, V. (Orgs.). **Corpora no Ensino de Línguas Estrangeiras.** São Paulo: Hub. p. 301-356, 2011.

BIBER, Douglas. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In: HEINE, Bernd; NARROG, Heiko (orgs). **The Oxford Handbook of Linguistic Analysis**, 2009.

BRASIL. **Manual do Candidato do Exame CELPE-Bras**. Ministério da Educação (MEC). Brasília, 2006. Disponível em: http://www.ufrgs.br/acervocelpebras/ Manuais/manual

BRASIL. **Documento-base do exame Celpe-Bras**. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020. Disponível em: http://www.ufrgs.br/acervocelpebras/arquivos/manuais/documento-base-do-exame-celpe-bras

CALLIES, Marcus; GÖTZ, Sandra. 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.

Corpus Brasileiro. **Corpus Brasileiro**, s.d. Disponível em: http://corpusbrasileiro.pucsp.br/cb/Inicial.html. Acesso em: 19 de novembro de 2021.

Corpus Brasileiro: corpus of Brazilian Portuguese. **Sketch Engine**, s.d. Disponível em: https://www.sketchengine.eu/corpus-brasileiro/. Acesso em: 19 de novembro de 2021.

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

DIVINO, L. S.; HANAUER, I. D.; SILVEIRA, J. L. O NÍVEL AVANÇADO SUPERIOR NO CELPE-BRAS: UMA ANÁLISE DE TEXTOS DE EXAMINANDOS. In: I Congresso de

Português como Língua Estrangeira, 1, 2021, virtual. Resumo. Columbia University: **CADERNO DE RESUMOS**. p.72. Disponível em: https://www.lrc.columbia.edu/wp-content/uploads/2021/06/I-Congresso-PLE-Columbia-University.pdf

EVERS, A. **Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras**. Dissertação de Mestrado em Letras. Universidade Federal do Rio Grande do Sul, UFRGS, 2013.

FORTES, M. **Uma compreensão etnometodológica do trabalho de fazer ser membro na fala-em-interação de entrevista de proficiência oral em português como língua adicional.** Tese de Doutorado, PPG- Letras – UFRGS, 2009.

GABLASOVA, Dana. Corpora for second language assessments. In: WINKE, Paula; BRUNFAUT, Brunfaut (Orgs.). **The Routledge Handbook of Second Language Acquisition and Language Testing**. Routledge, Londres, 2020.

GRANGER, Sylviane. The contribution of learner corpora to second language acquisition and foreign language teaching: a critical evaluation. In: AIJMER, Karin. (org.), **Corpora and Language Teaching**. Benjamins, 2009.

Granger, S. Phraseology, corpora and L2 research. In Granger, S. (ed.) **Perspectives on the L2 Phrasicon:** The View from Learner Corpora. Bristol: Multilingual Matters, p. 3-21, 2021.

HANAUER, Isadora Dahmer. **CARACTERIZAÇÃO DO AVANÇADO SUPERIOR EM TAREFA INTEGRADA DE ESCRITA E COMPREENSÃO ORAL NO CELPE-BRAS**. XXXII Salão de Iniciação Científica, UFRGS.Youtube, 4 dez. 2020. Disponível em: https://www.youtube.com/watch?v=Wdt3tVXApqY

HANAUER, Isadora Dahmer. **TAREFAS DE ESCRITA E COMPREENSÃO ORAL NO CELPE-BRAS: ANÁLISE DO AVANÇADO SUPERIOR BASEADA EM CORPUS.** XXXIII Salão de Iniciação Científica, UFRGS.Youtube, 28 ago. de 2021.Disponível em: https://www.youtube.com/watch?v=0dNBGCpXbdM

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

KILGARRIFF, Adam; RYCHLÝ, Pavel; SMRŽ, Pavel; TUGWELL, David. The Sketch Engine. In: **Proceedings of the XI EURALEX International Congress**. Universite de Bretagne-Sud, p. 105–116, 2004. Disponível em: https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf

KILGARRIFF, Adam; BAISA, Vít; BUŠTA, Jan; JAKUBÍCEK, Miloš; KOVÁR, Vojtech; MICHELFEIT, Jan; RYCHLÝ, Pavel; SUCHOMEL, Vít: The Sketch Engine: Ten Years On.

