SCHOFFEN, Juliana R.; STUMPF, Elisa M. Corpus Celpe-Bras (CorCel): contribuições para pesquisa, avaliação e ensino de Português como Língua Adicional. *ReVEL,* edição especial, v. 23, n. 22, p. 1-38, 2025. [www.revel.inf.br].

Corpus Celpe-Bras (CorCel): contribuições para pesquisa, avaliação e ensino de Português como Língua Adicional

Celpe-Bras Corpus (CorCel): contributions to research, assessment and teaching of Portuguese as an Additional Language

Juliana Roquele Schoffen[1]
Elisa Marchioro Stumpf[2]

julianaschoffen@gmail.com
elisa.stumpf@gmail.com

RESUMO: Este artigo apresenta o *Corpus Celpe-Bras* (CorCel), um corpus de textos escritos por examinandos do Celpe-Bras. O CorCel reúne 15.315 textos produzidos em quatro edições do exame e classificados de acordo com os níveis de proficiência avaliados, oferecendo uma base empírica inédita para pesquisas sobre o desempenho linguístico em português como língua adicional. Fundamentado em uma concepção discursiva de linguagem, o projeto entende os textos como práticas sociais situadas, articulando propósitos comunicativos, gêneros do discurso e contextos de produção. As análises já realizadas revelam diferenças significativas entre os textos avaliados com diferentes notas, especialmente em relação à extensão do texto, diversidade lexical, uso de conjunções, adequação ao gênero e configuração da interlocução, fornecendo evidências empíricas para o refinamento dos descritores e parâmetros de avaliação do exame. O CorCel amplia as possibilidades de pesquisa em Linguística Aplicada, avaliação de proficiência e ensino de português como língua adicional, favorecendo investigações sobre o desenvolvimento da proficiência escrita, o uso de recursos linguístico-discursivos e a validade de construto do Celpe-Bras. Além disso, o corpus servirá de base para o desenvolvimento de uma plataforma de avaliação automatizada e de ferramentas de *feedback* baseadas em inteligência artificial, contribuindo para práticas pedagógicas orientadas por dados empíricos e para o avanço das pesquisas sobre ensino e avaliação de português como língua adicional.
PALAVRAS-CHAVE: CorCel; Exame Celpe-Bras; português como língua adicional, avaliação de proficiência

ABSTRACT: This article presents the *Celpe-Bras Corpus* (CorCel), a collection of written texts produced by candidates of the Celpe-Bras exam, the official Brazilian Certificate of Proficiency in Portuguese as an Additional Language. The CorCel comprises 15,315 texts from four editions of the exam, rated according to assessed proficiency levels, and provides an unprecedented empirical basis for research on linguistic performance in Portuguese as an additional language. Grounded in a discursive conception of language, the project conceives texts as situated social practices that articulate communicative purposes, discourse genres, and production contexts. Analyses conducted so far reveal

[1] Doutora em Linguística Aplicada. Professora do Departamento de Letras Clássicas e Vernáculas e do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.
[2] Doutora em Análises Textuais, Discursivas e Enunciativas. Professora do Departamento de Línguas Modernas e do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

significant differences among texts rated at different proficiency levels, particularly regarding text length, lexical diversity, use of conjunctions, genre adequacy, and audience awareness. These findings offer empirical evidence to refine the exam's descriptors and rating parameters. The CorCel Corpus expands research possibilities in Applied Linguistics, proficiency assessment, and the teaching of Portuguese as an additional language, supporting investigations into the development of written proficiency, the use of linguistic and discursive resources, and the construct validity of Celpe-Bras assessment criteria. Furthermore, the corpus will serve as the foundation for developing an automated scoring platform and AI-based feedback tools, contributing to data-driven pedagogical practices and advancing research on the teaching and assessment of Portuguese as an additional language.
KEYWORDS: CorCel, Celpe-Bras Exam, Portuguese as an additional language, proficiency assessment.

Introdução

Este artigo apresenta o Corpus Celpe-Bras (CorCel), um corpus de textos escritos por examinandos do exame Celpe-Bras, o certificado oficial de proficiência em português como língua adicional do Brasil. O objetivo principal da compilação do CorCel é fornecer uma base empírica robusta para estudos sobre desempenho linguístico de falantes de português como língua adicional (Schoffen et al. 2024). O corpus compreende 15.315 textos produzidos em 4 edições do exame, abrangendo diferentes níveis de proficiência.

