revista HORIZONTES DE LINGUÍSTICA APLICADA

e-ISSN 2237-0951

ARTIGO ORIGINAL

Padrões de uso de recursos lexicais em diferentes níveis de proficiência: recuperação de informações do material de insumo em uma tarefa de leitura e escrita do Celpe-Bras

Patterns of lexical resources use at different proficiency levels: retrieval of input material information in a reading and writing Celpe-Bras task

Luiza Sarmento Divino[1], Juliana Roquele Schoffen[2]

1 Universidade Federal do Rio Grande do Sul - luiza.sarmento.divino@gmail.com
2 Universidade Federal do Rio Grande do Sul - julianaschoffen@gmail.com

Como citar o artigo.
DIVINO, L. S.; SCHOFFEN, J. R. Padrões de uso de recursos lexicais em diferentes níveis de proficiência: recuperação de informações do material de insumo em uma tarefa de leitura e escrita do Celpe-Bras. *Revista Horizontes de Linguística Aplicada*, ano 24, n. 2, AG3, 2025.

Resumo

Este estudo visa investigar a relação entre o material de insumo de uma tarefa de leitura e escrita da edição 2015/2 do Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) e os recursos lexicais utilizados pelos examinandos avaliados em diferentes níveis de proficiência. O *corpus* deste trabalho é composto por 2.268 textos divididos em cinco sub*corpora* de acordo com a sua nota final, podendo ser 1, 2, 3, 4 ou 5. O *software* utilizado para as análises foi o *Sketch Engine*, um conjunto de ferramentas *on-line* com uma gama de funções altamente flexíveis para manusear *corpora* (Kunilovskaya; Koviazina, 2017). A partir da ferramenta *N-Grams*, foram extraídas sequências de seis itens lexicais que se repetiram nos textos de cada sub*corpus*, para, então, serem comparadas com o material de insumo para averiguar a existência de cópia. Os resultados apontam para um aumento no número de *n-grams* conforme a nota diminui, indicando que, apesar de haver cópia em todos os níveis, quanto menor a proficiência, maior a necessidade de cópia exata das informações presentes no material de insumo sem que estas informações sejam adequadamente articuladas ao novo contexto de produção, corroborando os resultados de Sirianni (2016) e Mendel (2019).

Palavras-chave: Exame Celpe-Bras. Linguística de *Corpus*. Avaliação de Proficiência. Avaliação de habilidades integradas. Avaliação de leitura e escrita.

Abstract

This study aims to investigate the relationship between the input material of a reading and writing task from the 2015/2 edition of the *Certificado de Proficiência em Língua Portuguesa para Estrangeiros* (Celpe-Bras) and the lexical resources used by test-takers rated at different proficiency levels. The *corpus*

Fonte de financiamento: Nenhuma.
Conflito de interesse: As autoras declaram não haver.
Data de recebido: 15 Dez 2024. Data de aprovado: 06 Jun 2025.

consists of 2,268 texts divided into five subcorpora according to their final scores, ranging from 1 to 5. The software used for the analysis was Sketch Engine, an online suite of tools with highly flexible functions for *corpus* handling (Kunilovskaya; Koviazina, 2017). Using the *N-Grams* tool, sequences of six lexical items recurring in the texts of each sub*corpus* were extracted and compared with the input material to verify the presence of copying. The results indicate an increase in the number of *n-grams* as scores decrease, suggesting that, although copying occurs at all proficiency levels, lower proficiency correlates with a greater reliance on exact copying of information from the input material without adequate integration into the new production context, corroborating the findings of Sirianni (2016) and Mendel (2019).

Keywords: Celpe-Bras Exam. *Corpus* Linguistics. Proficiency Assessment. Integrated Skills Assessment. Reading and Writing Assessment.

1 INTRODUÇÃO

Neste artigo, propomo-nos a analisar a relação entre o material de insumo de uma tarefa de leitura e escrita da edição 2015/2 do Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) e o léxico utilizado pelos examinandos avaliados em diferentes níveis de proficiência à luz da Linguística de *Corpus* (LC). A LC se dedica à coleta e análise de *corpora*, conjuntos de textos eletrônicos organizados com critérios específicos (Sinclair, 1999). Seu propósito é investigar, empiricamente, aspectos da língua por meio de ferramentas de processamento computacional (Sardinha, 2000). Para Biber (2012), a LC representa uma abordagem metodológica que facilita o estudo da variação e do uso linguístico. Fundamentalmente, a LC é considerada uma disciplina que consiste em dados de frequência sobre como algum fenômeno linguístico é distribuído em um *corpus* (Gries, 2010) e como diferentes fenômenos linguísticos são combinados uns com os outros (Cushing, 2022).

Desde a virada do milênio, percebe-se o aumento do interesse no uso de LC para informar o desenvolvimento e a validação de exames de língua (Cushing, 2017; 2022). No Celpe-Bras, no entanto, a não existência de um *corpus* de textos produzidos e avaliados no exame e a dificuldade de conseguir dados para pesquisa (Vicentini, 2022; Tosatti, 2021) limitaram a utilização dessa metodologia. Apenas com a recente compilação do CorCel (Schoffen *et al.*, no prelo)[1] foi possível avançar com os trabalhos de descrição dos níveis de proficiência avaliados no exame utilizando LC. Até então, os estudos sobre o Celpe-Bras que fazem uso de ferramentas de análise da LC objetivaram detectar padrões lexicais (Evers, 2013; Divino 2021; 2024; Hanauer, 2023; Schoffen; Divino, 2023; Raupp, 2024) e coesivos (Evers, 2013; Kunrath, 2019; Sostruznik, 2023), bem como informações referentes à extensão dos textos e riqueza lexical (Evers, 2013; Divino, 2021; 2024; Hanauer, 2023) que pudessem melhor descrever os níveis de proficiência. Além disso, também olharam para a recuperação de informações do texto de insumo (Kunrath, 2019; Divino, 2021; Hanauer, 2023). Os *softwares* utilizados nesses estudos foram o Coh-Metrix-Port (Evers, 2013; Kunrath, 2019) e o *Sketch Engine (SkE)* (Divino, 2021; 2024; Hanauer, 2023; Schoffen; Divino, 2023; Sostruznik, 2023; Raupp, 2024).

O uso de tarefas integradas, que combinam as habilidades de compreensão e produção, são consideradas um avanço na área de avaliação de proficiência linguística, direcionando a definição de um construto mais válido para o letramento, e têm se consolidado como uma prática relevante em exames de larga escala ao longo das últimas décadas (Mendel, 2019). No entanto, uma questão importante que surge nesse contexto é até que ponto a recuperação de informações do texto de insumo, incluindo paráfrases e cópias, pode ser considerada adequada sem comprometer a avaliação da proficiência do examinando. A seguir, refletiremos sobre essa questão, analisando como o léxico utilizado pelos candidatos em

[1] Os textos que compõem o CorCel fazem parte dos dados disponibilizados pelo Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) ao grupo de pesquisa AVALIA - Avaliação de Uso da Linguagem (UFRGS) em 2017 para contemplar os objetivos de pesquisa do grupo, entre eles, a compilação de um *corpus* de textos avaliados. Os textos foram disponibilizados em forma de cópia digitalizada, precisando ser digitados, revisados e anonimizados com base em um tutorial (Schoffen *et al*., 2024) para assegurar a padronização durante o processo.

diferentes níveis de proficiência dialoga com o material de insumo e em que medida essa interação pode ser considerada um indicador da competência escrita.