In: **Lexicography ASIALEX**, Vol. 1, p. 7–36, 2014. Disponível em: http://link.springer.com/article/10.1007/s40607-014-0009-9

KUNRATH, Simone Paula. **Os descritores gerais e a progressão dos níveis de proficiência do Exame Celpe-Bras**. Tese de Doutorado em Letras. Universidade Federal do Rio Grande do Sul, UFRGS, 2019.

MARTINS, Cristina. Número e género nominais no desenvolvimento das interlínguas de aprendentes do português europeu como língua estrangeira. **Revista Científica da Universidade Eduardo Mondlane**, Séries: Letras e Ciências Sociais, Edição Especial, 2015.

MATTE, Marine Laísa; GOULART, Larissa. Lexical Bundles across levels of Proficiency in Portuguese as a Second Language: an examination of bundle function. **Letras de hoje**, Porto Alegre, v. 55, n. 4, p. 477-495, 2021.

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

MENDEL, Kaiane. **Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras**. Dissertação de Mestrado em Letras. Universidade Federal do Rio Grande do Sul, UFRGS, 2019.

MENDES, A., ANTUNES, S., JANSSEN, M. & GONÇALVES, A. The COPLE2 Corpus: a Learner Corpus for Portuguese. In: **Proceedings of LREC 2016**, Portorož, Slovenia, 2016. Disponível em: https://repositorio.ul.pt/bitstream/10451/30692/1/Mendes_et_al_COPLE2_LREC_2016.pdf

MILTON, James. The development of vocabulary breadth across the CEFR levels. In: I. Vedder, I. Bartning, & M. Martin (org.), **Communicative proficiency and linguistic development: intersections between SLA and language testing research**, p. 211-232, Second Language Acquisition and Testing in Europe Monograph Series 1, 2010.

MILTON, James. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In: Bardel, C. **L2 VOCABULARY ACQUISITION, KNOWLEDGE AND USE new perspectives on assessment and corpus analysis**, p. 57 - 78, 2013.

NAGASAWA, E. Y. **Português como Língua Adicional para fins específicos: preparação ao exame Celpe-Bras.** Dissertação de Mestrado em Letras. Universidade Federal do Rio Grande do Sul, 2018.

NAGASAWA, E. Y. **Número de examinandos do exame Celpe-Bras**. Acervo Celpe-Bras, 2021. Disponível em: http://www.ufrgs.br/acervocelpebras/dados-celpe-bras/numero-de-examinandos-homologados/view

NAGASAWA, Ellen Yurika; DIVINO, Luiza Sarmento; SCHOFFEN, Juliana Roquele. Relatos dos usuários sobre as contribuições do acervo Celpe-Bras para promoção da língua portuguesa. **Revista de Letras Juçara**, v. 3, p. 239-257, 2019.

OLIVEIRA, Letícia Machado. **O Nível Avançado Superior na Avaliação Integrada de Leitura e Escrita do Exame Celpe-Bras**. XXXII Salão de Iniciação Científica. Youtube, 1 set. de 2020. Disponível em: https://www.youtube.com/watch?v=WvHtM7vLuBw

PAQUOT, M. The phraseological dimension in interlanguage complexity research. **Second Language Research**, 35(1), p. 121–145, 2019.

QUEIROZ, V. S. **A competência discursiva em textos de participantes do Celpe-Bras: uma abordagem modular**. Dissertação de Mestrado em Linguística. Universidade Federal de Minas Gerais, UFMG, 2017.

RAYSON, P. **Matrix:** A statistical method and software tool for linguistic analysis through corpus comparison. Tese de doutorado. Universidade de Lancaster, 2002.

SARMENTO, S. **O uso dos verbos modais em manuais de aviação em inglês:** Um estudo baseado em corpus. Tese de doutorado. UFRGS: Porto Alegre, 2008.

SCARAMUCCI, Matilde Virginia Ricardi. Proficiência em LE: considerações terminológicas e conceituais. **Trabalhos em Linguística Aplicada**, Campinas, v. 36, n. 1, p. 11-22, 2000.