A compilação do CorCel busca promover um diálogo mais estreito entre pesquisa acadêmica e práticas pedagógicas, contribuindo para fomentar pesquisas em linguística aplicada e avaliação de proficiência, bem como apoiar o desenvolvimento de abordagens de ensino baseadas em dados de uso da língua. Os dados do CorCel podem fomentar pesquisas que aprofundem a compreensão sobre a relação entre desempenho linguístico e níveis de proficiência em português como língua adicional (PLA), que poderão se refletir na formulação de políticas linguísticas e práticas de ensino e avaliação mais informadas por dados empíricos, como elaboração de materiais didáticos, propostas curriculares, instrumentos de avaliação e descritores de proficiência.

O CorCel foi desenvolvido com o objetivo de criar um corpus representativo de textos escritos produzidos por examinandos do Celpe-Bras. A construção do corpus envolveu um conjunto de decisões metodológicas relativas à seleção e ao tratamento dos dados, bem como à definição de critérios éticos e linguísticos para sua organização. A proposta do CorCel articula-se a uma perspectiva discursiva de linguagem que entende o uso da língua como ação situada, mediada por gêneros do discurso. Isso significa que os textos reunidos no corpus são compreendidos não apenas como produtos linguísticos, mas como práticas sociais, resultantes da

interação entre propósitos comunicativos, interlocutores e contextos socioculturais, neste caso circunscritos duplamente, tanto pelos enunciados das tarefas (que propõem um contexto de produção para os textos a serem escritos) quanto pela situação avaliativa (que considera o propósito de performar de modo mais proficiente possível tendo em vista a interlocução institucional com um avaliador).

Este artigo descreve a organização e a estrutura do CorCel, os critérios de seleção e pré-processamento dos textos, além das possibilidades de análise que o corpus oferece para estudos sobre gêneros discursivos, estratégias discursivas e desenvolvimento da proficiência. Além desta introdução, este artigo está organizado em 7 seções, que se propõem a: 1- apresentar o exame Celpe-Bras; 2- relatar estudos sobre Linguística de Corpus e destacar a importância que esses estudos têm ganhado na área de avaliação de proficiência; 3- apresentar o CorCel, relatando os procedimentos realizados para compilação e preparação do corpus, bem como os metadados disponíveis; 4- relatar resultados de estudos já realizados sobre o CorCel; 5- apresentar possíveis contribuições do CorCel para os estudos sobre o Celpe-Bras e para a área de PLA; 6- elencar perspectivas futuras de pesquisas a partir da disponibilização do CorCel; e 7-apresentar as considerações finais.

1. O Celpe-Bras

O Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras), exame de proficiência em língua portuguesa desenvolvido e aplicado pelo Ministério da Educação do Brasil desde 1998, tem sido o objeto de grande parte das reflexões teóricas sobre avaliação de proficiência publicadas no Brasil. Elaborado com o objetivo de avaliar a proficiência necessária para ingresso na universidade pelos estudantes candidatos ao Programa de Estudante Convênio (PEC-G)[3], o exame foi se tornando relevante também para outros públicos e vem sendo objeto de uma consistente procura tanto de instituições interessadas em atuarem como postos aplicadores quanto de examinandos interessados em comprovar proficiência em língua portuguesa. O exame conta hoje com 146 instituições credenciadas como

[3] O Programa de Estudantes-Convênio de Graduação (PEC-G) é um programa do governo brasileiro com objetivo de oferecer a oportunidade de realizar estudos de graduação em Instituições de Ensino Superior brasileiras a estudantes de países em desenvolvimento com os quais o Brasil mantém acordo de cooperação educacional e/ou cultural. Disponível em: https://www.gov.br/mre/pt-br/assuntos/cultura-e-educacao/temas-educacionais/programas-de-estudo-para-estrangeiros/pec-g. Acesso em: 01 out. 2025.

postos aplicadores e vem sendo aplicado anualmente para um grande número de examinandos, como podemos ver no gráfico a seguir.