2 O EXAME CELPE-BRAS

O Celpe-Bras tem como objetivo avaliar a proficiência em língua portuguesa por meio do uso da língua, com a realização de tarefas autênticas que compartilham características com as tarefas possíveis de serem realizadas em situações reais de uso (Douglas, 2000). No Celpe-Bras, a proficiência é testada de maneira direta (Brasil, 2020) e o exame é composto por duas partes distintas: uma Parte Oral (PO) e uma Parte Escrita (PE).

O Celpe-Bras não se baseia na ideia de uma proficiência "única, absoluta [e] monolítica" (Scaramucci, 2000, p. 14). No exame, os falantes são considerados mais ou menos proficientes, a depender do seu desempenho para se adequar a determinado contexto de comunicação (Scaramucci, 2000). A noção de proficiência adotada no Celpe-Bras pressupõe o uso da língua em uma relação dialógica (Bakhtin, 2003) entre falantes, pensada como "uma relação (de sentido) que se estabelece entre enunciados na comunicação verbal" por meio de práticas sociais. Cada enunciado é, segundo Bakhtin (2003), individual e único, porém, as diferentes esferas de utilização da língua apresentam tipos relativamente estáveis de enunciados, denominados por Bakhtin como "gêneros do discurso", através dos quais materializam-se as práticas de linguagem (Dell'Isola; Pordeus, 2021). É por meio dos gêneros que as tarefas da PE do Celpe-Bras buscam avaliar proficiência.

A PE do Celpe-Bras é composta de quatro tarefas que integram compreensão e produção, como podemos ver no Quadro 1.

Quadro 1. Estrutura da Parte Escrita

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

Fonte: Brasil (2020, p. 35).

Cada tarefa apresenta um enunciado que indica "um propósito claro de comunicação (escrever um texto para reclamar, informar, discordar etc.), um enunciador (morador de um determinado bairro, gerente de uma empresa, internauta etc.) e um ou mais interlocutores (leitores de um jornal, o chefe, o prefeito da cidade etc.)" (Brasil, 2020, p. 36). As tarefas do Celpe-Bras são elaboradas com base em um material de insumo, prevendo uma avaliação que integra as habilidades de leitura ou compreensão oral e produção escrita (Mendel; Schoffen, 2017), em que os propósitos de leitura são determinados pelos propósitos de escrita (Scaramucci, 2016). Para que a tarefa seja adequadamente cumprida, o examinando precisa realizar a articulação das informações do texto de insumo com as demais exigências do enunciado (Pileggi, 2017), pois a "compreensão envolve não apenas o conteúdo temático do texto, mas também a situação de produção em que o texto a ser [recontextualizado] está inserido" (Mendel, 2019, p. 52). Em outros termos, pode-se afirmar que a integração de habilidades se dá, no exame, por meio da recontextualização de informações do material disponibilizado, apoiada nas expectativas de compreensão e produção determinadas no enunciado da tarefa (Mendel, 2019).

3 "AZULEJOS VALIOSOS": A TAREFA ANALISADA

A tarefa analisada neste artigo, intitulada "Azulejos valiosos", foi a Tarefa 4 da edição de 2015/2 do Celpe-Bras. Entre as informações contidas no enunciado, são determinados o locutor, um *morador de Belém*, e o interlocutor projetado, a *prefeitura municipal*. Como complemento da descrição do papel do examinando nesse contexto comunicativo, há a informação de que ele está *inconformado com a situação dos casarões históricos da cidade*, o que também complementa o propósito de *explicar o problema e argumentar sobre a necessidade de se tomarem medidas imediatas para solucioná-lo*. O gênero de produção também é explicitado no enunciado, sinalizando que a interação está circunscrita em uma *carta aberta* que será *publicada em jornais locais*.

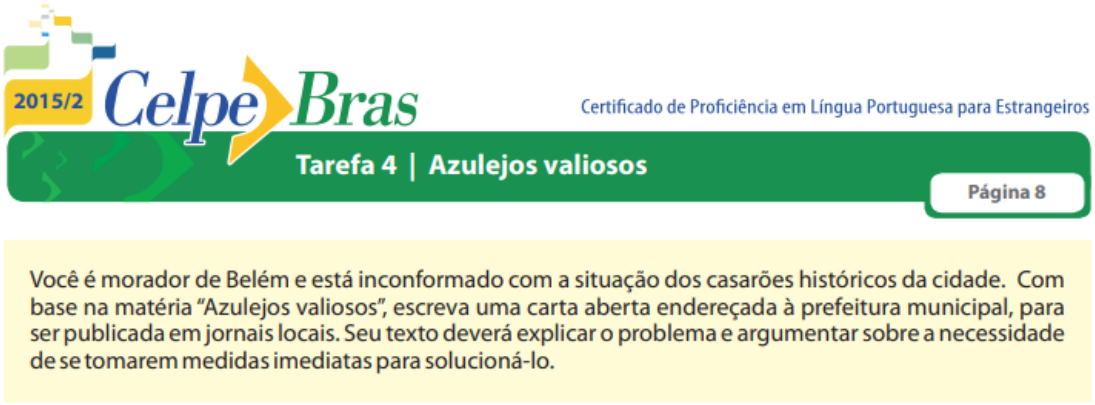

Figura 1. Enunciado da tarefa "Azulejos valiosos"
Fonte: Brasil (2015).

As condições de produção dessa tarefa indicam que ela deve ser feita com base na matéria "Azulejos valiosos", como pode ser observado na Figura 1. O texto em questão, publicado no jornal *Em Dia*, relata a situação de casarões históricos na cidade de Belém, que vêm sendo depredados e têm seus azulejos sendo alvo de roubo. No texto, são abordados, ainda, a possível existência de um mercado ilegal de azulejos e o processo de tombamento de um dos casarões.

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 Em Dia, jan. 2012 (adaptado).

Figura 2. Design da tarefa "Azulejos valiosos"
Fonte: Brasil (2015).

Estudos previamente realizados sobre os textos escritos em resposta a essa tarefa utilizando metodologia qualitativa chegaram a algumas conclusões sobre a utilização do material de insumo nos diferentes níveis de proficiência. Mendel (2019) afirma ser possível encontrar cópia de trechos do material de insumo nos textos de notas 3, 2 e 1. Em seu estudo, a autora encontrou trechos mais longos copiados nos textos de nota 1 do que nas demais notas. Os textos de notas mais baixas são, segundo a autora, mais dependentes do texto de insumo, não somente das informações em si, mas também da organização dessas informações e dos recursos linguísticos utilizados, ocasionando que, muitas vezes, os textos produzidos se assemelham bastante ao texto de insumo. Quanto às notas 4 e 5, Mendel (2019) afirma não haver cópia de trechos inteiros, de modo que as informações do texto de insumo são articuladas nos textos de ambas as notas, de maneira autoral nos textos de nota 5. Os resultados de Sirianni (2016) também apontam o impacto da cópia no cumprimento do propósito e na configuração do gênero solicitado, uma vez que, na nota 1, o texto produzido se assemelha muito ao texto de insumo, não havendo adaptação ao novo contexto de produção. Além disso, Sirianni (2016) comenta sobre o conteúdo dos trechos copiados, concluindo que textos de notas mais baixas tendem a copiar mais trechos que não contribuem diretamente para o cumprimento adequado da tarefa.