SCARAMUCCI, Matilde Virginia Ricardi. Validade e consequências sociais das avaliações em contextos de ensino de línguas. ***LINGVARVM ARENA***, v. 2, p. 103-120, 2011.

SCHLATTER, Margarete. **Celpe-Bras:** avaliação, ensino e formação de professores de português como língua adicional. 2014. Disponível em: http://www.ufrgs.br/acervocelpebras/um-pouco-de-historia. Acesso em: 02 de outubro de 2021.

SCHLATTER, Margarete; GARCEZ, Pedro M. Línguas adicionais (Espanhol e Inglês). In: RIO GRANDE DO SUL, Secretaria de Estado da Educação, Departamento Pedagógico. **Referenciais curriculares do Estado do Rio Grande do Sul:** linguagens, códigos e suas tecnologias. Porto Alegre: SE/DP, p. 127-172, 2009. Disponível em: https://servicos.educacao.rs.gov.br/dados/refer_curric_vol1.pdf

SCHLATTER, M.; SCARAMUCCI, M. V. R., PRATI, S., ACUÑA, L. Celpe-Bras e Celu: impactos da construção de parâmetros comuns de avaliação de proficiência em portugês e em espanhol. In: FONTANA, Mónica Zoppi (org.). **O português do Brasil como língua transnacional**. Campinas: RG Editora, 2009.

SCHLATTER, Margarete; ALMEIDA, Alexandre N.; FORTES, Melissa S; SCHOFFEN, Juliana R. Avaliação de desempenho e os conceitos de validade, confiabilidade e efeito

retroativo. In: NASCIMENTO, Valdir F.; NAUJORKS, Jane C.; REBELLO, Lúcia S.; SILVA, Deborah S. (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.

SCHOFFEN, J. R. **Avaliação de proficiência oral em língua estrangeira: descrição dos níveis de candidatos falantes de espanhol no exame Celpe-Bras.** Dissertação de Mestrado, PPG-Letras – UFRGS, 2003.

SCHOFFEN, Juliana Roquele. **Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras**. Tese (Doutorado em Linguística Aplicada) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SCHOFFEN, Juliana Roquele. O conceito de proficiência e o processo de avaliação da Parte Escrita do exame Celpe- Bras. **Inventário**, Salvador, 2021.

SCHOFFEN, Juliana Roquele; NAGASAWA, Ellen Yurika; SIRIANNI, Gabrielle Rodrigues; MACHADO, Bárbara Petry. Resgatando a história do exame Celpe-Bras: desenvolvimento, disponibilização e estudos sobre o Acervo de provas e documentos públicos do exame Celpe-Bras. **Cadernos do IL**, Porto Alegre, v. 55, p. 87-113, 2017.

SHEN, Wen. **Padrões de autocorreção e de reformulação de produções escritas por aprendentes de PLE**. Dissertação de Mestrado em Português como Língua Estrangeira e Língua Segunda. Universidade de Coimbra, Coimbra, 2017.

SILVEIRA, Julia Luiz da. **ANÁLISE DAS TAREFAS DE LEITURA E ESCRITA NO NÍVEL AVANÇADO SUPERIOR DO EXAME CELPE-BRAS.** XXXIII Salão de Iniciação Científica, UFRGS. Youtube, 29 ago. de 2021. Disponível em: https://www.youtube.com/watch?v=-YLlpvfyMh8

SIDI, W. A. **Níveis de proficiência em leitura e escrita de falantes de espanhol no exame CELPE-Bras**. Dissertação de Mestrado – Pós-Graduação em Letras, UFRGS, 2002.

SINCLAIR, J. **Corpus, Concordance, Collocation**. Oxford: OUP, 1991.

Simple maths. **Sketch Engine**, s.d. Disponível em: https://www.sketchengine.eu/documentation/simple-maths/. Acesso em: 21 de novembro de 2021.

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.** Trabalho de Conclusão de Curso em Licenciatura em Letras. Universidade Federal do Rio Grande do Sul, 2016.

SIRIANNI, Gabrielle Rodrigues. **Entre a certificação e a não certificação no Celpe-Bras: um estudo sobre os níveis de proficiência na Parte Escrita do exame.** Dissertação de Mestrado em Letras. Universidade Federal do Rio Grande do Sul, UFRGS, 2020.