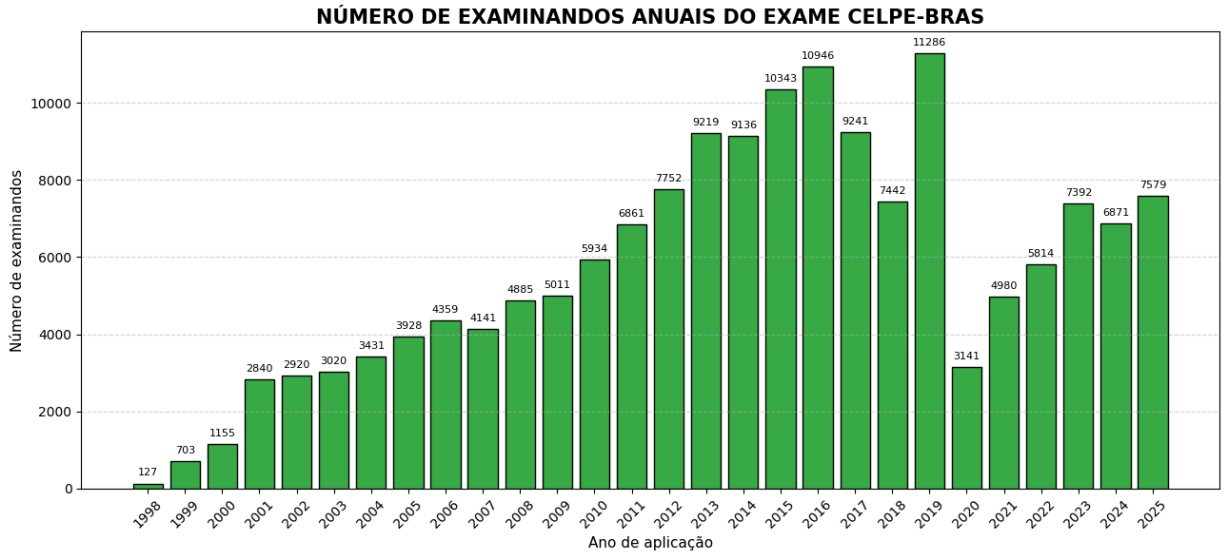

Figura 1: Número de examinandos anuais do exame Celpe-Bras
Fonte: https://www.ufrgs.br/acervocelpebras/dados-celpe-bras/

Fundamentado em uma "visão de uso da língua(gem) com propósitos sociais, construída social e localmente por seus participantes" (Inep 2020: 28), o exame entende que proficiência "implica ser capaz de engajar-se em diferentes situações de uso da língua portuguesa no mundo, mostrando adequação às demandas dos vários contextos" (Inep 2020: 28). O Celpe-Bras consiste em uma parte oral, que compreende uma interação de vinte minutos, e uma parte escrita, composta por quatro tarefas integradas de compreensão oral, leitura e produção de textos. A parte escrita do exame é baseada na noção bakhtiniana de gêneros do discurso (Bakhtin 2003), segundo a qual os textos sempre "são produzidos em uma situação comunicativa específica, por alguém com um papel específico nessa situação, e são endereçados a interlocutores com propósitos específicos e também produzidos em um contexto, que também produz sentido" (Schlatter et al. 2009: 105-106). Ser proficiente, dessa perspectiva, significa ser capaz de construir enunciados adequados para participar de contextos de comunicação em diferentes esferas de uso da linguagem (Schoffen 2009).

A partir do seu construto teórico e da discussão sobre proficiência que o exame ensejou no Brasil, o Celpe-Bras vem, ao longo dos anos, exercendo papel de

direcionador dos processos educacionais em português como língua adicional (Schoffen e Martins 2016; Dorigon 2016). Além das diversas pesquisas que mostram essa influência (Costa 2013; Mittelstadt 2013; Ohlweiler 2006; Yan 2008; Li 2009; Bortolini 2006, Nagasawa 2016; 2018; Martins 2022 para citar algumas), as *Propostas curriculares para o ensino de português no exterior* publicadas pelo Itamaraty (MRE 2020a; MRE 2020b) também utilizam os níveis certificados pelo Celpe-Bras como parâmetros para a organização dos currículos. Todos esses fatores configuram o Celpe-Bras como um exame de alta relevância, o que torna fundamental o incentivo às pesquisas sobre os diversos aspectos envolvidos nesse sistema de avaliação, especialmente sobre os níveis certificados pelo exame.