4 DESCRIÇÃO DO *CORPUS*

O *corpus* utilizado nesta pesquisa é composto por textos produzidos por examinandos em resposta à Tarefa 4 da edição de 2015/2 do exame Celpe-Bras, e faz parte do CorCel (Schoffen *et al.*, no prelo); por esse motivo foi atribuído o nome de CorCel20152t4. Cada texto produzido na PE do Celpe-Bras é avaliado com notas entre 0 e 5 de forma independente por dois avaliadores[2]. O critério para a seleção dos textos foi a condição de terem, como nota final, um valor inteiro (5, 4, 3, 2, 1 e 0). No total, o *corpus* deste trabalho é composto por 2.268 textos, contendo a totalidade de textos com notas finais 1, 2, 3, 4 e 5 recebidas pelos textos escritos na Tarefa 4 da edição 2015-2 do exame. Pela baixa quantidade de textos que compõem o sub*corpus* n0 (25), optamos por não utilizá-lo neste estudo. Os textos estão distribuídos conforme ilustrado na Tabela 1, contendo o número total de textos no sub*corpus* de cada nota (N_textos), o número total de palavras no sub*corpus* de cada nota (*Tokens*) (Biber, 2012) e o número de diferentes formas de palavras (*Types*) (Biber, 2012). Para fins de comparação, o número total de textos em cada nota foi normalizado para 100.

Tabela 1. Número de textos, número de *types*, número de *tokens* e dos sub*corpora* do CorCel20152t4.

Corpus	N_textos	*N_textos_ norm*	Tokens	*Tokens_ norm*	Types	*Types_ norm*
n5	237	*100*	47.616	*20.091,14*	5.663	*2.389,45*
n4	477	*100*	88.235	*18.497,90*	7.247	*1.519,28*
n3	715	*100*	123.033	*17.207,41*	8.838	*1.236,08*
n2	628	*100*	99.457	*15.837,10*	8.537	*1.359,39*
n1	211	*100*	30.290	*14.355,45*	4.044	*1.916,59*
Total	**2.268**					

Fonte: Adaptado de Divino (2024).

Vê-se, pela quantidade de tokens normalizados, que os textos do corpus n5 são, comparativamente, mais extensos que os textos do n4, que são mais extensos que os do n3, que são mais extensos que os do n2, que são mais extensos que os do n1. Os *types*, por outro lado, apresentam um comportamento um pouco diferente. Do corpus n5 ao n3, o número diminui. Cada palavra escrita de forma diferente, por mais semelhante que seja, é codificada como um *type* diferente no SkE, de modo que, quanto mais variedade há na grafia de uma mesma palavra, mais *types* serão contabilizados. Em outras palavras, os textos de notas mais baixas apresentam mais inadequações ortográficas que os demais textos, embora também haja certa quantidade de inadequações em todas as notas, que vão gradativamente ficando mais ocasionais conforme a nota aumenta**.**

5 PROCEDIMENTOS PARA AS ANÁLISES

Neste artigo, propomo-nos a analisar a relação entre o material de insumo de uma tarefa de leitura e escrita da edição de 2015/2 do Celpe-Bras e o léxico utilizado pelos examinandos de diferentes níveis de proficiência. Para tal, buscamos responder às perguntas: *(a) Quantos n-grams de seis itens lexicais podem ser encontrados em cada sub*corpus *e em quantos textos eles aparecem?;* e *(b) Qual a relação dos n-grams encontrados em cada subcorpus com o material de insumo da tarefa?.* O *software* utilizado foi o *Sketch Engine* (SkE), um conjunto de ferramentas *on-line* com uma gama de funções altamente flexíveis para manusear *corpora* (Kunilovskaya; Koviazina, 2017), utilizadas para gerar resumos automáticos do comportamento gramatical e da colocação de palavras, encontrar palavras semelhantes em termos de sinônimos e

[2] Para maiores informações sobre o processo de avaliação da PE do Celpe-Bras, ver Segat (2022).

diferenças de usos de uma mesma palavra, bem como realizar busca mais manual de itens no *corpus* (Kilgarriff *et al*., 2004).

Para responder à pergunta *(a)*, utilizamos a ferramenta *N-Grams* do *SkE*. *N-Grams* são sequências contíguas de dois a seis ítens lexicais, podendo incluir palavras e outros sinais gráficos, como pontuação e símbolos como "#", "@", "%", contam como elementos isolados, sendo incluídos na contagem. Nessa ferramenta, foi selecionada a configuração de buscar por palavra (*word*) como atributo nas configurações, levando em conta a maneira exata como foi escrita, em vez de *lemma*, que inclui as diferentes flexões de gênero e número. Foi escolhida a opção que permite que letras maiúsculas e minúsculas não sejam diferenciadas (A=a). A frequência mínima padrão da ferramenta é de duas ocorrências, que foi mantida. Com o objetivo de identificar a recuperação de trechos longos do material de insumo, a extensão escolhida para análise foi de seis elementos lexicais. Para esta busca, foram incluídas também não palavras.

Para responder à pergunta *(b)*, os *n-grams* foram comparados com o material de insumo, para averiguar se fragmentos do texto de insumo e do enunciado foram recuperados da exata maneira que estavam. Em cada sub*corpus*, os *n-grams* foram agrupados de acordo com os trechos do material de insumo ao qual faziam referência, com exceção de dois grupos de *n-grams* que não estão presentes no material. Os resultados de cada *corpora* foram posteriormente comparados entre si.

5.1 Organização dos *n-grams*

Para organizar os resultados, os *n-grams* foram divididos em grupos de informações, pois percebeu-se que mais de um *n-gram* fazia parte de um mesmo trecho do material de insumo. Por exemplo, no corpus de nota 5, foram encontrados os *n-grams* "uma das cidades brasileiras com maior", "das cidades brasileiras com maior variedade", "cidades brasileiras com maior variedade de", "brasileiras com maior variedade de azulejos" e "considerada uma das cidades brasileiras com", que fazem parte do trecho "considerada uma das cidades brasileiras com maior variedade de azulejos".

Os *n-grams* foram adicionados a um quadro, em que estes grupos de informações foram enumerados e intitulados de acordo com o trecho completo formado pela união de todos os *n-grams* encontrados, que fazem referência a este trecho. Além dos *n-grams* em si, o quadro também apresenta a quantidade de *n-grams* em cada grupo, a extensão total do trecho somando os itens lexicais de todos os *n-grams*, a quantidade de documentos em que o *n-gram* aparece (DOCF) e a frequência relativa dessa quantidade de textos em relação ao número total de textos do subcorpus de cada nota (Relative DOCF), ou seja, o percentual de textos em que os *n-grams* aparecem. Para que um *n-gram* fosse considerado, definiu-se um ponto de corte de Relative DOCF de 5%. A ordem dos grupos foi estabelecida de acordo com a Relative DOCF do *n-gram* mais frequente de cada grupo de *n-grams*, por mais que nem todos os *n-grams* daquele grupo tenham uma Relative DOCF alta.