SIRIANNI, Gabrielle Rodrigues; MENDEL, Kaiane; NAGASAWA, Ellen Yurika; SCHOFFEN, Juliana Roquele. Os desdobramentos do Acervo Celpe-Bras para ensino, aprendizagem, avaliação e pesquisa em Português como Língua Adicional. **BELT**, Porto Alegre, v. 10, n. 1, p. 1-19, jan.-jun. 2019.

SOUZA, Juliana Chaves. Aquisição dos tempos verbais do português no ensino de PLE. In: GIL, Beatriz Daruj; AMADO; Rosane de Sá (Orgs.). **Reflexões sobre o ensino de português para falantes de outras línguas**, São Paulo, p. 80-86, 2012.

SHOHAMY, E. **Language Policy:** hidden agendas and new approaches. London, UK: Routledge, 2006.

SOUZA NETO, Maurício José de. **CELPE-BRAS E CAPLE: A PROFICIÊNCIA EM PORTUGUÊS COMO LÍNGUA NÃO MATERNA EM PARALAXE**. Dissertação de Mestrado em Letras. Universidade Federal da Bahia, Salvador, 2019.

STICHINI, Catarina. **Aquisição dos Clíticos no Ensino Simultâneo de PE e PB a Alunos Universitários na Suécia**. Dissertação de Mestrado em Letras. Universidade do Porto, Porto, 2014.

VIANA, Vander. **Verbos modais em contraste: uma análise de corpus da escrita de universitários em inglês**. Dissertação de Mestrado em Letras. Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

VIANA, Vander. LINGUÍSTICA DE CORPUS: CONCEITOS, TÉCNICAS & ANÁLISES. In: TAGNIN, S.; VIANA, V. (Orgs.). **Corpora no Ensino de Línguas Estrangeiras**. São Paulo: Hub. p. 25-96. 2011.

ANEXOS

Anexo 1: Parâmetros de avaliação da Parte Escrita

Parâmetros de Avaliação da parte escrita
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: BRASIL, 2020.

Anexo 2: Os descritores gerais dos níveis de proficiência

Avançado Superior

O examinando que atinge o nível Avançado Superior é capaz de produzir textos escritos claros e coesos de diferentes gêneros discursivos sobre assuntos variados, configurando a interlocução de forma adequada e consistente, utilizando recursos lexicais e gramaticais apropriados aos gêneros produzidos. É capaz de recontextualizar, com propriedade, informações relevantes obtidas a partir da interpretação de textos de diferentes gêneros orais e escritos, demonstrando compreensão eficiente e seletiva. Eventuais inadequações pontuais não comprometem o bom cumprimento dos propósitos dos textos produzidos.

É capaz de interagir oralmente com muita autonomia e desenvoltura, utilizando vocabulário amplo e adequado e variedade também ampla de estruturas para expressar ideias e opiniões sobre assuntos variados, contribuindo muito para o desenvolvimento da interação. Apresenta fluência, sem interrupções do fluxo natural da conversa, e pronúncia adequada. Demonstra compreensão do fluxo natural da fala do interlocutor, com rara necessidade de repetição e/ou reestruturação.

Avançado

O examinando que atinge o nível Avançado é capaz de produzir textos escritos claros e coesos de diferentes gêneros discursivos sobre assuntos variados, configurando a interlocução de forma adequada, utilizando recursos lexicais e gramaticais apropriados aos gêneros produzidos. É capaz de recontextualizar adequadamente informações relevantes obtidas a partir da interpretação de textos de diferentes gêneros orais e escritos, demonstrando compreensão eficiente. Inadequações pontuais podem fragilizar partes do texto, ainda que não comprometam o cumprimento dos propósitos dos textos produzidos.

É capaz de interagir oralmente com autonomia e desenvoltura para a expressão de ideias e opiniões sobre assuntos variados, contribuindo para o desenvolvimento da interação. Demonstra fluência, com poucas interrupções do fluxo natural da conversa. Seu vocabulário é amplo e adequado, com poucas interferências de outras línguas. Utiliza uma variedade ampla e adequada de estruturas, com poucas inadequações no uso de estruturas complexas e raras inadequações no uso de estruturas básicas. Sua pronúncia pode apresentar algumas inadequações e/ou interferências de outras línguas. Demonstra compreensão do fluxo natural da fala do interlocutor, com alguma necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala.