A partir de uma única prova, o Celpe-Bras certifica quatro níveis de proficiência: Intermediário, Intermediário Superior, Avançado e Avançado Superior. A parte escrita, origem dos dados compilados no CorCel, é composta por tarefas que avaliam compreensão e produção de diferentes gêneros do discurso de forma integrada. As tarefas são definidas pelo exame como "'convite' para o participante usar a língua em diversos contextos, desempenhando papéis com variados propósitos e distintos interlocutores, produzindo textos de uma série de gêneros discursivos, que circulam em diversos suportes" (Inep 2020: 31). De acordo com o Documento Base do Celpe-Bras,

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

A avaliação da Parte Escrita do Celpe-Bras é realizada em uma escala de 0 a 5, utilizando-se parâmetros de avaliação holísticos e singularizáveis para cada tarefa. Cada texto é avaliado independentemente por dois avaliadores e, em caso de discrepância maior do que um ponto entre as notas atribuídas, o texto é encaminhado para reavaliação. A nota final atribuída ao texto é a média aritmética entre as notas dos dois avaliadores (ou do terceiro, em caso de reavaliação). A nota final da parte escrita é a média entre as notas finais das quatro tarefas e o nível de certificação atribuído a cada examinando corresponde à nota mais baixa atingida entre as partes escrita e oral do exame.

Apesar de muitos terem sido os estudos já realizados sobre o exame Celpe-Bras[4], apenas alguns analisaram textos e interações orais produzidos por examinandos antes da compilação do CorCel (Sidi 2002; Schoffen 2009; Gomes 2009; Tosatti 2021; Vicentini 2022, para a parte escrita; Schoffen 2003; Fortes 2009, para a parte oral). Todos esses estudos empregam metodologias qualitativas e analisam um número limitado de amostras. Apesar de os estudos já realizados terem colaborado para o aprimoramento do exame ao longo dos anos[5], a falta de um corpus robusto de textos produzidos e avaliados no Celpe-Bras tem impedido a realização de estudos quantitativos, envolvendo um número maior de textos, até a compilação do CorCel[6]. Esse fato tem limitado o uso de métodos e ferramentas automatizadas para descrever as características de cada nível de proficiência, mais especificamente as ferramentas de Linguística de Corpus (LC), que têm sido usadas consistentemente na área de avaliação de proficiência nas últimas décadas (Cushing 2017; Banerjee, Franceschina e Smith 2007; Biber e Gray, 2013; Paquot 2019). A compilação do CorCel, portanto, vem atender a uma antiga demanda da área e oferece novas possibilidades para a pesquisa em avaliação de proficiência em língua portuguesa.

2. Estudos em Linguística de Corpus na área de avaliação de proficiência

Avaliações de larga escala, como concursos para cargos públicos e processos seletivos para universidades, bem como exames de proficiência em línguas, têm uma importância muito grande na sociedade contemporânea, visto que são mecanismos definidores de possibilidades de acesso a oportunidades acadêmicas e profissionais (*gatekeepers*). Esses exames são considerados de alto impacto (*high stakes*), uma vez que costumam ter múltiplos efeitos na sociedade, especialmente no campo educacional, que podem ser verificados na oferta e no currículo de cursos, na elaboração de materiais didáticos, na formação de professores e no próprio processo de ensino-aprendizagem. Esses impactos, chamados pela literatura da área de Efeitos Retroativos (Alderson e Wall 1993; Scaramucci 2004; 2011; Taylor 2005; Wall 2012;

[4] A lista de estudos já realizados sobre o Celpe-Bras pode ser acessada em https://www.ufrgs.br/acervocelpebras/pesquisas/.
[5] Ver Inep (2020) para um histórico de mudanças nas grades de avaliação da parte escrita e da parte oral.
[6] Até a compilação do CorCel, apenas o estudo de Evers (2013) havia utilizado ferramentas de Linguística de Corpus para identificar elementos coesivos e lexicais em textos de diferentes níveis de proficiência produzidos por examinandos. O estudo analisou o corpus compilado para o trabalho de Schoffen (2009), contendo 181 textos.

Cheng e Sultana 2022), enfatizam a necessidade de elaboração e aprimoramento de sistemas de avaliação pautados em conceitos de validade (Chapelle 2012), confiabilidade (Weigle 2002; Schlatter et al. 2005) e ética (Walters 2022; Harding e Fulcher 2022, Shohamy 2006).