Os sinais de pontuação e o símbolo "%" estão separados das palavras, pois foram contabilizados como um item individual, e, como já mencionado, fazem parte dos *n-grams*. Nos casos em que o *n-gram* é muito semelhante mas não é exatamente idêntico ao material de insumo, seja pela adição de uma vírgula ou pela utilização de uma palavra diferente, optamos por adicioná-lo à contagem, destacando-o em itálico. Para ilustrar, o Quadro 2 contém o exemplo do trecho acima mencionado.

Quadro 2. *N-grams* do grupo (3) de informações do *corpus* n5

***N-Grams* encontrados no corpus n5**	**DOCF**	**Relative DOCF (%)**
(3) foi considerada uma das cidades brasileiras com maior variedade de azulejos, que coloriam as fachadas e o interior de residências (21 itens)		
13 n-grams		
uma das cidades brasileiras com maior	47	19,83
cidades brasileiras com maior variedade de	46	19,41
das cidades brasileiras com maior variedade	46	19,41
brasileiras com maior variedade de azulejos	45	18,99
considerada uma das cidades brasileiras com	36	15,19
com maior variedade de azulejos ,	23	9,7
coloriam as fachadas e o interior	20	8,44
que coloriam as fachadas e o	20	8,44
foi considerada uma das cidades brasileiras	20	8,44
com maior variedade de azulejos .	15	6,33
azulejos que coloriam as fachadas e	15	6,33
é considerada uma das cidades brasileiras	14	5,91
as fachadas e o interior das	12	5,06

Fonte: Adaptado de Divino (2024).

Quadros e tabelas como essas foram elaborados para cada subcorpus, contendo todos os grupos de informações encontrados a partir dos *n-grams* presentes em cada um. Para comparar os resultados, foi elaborado o Quadro 3 contendo a extensão dos trechos encontrados nos *n-grams* de cada corpus, bem como a quantidade de itens lexicais, que será apresentado a seguir.

6 ANÁLISE DOS DADOS

6.1 Quantos *n-grams* de seis itens lexicais podem ser encontrados em cada sub*corpus* e em quantos textos eles aparecem?

Com extensão de seis elementos lexicais e com frequência mínima de duas ocorrências, foram encontrados 1.131 *n-grams* no *corpus* n5, 2.782 no *corpus* n4, 4.314 no *corpus* n3, 3.148 no *corpus* n2 e 1.066 no *corpus* n1. O *corpus* n1 é o único a apresentar *n-grams* com esta extensão em mais de 30% dos textos, com dois *n-grams*. Com *Relative DOCF* superior a 20%, o *corpus* n5 contém dois *n-grams*, o *corpus* n4 contém cinco, o *corpus* n3 e o *corpus* n2 contêm oito cada, e o *corpus* n1 contém três. Em pelo menos 10% dos textos, o *corpus* n5 apresenta 21 *n-grams*, o *corpus* n4 apresenta 27, o *corpus* n3 e o *corpus* n2 apresentam 28 cada, e o *corpus* n1 apresenta 42. Presente em mais de 5% dos textos, o *corpus* n5 contém 59 *n-grams*, o *corpus* n4 contém 69, o *corpus* n3 contém 83, o *corpus* n2 contém 80 e o *corpus* n1 contém 129.

No Quadro 3, são apresentadas as extensões dos trechos encontrados a partir dos *n-grams* que incidem em pelo menos 5% dos textos de cada *corpus*. Na coluna da esquerda, há o número do grupo, e na coluna da direita há toda a extensão do trecho copiado, bem como a quantidade de itens contidos no trecho. O grupo (1) contém duas informações sobre o número de itens, separadas por uma barra. A primeira refere-se à quantidade de itens que podem ser reconhecidos como pertencentes à primeira parte em que a sequência "o palacete vitor maria da silva" aparece no texto de insumo, e a segunda refere-se à segunda parte. Nos grupos (2), (6) e (12), o trecho encontrado no *corpus* n1 está em amarelo, pois estes são os

grupos cujos *n-grams* se sobrepuseram e extrapolaram o ponto final. Por este motivo, o ponto que separa uma informação da outra foi adicionado em ambos os trechos.

Quadro 3. Extensão dos trechos encontrados nos *n-grams* em cada *corpus*

Grupos	Extensão dos trechos encontrados nos *n-grams* em cada corpus
(1)	**n5** o palacete vitor maria da silva (,) **(7 itens/6 itens)** **n4** (é) o palacete vitor maria da silva (,) **(8 itens/6 itens)** **n3** (é) o palacete vitor maria da silva (,) [...] um dos interiores mais bonitos da cidade **(8 itens/13 itens)** **n2** (é) o palacete vitor maria da silva (,) tem um dos interiores mais bonitos da cidade **(8 itens/14 itens)** **n1** (é) o palacete vitor maria da silva (,) tem um dos interiores mais bonitos da cidade , **(8 itens/15 itens)**
(2)	**n5** da década de 1970 para cá , mais de 50% dos azulejos se perderam . **(16 itens)** **n4** da década de 1970 para cá , mais de 50% dos azulejos se perderam . **(16 itens)** **n3** da década de 1970 para cá , mais de 50% dos azulejos se perderam . **(16 itens)** **n2** da década de 1970 para cá , mais de 50% dos azulejos se perderam . **(16 itens)** **n1** . da década de 1970 para cá , no entanto , mais de 50% dos azulejos se perderam . **(20 itens)**
(3)	**n5** foi considerada uma das cidades brasileiras com maior variedade de azulejos , que coloriam as fachadas e o interior de residências **(21 itens)** **n4** já foi considerada uma das cidades brasileiras com maior variedade de azulejos , que coloriam as fachadas e o interior de residências **(22 itens)** **n3** já foi considerada uma das cidades brasileiras com maior variedade de azulejos , que coloriam as fachadas e o interior de residências **(22 itens)** **n2** já foi considerada uma das cidades brasileiras com maior variedade de azulejos , que coloriam as fachadas e o interior de residências **(22 itens)** **n1** já foi considerada uma das cidades brasileiras com maior variedade de azulejos , que coloriam as fachadas e o interior de residências **(22 itens)**
(4)	**n5** belém, 20 de outubro de 2015 **(7 itens)** **n4** belém, 20 de outubro de 2015 **(7 itens)** **n3** . belém, 20 de outubro de 2015 **(8 itens)** **n2** belém, 20 de outubro de 2015 **(7 itens)** **n1** belém, 20 de outubro de 2015 **(7 itens)**
(5)	**n5** departamento do patrimônio histórico , artístico e cultural **(9 itens)** **n4** o departamento do patrimônio histórico , artístico e cultural [...] o processo de tombamento do casarão **(15 itens)** **n3** o departamento do patrimônio histórico , artístico e cultural [...] iniciou o processo de tombamento do casarão **(16 itens)** **n2** o departamento do patrimônio histórico , artístico e cultural [...] o processo de tombamento do casarão **(15 itens)** **n1** o departamento do patrimônio histórico , artístico e cultural [...] iniciou o processo de tombamento do casarão **(16 itens)**
(6)	**n5** desde fevereiro , pelo menos quatro casarões foram alvo de vandalismo . **(12 itens)** **n4** fevereiro , pelo menos quatro casarões foram alvo de vandalismo . **(11 itens)** **n3** desde fevereiro , pelo menos quatro casarões foram alvo de vandalismo . **(12 itens)** **n2** . desde fevereiro , pelo menos quatro casarões foram alvo de vandalismo . **(13 itens)** **n1** . desde fevereiro , pelo menos quatro casarões foram alvo de vandalismo . **(13 itens)**
(7)	**n5** tudo indica que há um mercado de azulejos na cidade **(10 itens)** **n4** tudo indica que há um mercado de azulejos na cidade , **(11 itens)** **n3** tudo indica que há um mercado de azulejos na cidade , **(11 itens)** **n2** tudo indica que há um mercado de azulejos na cidade , **(11 itens)** **n1** tudo indica que há um mercado de azulejos na cidade , **(11 itens)**