Intermediário Superior

O examinando que atinge o nível Intermediário Superior é capaz de produzir textos escritos de diferentes gêneros discursivos sobre assuntos variados, podendo configurar a interlocução de forma nem sempre adequada e mobilizando recursos lexicais e gramaticais nem sempre apropriados aos gêneros produzidos, podendo apresentar problemas de clareza, coesão e/ou inadequações que podem comprometer a fluidez da leitura. É capaz de recontextualizar, ainda que com equívocos, informações a partir da interpretação de textos de diferentes gêneros orais e escritos, podendo demonstrar problemas de compreensão. Inadequações podem dificultar o cumprimento dos propósitos dos textos produzidos.

Intermediário Superior

É capaz de interagir oralmente para a expressão de ideias e opiniões sobre assuntos variados. Demonstra fluência, com algumas pausas e hesitações que às vezes interrompem o fluxo da conversa. Seu vocabulário é adequado, embora apresente algumas interferências de outras línguas. Apresenta algumas inadequações no uso de estruturas complexas e poucas no uso de estruturas básicas. Sua pronúncia contém inadequações e/ou interferências de outras línguas. Demonstra alguns problemas de compreensão do fluxo natural da fala do interlocutor, com necessidade de repetição e/ou reestruturação ocasionada por palavras de uso frequente, em ritmo normal da fala.

Intermediário

O examinando que atinge o nível Intermediário é capaz de produzir textos escritos sobre assuntos variados que, com dificuldade, podem ser reconhecidos como pertencentes a determinados gêneros discursivos, podendo não configurar adequadamente a interlocução. Os recursos lexicais e gramaticais mobilizados são limitados, podendo apresentar problemas de clareza e coesão e/ou inadequações frequentes que comprometem mais frequentemente a fluidez da leitura. É capaz de selecionar algumas informações a partir da interpretação de textos de diferentes gêneros orais e escritos, evidenciando problemas de compreensão e dificuldades no trabalho de recontextualização que podem levar ao cumprimento parcial dos propósitos dos textos produzidos. É capaz de interagir oralmente para a expressão de ideias e opiniões sobre assuntos variados. Apresenta poucas hesitações, com algumas interrupções no fluxo da conversa. Seu vocabulário pode apresentar limitações que podem comprometer o desenvolvimento da interação. Utiliza variedade limitada de estruturas, com algumas inadequações em estruturas complexas e poucas inadequações em estruturas básicas. Sua pronúncia contém inadequações e/ou interferências frequentes de outras línguas. Demonstra alguns problemas de compreensão do fluxo da fala, com necessidade frequente de repetição e/ou reestruturação ocasionada por palavras de uso frequente em nível normal de fala.

Fonte: BRASIL, 2020

Anexo 3: Tarefa IV de 2015/2, “Azulejos valiosos”

2015/2 CelpeBras

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.

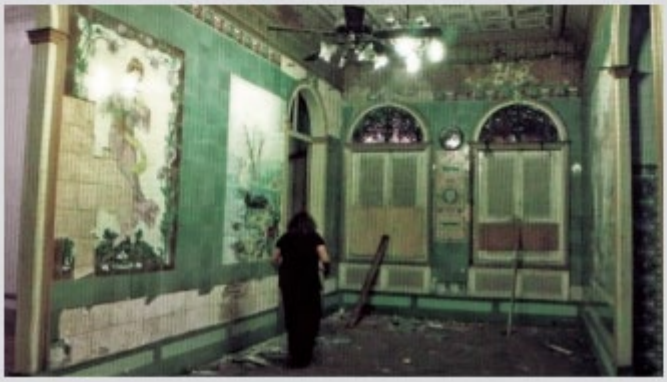
Azulejos valiosos

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 parense, 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 Toscado, 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 superindentende 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 **Eu Dia**, jun. 2012 (adaptado).