Exames internacionais de proficiência em larga escala têm se embasado nas pesquisas acadêmicas desenvolvidas ao longo dos anos para desenvolver instrumentos de avaliação mais autênticos e válidos (Weigle 2002), tendo vários exames passado por reformulações em seus instrumentos e grades de avaliação para atender às discussões contemporâneas sobre validade e ética na avaliação (Cumming et al. 2006; Guo 2011; Gebril e Plakans 2013). No que diz respeito às práticas avaliativas do Celpe-Bras, os resultados de pesquisas acadêmicas (Sidi 2002; Lima 2008; Gomes 2009; Schoffen 2003; 2009; Fortes 2009; Evers 2013; Sirianni 2016; 2020; Queiroz 2017; Schoffen et al. 2018; Kunrath 2019; Mendel 2017; 2019; Souza Neto 2018, para mencionar algumas) também têm contribuído muito para o aprimoramento dos instrumentos e das práticas de avaliação. Dentre as contribuições dessas pesquisas, destacam-se a descrição mais refinada dos níveis de proficiência avaliados no exame, o aprimoramento das grades de avaliação e uma descrição mais detalhada das especificações do Celpe-Bras, a fim de oferecer aos usuários uma certificação que reflita as necessidades contemporâneas de uso da língua, aumentando assim sua validade (ver histórico detalhado em Inep 2020).

Internacionalmente, o uso de ferramentas automáticas de análise possibilitadas pelas novas tecnologias tem auxiliado muito a área de avaliação de proficiência em línguas, especialmente as análises realizadas utilizando-se de ferramentas da Linguística de Corpus (LC). Segundo Gablasova (2019: 45), "corpora de linguagem são conjuntos de dados eletrônicos que contêm amostras de uso da língua e fornecem evidências sobre como a língua é usada na comunicação real". Para a autora, a possibilidade de analisar de forma robusta a produção de aprendizes e a possibilidade de encontrar padrões de uso da linguagem torna os métodos de análise de corpus um recurso valioso também em exames de proficiência, contribuindo para o objetivo central desses testes, qual seja, reunir de forma sistemática exemplos de uso da língua para fazer inferências sobre a habilidade linguística de um determinado falante e sua capacidade de usar a língua (Chapelle e Plakans 2013).

Desde a década de 1990, quando Alderson (1996) identificou potenciais aplicações da análise de corpus para os exames de proficiência, diversos órgãos

responsáveis por exames internacionais (*British Council, Cambridge Assessment English, ETS, Pearson e Trinity College London*, para citar alguns) têm utilizado métodos da LC para aprimorar seus testes (Gablasova 2019). Mais recentemente, o uso das abordagens analíticas baseadas em corpus tem sido relatado em muitos livros e periódicos especializados em avaliação de línguas (Cushing 2017; Paquot 2018; Staples et al. 2018; Taylor e Baker 2008; Barker et al. 2015; Cushing 2022) e análises baseadas em corpus têm sido usadas em uma gama crescente de aplicações nas avaliações de proficiência, contribuindo para cada estágio do ciclo de desenvolvimento de um teste: a definição do seu construto, o desenvolvimento dos itens, a validação e a revisão (Cushing 2017; Green e Fulcher 2019). Especificamente sobre a distinção de níveis de proficiência, a análise de corpora tem servido para refinar a descrição de itens lexicais, gramaticais e discursivos utilizados em cada nível (Barkaoui et al. 2007; Kennedy e Thorp 2007; Read e Nation 2006, com corpus do IELTS; Biber e Gray 2013, com corpus do TOEFL, apenas para citar alguns).

A área de LC também tem experimentado um crescimento considerável no Brasil nos últimos anos, tanto nos estudos de Terminologia e Lexicografia quanto nos estudos de Linguística Aplicada, com foco na formação de professores e na elaboração de materiais didáticos de línguas adicionais (Sardinha 2004; Sarmento et al. 2014; Finatto et al. 2018). Apesar de a maioria dos estudos ainda estar focada na língua inglesa, pela ampla disponibilidade de corpora e ferramentas de pesquisa já desenvolvidas para essa língua, atualmente estudos sobre a descrição da língua portuguesa também vêm sendo realizados (Kuhn 2017; Sardinha 2017; Goulart 2020; Matte e Goulart 2021; Nagasawa e Schlatter 2021; Matte e Stumpf 2022, Nagasawa 2023; para citar alguns).