(8)	**n5** laboratório de conservação e restauração da ufpa **(7 itens)** **n4** os azulejos foram encontrados dias depois [...] no laboratório de conservação e restauração da ufpa **(14 itens)** **n3** no laboratório de conservação e restauração da ufpa **(8 itens)** **n2** laboratório de conservação e restauração da ufpa **(7 itens)** **n1** laboratório de conservação e restauração da ufpa **(7 itens)**
(9)	**n5** na virada do século xix para o xx , **(9 itens)** **n4** na virada do século xix para o xx , **(9 itens)** **n3** na virada do século xix para o xx , **(9 itens)** **n2** na virada do século xix para o xx **(8 itens)** **n1** parte deles foi importada da europa na virada do século xix para o xx , **(15 itens)**
(10)	**n5** com a situação dos casarões históricos da cidade **(8 itens)** **n4** com a situação dos casarões históricos da cidade . **(9 itens)** **n3** com a situação dos casarões históricos da cidade . **(9 itens)** **n2** com a situação dos casarões históricos da cidade . **(9 itens)** **n1** inconformado com a situação dos casarões históricos da cidade . **(10 itens)**
(11)	**n5** aumentam cada vez mais de valor **(6 itens)** **n4** restam aumentam cada vez mais de valor **(7 itens)** **n3** os poucos exemplares de azulejos que ainda restam aumentam cada vez mais de valor **(15 itens)** **n2** exemplares de azulejos que ainda restam aumentam cada vez mais de valor . **(13 itens)** **n1** os poucos exemplares de azulejos que ainda restam aumentam cada vez mais de valor . **(15 itens)**
(12)	**n5 - (0 itens)** **n4** a situação parece ter se agravado **(6 itens)** **n3** a situação parece ter se agravado **(6 itens)** **n2** a situação parece ter se agravado **(6 itens)** **n1** . este ano , a situação parece ter se agravado . **(11 itens)**
(13)	**n5 - (0 itens)** **n4 - (0 itens)** **n3** devem ter sido encarregadas de roubar **(6 itens)** **n2** as pessoas que invadiram devem ter sido encarregadas de roubar azulejos **(11 itens)** **n1** as pessoas que invadiram devem ter sido encarregadas de roubar azulejos **(11 itens)**
(14)	**n5 - (0 itens)** **n4 - (0 itens)** **n3** tiveram azulejos do século xix furtados **(6 itens)** **n2** a situação na cidade causa preocupação [...] tiveram azulejos do século XIX furtados . **(13 itens)** **n1** a proteção do palacete parece encaminhada, mas a situação na cidade causa preocupação, já que outros três casarões tiveram azulejos do século xix furtados **(26 itens)**
(15)	**n5 - (0 itens)** **n4 - (0 itens)** **n3 - (0 itens)** **n2 - (0 itens)** **n1** o assunto vem se espalhando pela capital [...], e há quem suspeite de encomenda de roubos . **(17 itens)**
(16)	**n5 - (0 itens)** **n4 - (0 itens)** **n3 - (0 itens)** **n2 - (0 itens)** **n1** os azulejos foram encontrados dias depois **(6 itens)**

Fonte: Divino (2024, p. 114-115).

Diferentemente do que foi encontrado numa quantidade de, pelo menos, 5% dos textos dos *corpora* n5, n4, n3 e n2, no *corpus* n1 foram encontrados *n-grams* com sequências separadas por ponto final, como comentado anteriormente. Nos outros *corpora*, as sequências, quando contam com a presença de um ponto, apresentam-no no início ou no final. No corpus n1, a cópia extrapola os pontos restritos ao início e ao final dos *n-grams* encontrados nos demais corpora, causando a sobreposição dos grupos (2), (6) e (12), sublinhados de amarelo no Quadro 3, de modo que se pode encontrar todo o trecho "da década de 1970 para cá, 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.", com três informações juntas, o que não foi encontrado em nenhum outro *corpus*. Há cinco *n-grams* em que isto acontece, "dos azulejos se perderam . este", "azulejos se perderam . este ano", "parece ter se agravado. desde", "ter se agravado. desde fevereiro" e "agravado . desde fevereiro , pelo".

Além disso, outros grupos que já faziam parte dos outros *corpora* têm suas sequências expandidas no *corpus* n1, como é o caso do grupo (1), do grupo (5), que têm o termo "iniciou", assim como o *corpus* n3, com a inclusão da vírgula no final, do grupo (9), com adição de "parte deles foi importada da europa", do grupo (10), com adição de "inconformado", do grupo (12), com adição de "os poucos", do grupo (11), e do grupo (14), com adição de "a proteção do palacete parece encaminhada, mas" no início e ", já que outros três casarões" no meio. Foram encontrados no *corpus* n1, também, *n-grams* que não constavam nos demais *corpora*, como é o caso do grupo (15) "o assunto vem se espalhando pela capital [paraense], e há quem suspeite de encomenda de roubos .", e do grupo (16) "os azulejos foram encontrados dias depois".

6.2 Qual a relação dos *n-grams* encontrados em cada sub*corpus* com o material de insumo da tarefa?

O Quadro 3 permite que se faça uma comparação entre os trechos que aparecem em sequência em cada *corpus*. Em comum a todos os *corpora*, há 11 grupos (do grupo 1 ao grupo 11), dos quais apenas o grupo (4) não está presente no material de insumo, porém, percebe-se que há examinandos nivelados em todas as notas que incluíram, em seus textos, a informação da mesma maneira. Por se tratar de uma informação padronizada, percebe-se que sua extensão, em todos os *corpora*, é muito semelhante. Além desse grupo, os grupos (3), (7) e (10) também apresentam extensões semelhantes em todos os *corpora*. Apesar de os grupos (2) e (6) também apresentarem extensões semelhantes, eles possuem um comportamento diferente no *corpus* n1, pelo fato de estarem sobrepostos um ao outro, juntamente com o grupo (12).

Os grupos (1) e (11) apresentam extensões aproximadas nos *corpora* n5 e n4, ao mesmo tempo que, com fragmentos um pouco mais longos, também apresentam extensões aproximadas nos *corpora* n3, n2 e n1. O grupo (5) é particularmente mais curto no *corpus* n5 do que nos demais *corpora*, ao passo que o grupo (8) é mais extenso no *corpus* n4, e o grupo (9) é mais extenso no *corpus* n1. A partir do grupo (12), os trechos passam a não aparecer mais em todos os *corpora*. O grupo (12) aparece em todos os *corpora*, menos no *corpus* n5, contendo a mesma quantidade de itens nos *corpora* n4, n3 e n1, sendo mais longo no n1. Os grupos (13) e (14) não aparecem nem no *corpus* n5, nem no *corpus* n4, e ambos apresentam exatamente a mesma extensão no *corpus* n3, e uma extensão aproximada no *corpus* n2. Apesar de o grupo (13) aparecer com uma extensão similar nos *corpora* n2 e n1, o grupo (14) apresenta um trecho especialmente longo no *corpus* n1. Algo parecido ocorre com o grupo (15), com o agravante de que essa informação foi encontrada apenas no *corpus* n1, assim como o grupo (16).