Anexo 4: Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_ Celpe-Bras 2015-2 T4 Nota 5

	Frecuency		Document frequency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
ciudade	46	0	26	0	4.14013	0
seguridade	23	0	19	0	3.02548	0
debe	19	0	15	0	2.38854	0
montando	16	0	13	0	2.07006	0
casaroes	16	0	12	0	1.91083	0
superindentende	16	0	16	0	2.54777	0
vitimas	15	0	13	0	2.07006	0
actos	15	0	13	0	2.07006	0
pido	15	0	15	0	2.38854	0
augusto	14	0	14	0	2.2293	0
solicitamos	14	0	12	0	1.91083	0
bonito	13	0	12	0	1.91083	0
importadas	13	0	13	0	2.07006	0
montenegro	12	0	12	0	1.91083	0
toscado	12	0	12	0	1.91083	0
la	11	0	11	0	1.75159	0
gran	11	0	10	0	1.59236	0
thais	11	0	10	0	1.59236	0
parana	11	0	8	0	1.27389	0
batizado	11	0	10	0	1.59236	0
thaís	11	0	3	0	0.47771	0
inicio	10	0	10	0	1.59236	0
podam	10	0	9	0	1.43312	0
voces	10	0	10	0	1.59236	0
gram	10	0	9	0	1.43312	0
control	10	0	9	0	1.43312	0
octubro	10	0	10	0	1.59236	0
proceso	10	0	9	0	1.43312	0
proteçao	10	0	10	0	1.59236	0
puntuais	10	0	8	0	1.27389	0
areas	9	0	9	0	1.43312	0
coordenadora	9	0	8	0	1.27389	0
situaçao	9	0	8	0	1.27389	0
ref	9	0	9	0	1.43312	0
pronto	9	0	8	0	1.27389	0
pretendemos	9	0	9	0	1.43312	0
belen	9	0	7	0	1.11465	0
inspetor	9	0	9	0	1.43312	0
achamos	9	0	8	0	1.27389	0
bonitas	9	0	9	0	1.43312	0
debem	8	0	8	0	1.27389	0
diferente	8	0	6	0	0.95541	0

importado	8	0	8	0	1.27389	0
saibam	8	0	8	0	1.27389	0
alguem	8	0	8	0	1.27389	0
população	8	0	6	0	0.95541	0
historicas	8	0	8	0	1.27389	0
presidente	7	0	3	0	0.47771	0
corresponda	7	0	7	0	1.11465	0
imagem	7	0	7	0	1.11465	0
titulo	7	0	7	0	1.11465	0

Anexo 5: Palavras-chave Celpe-Bras 2015-2 T4 Nota 5_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	828	1388	230	18	97.04641	62.06897
casarões	490	814	202	12	85.23207	41.37931
vandalismo	192	1231	156	16	65.82278	55.17241
palacete	120	517	102	15	43.03797	51.72414
atenciosamente	117	772	117	14	49.36709	48.27586
roubos	220	5053	145	21	61.18143	72.41379
belem	41	138	22	9	9.2827	31.03448
dphac	29	0	27	0	11.39241	0
coloriam	28	23	28	7	11.81435	24.13793
prezados	42	611	42	13	17.72152	44.82759
belém	570	22826	215	23	90.7173	79.31034
depredações	39	549	35	12	14.76793	41.37931
patrimonio	35	385	26	8	10.97046	27.58621
fachadas	70	2001	67	20	28.27004	68.96552
valiosos	75	2288	62	20	26.16034	68.96552
roubados	78	2767	60	20	25.31646	68.96552
prezado	44	1160	43	16	18.14346	55.17241
palacetes	26	231	19	12	8.01688	41.37931
furtados	32	547	26	11	10.97046	37.93103
vitor	105	4380	99	20	41.77215	68.96552
lacore	15	0	14	0	5.90717	0
depredação	27	1074	23	16	9.70464	55.17241
inconformidade	19	495	19	10	8.01688	34.48276
tombamento	48	3130	46	16	19.40928	55.17241
depredados	16	340	16	9	6.75105	31.03448
património	19	655	12	6	5.06329	20.68966
morador	102	8670	82	21	34.59916	72.41379
imediatas	70	5713	66	19	27.8481	65.51724
furtos	38	3116	29	16	12.23629	55.17241
depredadas	11	105	11	9	4.64135	31.03448
moradora	37	3087	34	17	14.34599	58.62069
ufpa	40	3580	39	11	16.4557	37.93103
escrevo	31	2573	30	19	12.65823	65.51724
cacos	17	993	17	20	7.173	68.96552
cordialmente	12	389	12	13	5.06329	44.82759
inconformado	15	811	15	15	6.32911	51.72414
azulejo	12	426	10	15	4.21941	51.72414
historicos	9	57	6	5	2.53165	17.24138
patrimônios	18	1289	13	18	5.48523	62.06897
encarregadas	17	1168	17	13	7.173	44.82759
desqualificação	22	1865	22	11	9.2827	37.93103
necessarias	9	113	8	6	3.37553	20.68966