Além de estudos teóricos, já existem também disponíveis diversos corpora de português brasileiro, como o CETENFolha (Linguateca s.d.), com edições de 1994 da Folha de São Paulo, e o Corpus Brasileiro (Sardinha 2010), com cerca de um bilhão de palavras de textos organizados por gênero. O Corpus do Português (Davies e Ferreira 2006) inclui textos de vários países lusófonos e foi ampliado com os corpora Web/Dialects e NOW, cada um com um bilhão de palavras. O corpus Carolina é o mais recente, com 823 milhões de tokens de textos brasileiros de diversos gêneros coletados entre 1970 e 2021 (Crespo et al. 2023). Há ainda o *Portuguese Trends*, disponível no *Sketch Engine* (Kilgariff et al. 2004), composto principalmente por textos jornalísticos brasileiros e atualizado diariamente.

Nagasawa (2023) compilou o "*corpus* Celpe", contendo todos os textos de insumo, os títulos e os enunciados das tarefas 3 e 4 aplicadas no Celpe-Bras entre 1998 e 2020, totalizando 45 edições (90 tarefas). O corpus está disponível com acesso aberto em https://picoral.shinyapps.io/insumo-celpe/ e foi analisado utilizando métricas de complexidade textual selecionadas e validadas a partir das métricas disponíveis na ferramenta NILC-Metrix (Leal et al. 2024). Até a compilação do CorCel, este era o único corpus do exame Celpe-Bras já compilado[7].

Em Portugal, vários corpora de português como língua adicional em diversos níveis de proficiência já foram compilados a partir de textos escritos para o exame de proficiência em português elaborado e aplicado pelo Centro de Avaliação de Português Língua Estrangeira (CAPLE) ou escritos por estudantes em instituições de ensino portuguesas. O projeto "Recolha de Dados de Aprendizagem do Português como Língua Estrangeira", realizado pelo Instituto Camões e pelo Centro de Linguística da Universidade de Lisboa, reuniu 470 textos, classificados do nível A1 ao C1 segundo o Quadro Europeu Comum de Referência (QECR), escritos por 397 falantes de 28 línguas diferentes. O Corpus de Aquisição de L2 (CAL2), compilado pelo Centro de Linguística da Universidade NOVA de Lisboa (CLUNL), é organizado por língua materna e nível de proficiência (iniciante, intermediário e avançado). A seção escrita contém 1.607 textos de 1.184 informantes, totalizando 314.817 palavras. O Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) (Martins et al. 2019), compilado no CELGA-ILTEC da Universidade de Coimbra, reúne textos de 458 estudantes, falantes de 39 línguas diferentes, de níveis variados (A1 a C1) em cursos de português como língua adicional na Faculdade de Letras. O Corpus de Português como Língua Estrangeira/Língua Segunda (COPLE2) (Mendes et al., 2016) consiste em 966 textos produzidos em contextos de avaliação, incluindo exames administrados pelo CAPLE. Os textos são organizados em cinco níveis de proficiência do QECR (A1, A2, B1, B2 e C1) e foram produzidos por falantes de 14 línguas diferentes.

Dos corpora de português como língua adicional de que temos notícia, apenas o *Macaws – Multilingual Academic Corpus of Assignments – Writing and Speech* (Sommer-Farias et al. 2022) não foi compilado em Portugal. Trata-se de um projeto

[7] Apesar de serem de naturezas diferentes (o "*corpus* Celpe" é composto pelos textos de insumo, títulos e enunciados das tarefas do Celpe-Bras e o CorCel por textos produzidos por examinandos em resposta a essas tarefas), entendemos os dois corpora como complementares e possíveis de serem utilizados conjuntamente em pesquisas futuras.

da Universidade do Arizona e é composto por textos escritos por estudantes em resposta a diferentes tarefas nos cursos de português e russo da universidade. O subcorpus de português contém, até o momento, 2.294 textos e 545.763 palavras. Podemos perceber, assim, a importância e o ineditismo do CorCel, por ser o primeiro corpus compilado a partir de textos produzidos e avaliados no Celpe-Bras, e também pelo número de textos compilados, muito maior do que o dos demais corpora.

Além dos estudos de corpora, em parceria com colaboradores da área das ciências da computação (processamento da linguagem natural), muitas pesquisas têm alcançado progressos no desenvolvimento de programas que possibilitam a análise da construção coesiva de textos escritos com métricas próprias da língua portuguesa (Leal et al. 2024; Silva et al. 2021), com resultados para discussão sobre simplificação textual e sobre complexidade de leitura. Essas ferramentas possibilitaram as análises apresentadas em Nagasawa e Schlatter (2021), que analisaram a complexidade dos enunciados de tarefas de produção escrita do Celpe-Bras e do Enem, e em Nagasawa (2023), que discute a complexidade de leitura dos textos de insumo do Celpe-Bras.