Entre os grupos de informações localizados no material de insumo, o *corpus* n5 deu conta de 10 trechos, fragmentados nos *n-grams*, o *corpus* n4 deu conta de 11, o *corpus* n3 e o *corpus* n2 deram conta de 13, e o *corpus* n1 deu conta de 15. De modo geral, há mais *n-grams* referentes

a cada trecho nas notas mais baixas do que nas notas mais altas, o que faz com que uma parte maior do trecho do texto de insumo esteja localizada nesses *corpora*. Comparativamente, enquanto o *corpus* n5 apresenta cópia de partes mais localizadas dos trechos (por exemplo, "aumentam cada vez mais de valor"), o *corpus* n1 apresenta a cópia de mais partes de um mesmo trecho (por exemplo, "os poucos exemplares de azulejos que ainda restam aumentam cada vez mais de valor.").

Embora haja cópia em todos os *corpora*, os fatos acima apresentados permitem que se considere que os examinandos que tiram notas mais altas, além de copiarem fragmentos mais curtos do texto de insumo e do enunciado, escolhem copiar informações essenciais para o cumprimento da tarefa. Já os examinandos que tiram notas mais baixas copiam fragmentos mais longos do material de insumo e, embora copiem informações que são copiadas também por examinandos com notas mais altas, realizam a cópia de fragmentos que talvez não sejam tão relevantes para o cumprimento da tarefa. As frequências da maioria dos *n-grams* longos são baixas, e os *corpora* com mais *n-grams* apresentam muitos destes com frequência abaixo de 10%. Pode-se supor, com isto, que os textos com notas mais baixas copiem trechos mais longos de informações diferentes, não apresentando, entre si, um padrão nas escolhas e trechos copiados. O padrão está no fato de que há mais cópia.

Entre as sequências mais longas (6 *n-grams*) presentes em cada *corpus*, encontram-se cópias exatas de trechos presentes no texto de insumo. Percebe-se que nos *corpora* de notas mais baixas, há mais trechos copiados exatamente como aparecem no material de insumo. No *corpus* n5, foram identificados 2 *n-grams* dessa extensão com *Relative DOCF* superior a 20%, o que não acontece nos demais *corpora*. No *corpus* n4, há sete *n-grams* com incidência em, pelo menos, 20% dos textos, e nos *corpora* n3 e n2, há oito. No *corpus* n1, embora haja apenas três *n-grams* com incidência acima de 20%, dois deles estão em mais de 30% dos textos, o que não ocorre em nenhum outro *corpus*. Entre os cinco primeiros *n-grams* de seis de todas as notas observadas, estão "o palacete vitor maria da silva" e a informação sobre Belém ser "uma das cidades com maior variedade de azulejos" do Brasil. Foram encontrados, também, *n-grams* que, sobrepostos, indicam trechos com um maior número de elementos copiados exatamente da forma como estão no material de insumo.

7 DISCUSSÃO DOS RESULTADOS

7.1 Qual a relação do léxico usado em cada sub*corpus* com o material de insumo da tarefa?

Por um lado, os resultados deste trabalho corroboram Mendel (2019) em relação à cópia de trechos do texto de insumo em textos de notas 3, 2 e 1, mas, por outro lado, os achados deste trabalho afirmam que também é possível que haja cópia de trechos com extensão de seis itens lexicais em textos avaliados com nota 4 e nota 5.

Quando olhamos para os *n-grams* presentes em, pelo menos, 5% dos textos do *corpus* n5 e n4, encontramos uma quantidade de 59 *n-grams* no *corpus* n5 e 69 no *corpus* n4. Esses dados indicam que há cópias de trechos de, pelo menos, seis itens em sequência em textos avaliados com 5 e 4. A sobreposição de *n-grams* indica que pode haver cópia, inclusive, de fragmentos mais longos nesses dois *corpora*. Relativo especificamente à extensão e à quantidade de trechos, por mais que tenham sido encontrados *n-grams* que, somados uns aos outros, correspondem a fragmentos relativamente extensos do material de insumo (de até 22 itens), eles correspondem a informações localizadas do material. Quanto aos grupos de informação, o *corpus* n5 apresenta *n-grams* que se encaixam em 10 grupos, e o *corpus* n4, em 11. Em quase todos os casos, os trechos copiados estão diretamente relacionados ao conteúdo informacional necessário para cumprir o propósito comunicativo solicitado na tarefa. Esse

fato permite que se possa inferir que, nesses dois *corpora*, a cópia se dá, majoritariamente, de informações imprescindíveis para o cumprimento da tarefa.

Quanto aos *corpora* n3 e n2, estes apresentam iguais quantidades de *n-grams* quando levamos em conta a ocorrência em ao menos 20% dos textos, com oito *n-grams* em cada, e o mesmo ocorre quando levamos em conta 10% dos textos, com 28 *n-grams* em cada. Em pelo menos 5% dos textos, ambos os *corpora* apresentam quantidades maiores de *n-grams* do que as encontradas nos *corpora* n4 e, sobretudo, n5, com 83 no *corpus* n3 e 80 no *corpus* n2. Entre os grupos de informações encontrados entre os *n-grams* do *corpus* n3 em relação aos *corpora* n5 e n4, não há uma diferença muito grande quanto à extensão dos fragmentos, no entanto, o *corpus* n3 apresenta *n-grams* que correspondem a 13 grupos de informações. A partir desses resultados, é possível concordar com a afirmação de Sirianni (2016, p. 44), de que, entre os textos que receberam nota 3 nessa tarefa, "alguns textos trazem mais informações do que o que seria necessário e que essas informações trazidas não teriam tanta relevância para o cumprimento da tarefa", visto que há uma maior quantidade de fragmentos do material de insumo, incorporados *ipsis litteris* aos textos, que não têm uma relação tão direta com o cumprimento do propósito comunicativo da tarefa.

No *corpus* n2, os resultados corroboram Mendel (2019, p. 218) na declaração de que, no conjunto de textos que receberam essa nota, alguns "textos são bastante dependentes não apenas das informações, mas também da organização dessas informações e dos recursos linguísticos do material de insumo, o que pode incluir a presença de alguns trechos copiados", pois alguns *n-grams* presentes nesse *corpus*, quando agrupados, dão conta de trechos mais extensos do material de insumo se comparados aos *corpora* n5 e n4. Os textos do *corpus* n2 podem apresentar "problemas de atribuir um novo contexto às informações do material de insumo" (Mendel, 2019, p. 118), podendo fazer com que o texto assuma um caráter mais informativo, como é típico em reportagens, como é o caso do gênero do texto de insumo, o que ocasiona o cumprimento parcial do propósito (solicitar) e a configuração não tão adequada do gênero solicitado na tarefa (carta aberta) (Sirianni, 2016). Tanto no *corpus* n3 quanto no *corpus* n2, os fragmentos copiados são mais espalhados no texto de insumo, havendo cópias de trechos que vão além do que seria essencial para o cumprimento da tarefa. Pode-se inferir que, comparativamente aos *corpora* n5 e n4, os *corpora* n3 e n2 contêm uma quantidade maior de informações secundárias copiadas nos textos.