restaurados	17	1360	17	18	7.173	62.06897
paraense	41	4881	38	15	16.03376	51.72414
tombadas	9	227	8	11	3.37553	37.93103
necessario	10	408	10	9	4.21941	31.03448
casarão	20	1981	18	17	7.59494	58.62069
prezada	9	277	9	11	3.79747	37.93103
roubando	15	1287	14	21	5.90717	72.41379
perfeitura	7	2	3	2	1.26582	6.89655

Anexo 6: Palavras-chave Celpe-Bras 2015-2 T4 Nota 2_Corpus Brasileiro

	Frequency		Document fecuency		Relative DOCF	
Item	**Focus**	**Reference**	**Focus**	**Reference**	**Focus**	**Reference**
azulejos	2169	1388	607	18	96.65605	62.06897
casarões	1012	814	482	12	76.75159	41.37931
palacete	321	517	236	15	37.57962	51.72414
vandalismo	446	1231	365	16	58.12102	55.17241
belem	225	138	140	9	22.29299	31.03448
atenciosamente	210	772	210	14	33.43949	48.27586
patrimonio	141	385	109	8	17.35669	27.58621
valiosos	313	2288	226	20	35.98726	68.96552
roubos	554	5053	360	21	57.32484	72.41379
coloriam	103	23	101	7	16.0828	24.13793
furtados	110	547	102	11	16.24204	37.93103
historicos	70	57	62	5	9.87261	17.24138
fachadas	169	2001	159	20	25.31847	68.96552
depredadas	62	105	61	9	9.71338	31.03448
historico	73	358	57	8	9.07643	27.58621
depredações	82	549	77	12	12.26115	41.37931
dphac	54	0	53	0	8.43949	0
prezados	73	611	73	13	11.6242	44.82759
belém	957	22826	455	23	72.45223	79.31034
vitor	220	4380	197	20	31.36943	68.96552
ciudade	46	23	26	6	4.14013	20.68966
lacore	43	0	39	0	6.21019	0
roubados	138	2767	125	20	19.90446	68.96552
prezado	81	1160	80	16	12.73885	55.17241
decada	45	150	45	8	7.16561	27.58621
agravado	111	2070	111	16	17.67516	55.17241
estam	40	56	35	6	5.57325	20.68966
cacos	70	993	69	20	10.98726	68.96552
solucioná-lo	44	205	44	12	7.00637	41.37931
moradora	124	3087	96	17	15.28662	58.62069
seculo	39	199	37	8	5.89172	27.58621
paraense	173	4881	152	15	24.20382	51.72414
tombadas	38	227	37	11	5.89172	37.93103
laboratorio	44	469	42	9	6.6879	31.03448
casarão	85	1981	82	17	13.05732	58.62069
morador	267	8670	219	21	34.87261	72.41379
azulejo	41	426	39	15	6.21019	51.72414
imediatas	179	5713	159	19	25.31847	65.51724
encarregadas	60	1168	59	13	9.3949	44.82759
patrimonios	29	10	23	4	3.66242	13.7931
necessario	39	408	35	9	5.57325	31.03448
tombamento	105	3130	103	16	16.40127	55.17241

suspeite	29	77	29	11	4.61783	37.93103
inconformado	46	811	43	15	6.84713	51.72414
tomarem	74	2012	66	20	10.50955	68.96552
palacetes	29	231	27	12	4.29936	41.37931
bonitos	59	1670	57	20	9.07643	68.96552
artistico	25	69	24	6	3.82166	20.68966
inconformidade	33	495	33	10	5.25478	34.48276
ufpa	94	3580	93	11	14.80892	37.93103