O *corpus* n1 apresenta traços singulares em relação aos demais *corpora*, sendo o único *corpus* a: conter *n-grams* de seis itens em mais de 30% dos textos; não incluir nenhum trecho relacionado ao Palacete Vitor Maria da Silva em uma quantidade de 20% dos textos ou mais; ter mais de 30 *n-grams* em, pelo menos, 10% dos textos, totalizando 42; ter mais de 90 *n-grams* em, pelo 5% dos textos, totalizando 129; e apresentar *n-grams* que indicam que houve cópia de trechos do material de insumo que são separados por um ponto final entre duas palavras. A separação dos fragmentos do material de insumo em grupos foi delimitada durante o agrupamento dos *n-grams* nos *corpora* n5, n4, n3 e n2, estabelecendo que os grupos não ultrapassariam o ponto final, pois, até então, nenhum dos *corpora* havia apresentado *n-grams* com essa característica. Quando os *n-grams* do *corpus* n1 foram agrupados, foi encontrada a sobreposição de três dos grupos previamente estabelecidos, em que a sobreposição destes *n-grams* corresponde a um fragmento de 38 itens em sequência, de forma que se pode concordar com a afirmação de que "o nível 1 do Celpe-Bras [...] se caracteriza por textos com trechos longos copiados" (Mendel, 2019, p. 131). Além de *n-grams* que incluem pontos finais, o *corpus* n1 também apresenta *n-grams* que dizem respeito a 15 trechos do material de insumo, dois a mais do que a quantidade encontrada nos *corpus* n3 e n2. Consequentemente, isto indica que há ainda mais informações copiadas que não se relacionam diretamente com o que é imprescindível para cumprir com a tarefa, o que está de acordo com as conclusões de Sirianni (2016), quando a autora sugere que algumas informações recuperadas do texto de insumo podem ser até consideradas irrelevantes para o cumprimento do propósito.

Apesar de todos os *corpora* apresentarem *n-grams* com esta *Relative DOCF* de 20%, a presença dos *n-grams* aumenta gradualmente do *corpus* n5 ao *corpus* n3/n2. O fato de ser possível agrupar os *n-grams* encontrados em, pelo menos, 5% dos textos, indica certa similaridade, em todos os *corpora*, das partes do material de insumo (texto e enunciado) de onde informações foram incorporadas às produções. Em todos os *corpora*, foi identificada a cópia de trechos diretamente relacionados ao conteúdo informacional definido como essencial para o cumprimento da tarefa, indicando que a simples cópia de informações relevantes para o contexto comunicativo não é suficiente para cumprir satisfatoriamente a tarefa (Mendel, 2019). Não se pode deixar de ressaltar que 5% é um percentual baixo para afirmar que um *corpus* apresenta a recuperação recorrente de alguma informação específica. O que se pode inferir, a partir disso, é que o padrão está no fato de haver mais cópia de trechos diversos conforme a nota baixa, especialmente no *corpus* n1, que parece fazer mais uso de informações transcritas exatamente da maneira como estão no material de insumo.

8 CONSIDERAÇÕES FINAIS

Dados provenientes de *corpora* de línguas adicionais podem revelar características importantes sobre o desempenho no uso da língua em diferentes níveis de proficiência (Mendes *et al.*, 2016). Resultados advindos de estudos que buscam uma classificação confiável do desempenho dos examinandos utilizando análise de *corpus* são úteis não apenas para que se tenha um maior entendimento do significado das classificações de proficiência em um nível empírico (Callies; Díez-Bedmar; Zaytseva, 2014; Barker; Salamoura; Saville; 2015; Callies; Götz, 2015), mas também para que se possa aumentar a confiabilidade e a coerência de testes de proficiência (Wisniewski, 2017).

Como é possível verificar nos resultados apresentados neste artigo, análises de corpora podem auxiliar na descrição dos diferentes níveis avaliados em exames de proficiência. No caso aqui apresentado, é possível afirmar que o padrão de cópia de informações do texto de insumo varia consideravelmente entre os diferentes níveis, apresentando aumento nas notas mais baixas, tanto em relação à extensão da cópia quanto em relação aos grupos de informações copiadas. Em que pese as tarefas integradas solicitarem recuperação de informações do material de insumo para caracterizar o adequado cumprimento do propósito comunicativo solicitado, é possível afirmar que cópias literais do material de insumo são consideradas adequadas apenas quando recuperam informações imprescindíveis e padronizadas, como é o caso do nome do casarão histórico na tarefa analisada neste artigo. Cópias longas de vários dos trechos do material de insumo não são bem avaliadas, visto que textos com essas características receberam nota 1 em grande parte dos dados analisados.

Estudos como este, que usam análise de *corpus* para descrever os níveis de proficiência, ainda são recentes sobre o Celpe-Bras, apesar de já serem abundantes com dados de exames internacionais há mais de 20 anos. A compilação do CorCel (Schoffen *et al.*, no prelo) tem potencial para fomentar mais pesquisas que utilizem essa metodologia, o que pode acelerar os processos de descrição dos níveis de proficiência avaliados no exame, contribuindo para a sua validade. Pesquisas como esta têm também o poder de contribuir tanto para que professores tenham mais acesso a características definidoras dos gêneros do discurso avaliados pelo exame e possam aprimorar seus materiais didáticos com cada vez mais dados empíricos, quanto para avaliadores, dando a eles mais subsídios para compreender mais profundamente os níveis de proficiência descritos nos parâmetros de avaliação.

REFERÊNCIAS

BAKHTIN, M. M. Estética da criação verbal. São Paulo: Martins Fontes, 2003.

BARKER, F.; SALAMOURA, A.; SAVILLE, N. Learner corpora and language testing. In: GRANGER, S.; GILQUIN, G.; MEUNIER, F. (Ed.). The Cambridge handbook of learner corpus research. Cambridge: Cambridge University, 2015. p. 511–534.

BIBER, D. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In: HEINE, B.; NARROG, H. (Ed.). The Oxford Handbook of Linguistic Analysis. Oxford: Oxford Academic, 2012. p. 159-192.

BRASIL. Instituto Nacional de Estudos e Pesquisas Anísio Teixeira. Documento-base do exame Celpe-Bras. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020.

BRASIL. Instituto Nacional de Estudos e Pesquisas Anísio Teixeira. Celpe-Bras – Caderno de questões: parte escrita. Edição 2015/II. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2015. Disponível em: https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/PARTE-ESCRITA_CELPEBRAS_2015_II_Caderno-de-questoes.pdf. Acesso em: 1 jul. 2025.

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

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

CUSHING, S. T. Corpus linguistics in language testing research. Language Testing, [S.l], v. 34, n. 4, p. 441-449, Sept. 2017. Disponível em: https://doi.org/10.1177/0265532217713044. Acesso em: 15 dez. 2024.

CUSHING, S. T. Corpus Linguistics and Language Testing. In: FULCHER, G.; HARDING, L. (Ed.). The Routledge Handbook of Language Testing. London; New York: Routledge, 2022. p. 545-560.

DELL'ISOLA, R. L. P., PORDEUS, I. R. Os enunciados de tarefas integradas de leitura e escrita do Exame Celpe-Bras. Inventário, Niterói-RJ, p. 65–78, 2021. Disponível em: https://periodicos.ufba.br/index.php/inventario/article/view/27218. Acesso em: 15 dez. 2024.

DIVINO, L. S. Índices lexicais de análise para a caracterização dos níveis intermediário e avançado superior no Exame Celpe-Bras. Orientadora: Juliana Roquele Schoffen. 2021. 68f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021. Disponível em: https://lume.ufrgs.br/handle/10183/235361. Acesso em: 9 jul. 2024.

DIVINO, L. S. Contribuições da linguística de corpus para a definição de níveis de proficiência escrita no exame Celpe-Bras. 2024. 195f. Dissertação (Mestrado em Letras) –Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024. Disponível em: http://hdl.handle.net/10183/282481. Acesso em: 1 jul. 2025.

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

EVERS, A. Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras. 2013. 174f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: https://lume.ufrgs.br/handle/10183/79447. Acesso em: 15 dez. 2024.

GRIES, S. T. Useful statistics for corpus linguistics. In: SÁNCHEZ, A.; ALMELA, M. (Ed.). A mosaic of corpus linguistics: selected approaches. Frankfurt am Main: Peter Lang, 2010. p. 269-291.

HANAUER, I. D. Caracterização dos níveis intermediário e avançado superior do exame Celpe-Bras em produções escritas de examinandos no gênero carta/e-mail: contribuições de uma análise guiada por corpus. 2023. 84f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

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

KUNILOVSKAYA, M.; KOVIAZINA, M. Sketch engine: A toolbox for linguistic discovery. Journal of Linguistics/Jazykovedný casopis, Alemanha, v. 68, n. 3, p. 503-507, dez. 2017. Disponível em: https://www.researchgate.net/publication/324760982_Sketch_Engine_A_Toolbox_for_Linguistic_Discovery. Acesso em: 01 jul. 2025.

KUNRATH, S. P. Os descritores gerais e a progressão dos níveis de proficiência do Exame Celpe-Bras. 2019. 198f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019. Disponível em: https://lume.ufrgs.br/handle/10183/239041. Acesso em: 15 dez. 2024.

MENDEL, K. Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras. 2019. 177f. Dissertação (Mestrado em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019. Disponível em: https://lume.ufrgs.br/handle/10183/217789. Acesso em: 15 dez. 2024.

MENDES, A. et al. The COPLE2 Corpus: a Learner Corpus for Portuguese. Proceedings of LREC 2016, Portorož, Slovenia, p. 3207–3214, May 2016. Disponível em: https://aclanthology.org/L16-1511/. Acesso em: 1 jul. 2025.

PILEGGI, M. G. S. Integração de habilidades: perspectiva histórico-teórica e operacionalização no exame Celpe-Bras. Estudos Linguísticos, São Paulo, v. 46, n. 2, p. 577-592, 2017.

RAUPP, A. M. Características das produções escritas do exame Celpe-Bras na Tarefa 3 de 2016-2: uma pesquisa guiada por corpus. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024.

SARDINHA, T. B. Linguística de Corpus: histórico e problemática. DELTA., São Paulo, v. 16, n. 2, p. 323-367, 2000. Disponível em: https://doi.org/10.1590/S0102-44502000000200005. Acesso em: 15 dez. 2024.

SCARAMUCCI, M. V. R. Proficiência em LE: considerações terminológicas e conceituais. Trabalhos em Linguística Aplicada, Campinas-SP, v. 36, n. 1, p. 11-22, 2000. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/tla/article/view/8639310. Acesso em: 15 dez. 2024.

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

SCHOFFEN, J. et al. Compilation and tagging of a corpus with Celpe-Bras texts. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics, 2024. p. 627-632. Disponível em: https://aclanthology.org/2024.propor-1.72.pdf. Acesso em: 1 jul. 2024.

SCHOFFEN, J. et al. CorCel: a Brazilian Portuguese corpus of Celpe-Bras exam written texts. [S.l,; s.n.e.], 2025. No prelo.

SCHOFFEN, J.; DIVINO, L. Contribuição da Linguística de Corpus para a formação de professores de PLA: sugestões a partir da descrição dos níveis de proficiência avaliados em uma tarefa da Parte Escrita do Celpe-Bras. Letras de Hoje, [S. l.], v. 58, n. 1, p. e44904, 2023. DOI: 10.15448/1984-7726.2023.1.44904. Disponível em: https://revistaseletronicas.pucrs.br/fale/article/view/44904. Acesso em: 1 jul. 2025.

SEGAT, G. Estudos sobre Confiabilidade em Exames de Proficiência: o processo de atribuição de notas e a reavaliação na Parte Escrita do Celpe-Bras. 2023. 163f. Dissertação (Mestrado em Linguística Aplicada) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

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

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. 2016. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: https://lume.ufrgs.br/handle/10183/157769. Acesso em: 15 dez. 2024.

SOSTRUZNIK, J. L. O uso de conjunções em produções escritas no exame Celpe-Bras: um estudo baseado em corpus. 2013. 76f. Trabalho de Conclusão de Curso (Graduação em Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023. Disponível em: https://lume.ufrgs.br/handle/10183/270404. Acesso em: 15 dez. 2024.

MENDEL, K.; SCHOFFEN, J. R. Integrated assessment in the Celpe-Bras exam and tasks of reading and writing. BELT - Brazilian English Language Teaching Journal, [S. l.], v. 8, n. 2, p. 148–170, 2017. DOI: 10.15448/2178-3640.2017.2.28568. Disponível em: https://revistaseletronicas.pucrs.br/belt/article/view/28568. Acesso em: 1 jul. 2025.

TOSATTI, N. M. O desempenho de estudantes de países africanos de língua oficial portuguesa no Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). 2021. 260f. Tese

(Doutorado em Linguística Aplicada) – Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: https://repositorio.ufmg.br/handle/1843/36528. Acesso em: 15 dez. 2024.

VICENTINI, M. P. As tarefas de compreensão oral para produção escrita no Celpe-Bras: construto e operacionalização. Revista Estudos Linguísticos, São Paulo, v. 48, n. 1, p. 561-580, abr. 2019. Disponível em: https://doi.org/10.21165/el.v48i1.2141. Acesso em: 15 dez. 2024.

VICENTINI, M. P. As dimensões do construto compreensão oral para produção escrita no Exame Celpe-Bras: percepções, processos, estratégias e desempenhos e examinandos. 2022. 241f. Tese (Doutorado em Linguística Aplicada) – Universidade Estadual de Campinas, Campinas-SP, 2022. Disponível em: https://repositorio.unicamp.br/acervo/detalhe/1253389. Acesso em: 1 jul. 2025.

WISNIEWSKI, K. Empirical learner language and the levels of the Common European Framework of Reference. Language Learning, Michigan, v. 67, n. 1, p. 232-253, Jan. 2017. Disponível em: http://dx.doi.org/10.1111/lang.12223. Acesso em: 1 jul. 2025.

Contribuição dos autores

Luiza Sarmento Divino e Juliana Roquele Schoffen participaram conjuntamente da redação do artigo, desde a concepção e análise dos dados até a revisão final do texto, sendo o trabalho baseado na dissertação de mestrado de Luiza, sob orientação de Juliana.