Como já vimos, os estudos realizados com textos produzidos no Celpe-Bras ao longo dos últimos anos, apesar de já serem muitos, são, em sua grande maioria, de natureza qualitativa, visto a inexistência, até então, de um corpus de textos produzidos por examinandos. Para suprir essa lacuna e ampliar as possibilidades de pesquisa sobre o exame, decidiu-se compilar um corpus de estudo com dados de textos de quatro edições do Celpe-Bras, o CorCel, que estamos apresentando neste artigo. Os textos do CorCel já foram objeto de alguns estudos que buscaram ampliar e refinar a descrição dos níveis de proficiência avaliados no Celpe-Bras, tanto enfocando recursos gramaticais (Kunrath 2019; Soztrusnik 2023) quanto lexicais (Divino 2021; 2024; Hanauer 2023; Raupp 2024; Xavier 2025). Esses estudos, cujos resultados serão apresentados mais adiante, utilizaram ferramentas de LC como a geração de *word lists* (listas com as palavras mais frequentes), identificação de *keywords* (palavras mais frequentes em um corpus em comparação com outro) e *n-grams* (sequências contíguas de palavras) para analisar diferentes tarefas que compõem o corpus.

3. O CorCel

O Corpus Celpe-Bras (CorCel) é um projeto desenvolvido pelo grupo de pesquisa Avalia[8], que atua no Instituto de Letras da UFRGS, e visa compilar um corpus de mais de 70.000 textos produzidos em quatro edições do Celpe-Bras: 2015-2, 2016-1, 2016-2 e 2017-2[9]. O projeto CorCel está ancorado em uma concepção de linguagem de base discursiva, em que o uso da língua é compreendido como prática social situada. Essa perspectiva, amplamente inspirada em trabalhos bakhtinianos, reconhece que todo ato de linguagem se realiza por meio de gêneros do discurso, entendidos como tipos relativamente estáveis de enunciados que refletem e refratam práticas sociais específicas. No campo da avaliação, essa concepção implica considerar que os textos produzidos pelos examinandos em exames de proficiência não são apenas evidências de estruturas gramaticais internalizadas, mas manifestações concretas de sua capacidade de agir linguisticamente em contextos comunicativos diversos. Em outras palavras, a proficiência não se reduz à correção formal ou à complexidade linguística: envolve também a adequação ao gênero, ao propósito comunicativo e ao interlocutor. Isso se reflete na própria organização do corpus, que é feita em número de textos e não de palavras. Também é importante destacar que não está entre os objetivos do CorCel incentivar pesquisas relacionadas somente à análise de erros que, em uma visão tradicional, dizem respeito apenas à inadequação formal. Os textos avaliados nos diferentes níveis de proficiência no Celpe-Bras podem apresentar inadequações linguísticas; no entanto, não são apenas essas inadequações as responsáveis pelas notas atribuídas aos textos. De acordo com os parâmetros de avaliação do exame, embasados por uma orientação teórica discursiva e dialógica, as notas são atribuídas aos textos a partir da consistência da configuração da interlocução e do cumprimento do propósito solicitados na tarefa dentro do gênero discursivo proposto, de maneira que as formas estão subordinadas ao uso na interação proposta no enunciado da tarefa.

O CorCel se configura como um corpus de língua adicional, em que os textos compilados foram produzidos por examinandos do Celpe-Bras e avaliados por equipes de avaliação experientes nas avaliações do exame. Diferentemente de outros corpora produzidos em condições de exame, o CorCel não é considerado um "corpus de aprendizes" (*learner corpus*), nomenclatura que tradicionalmente engloba tais

[8] https://www.ufrgs.br/grupoavalia/

[9] Os dados do CorCel fazem parte do banco do Celpe-Bras, sob a responsabilidade do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Em 2017, a pesquisadora líder do grupo fez solicitação formal de dados ao Inep e recebeu autorização para utilizá-los para fins de pesquisa.

corpora e outros compilados em situação de ensino de línguas. Entendemos que o fato de terem nascido fora do Brasil e não serem falantes socializados em português desde a infância não justifica categorizar os examinandos do Celpe-Bras como "aprendizes". Nesse sentido, concordamos com Jenkins (2005) e com Divino (2024), entendendo que "atribuir o rótulo de *aprendiz* a qualquer estrangeiro é, ao nosso ver, concordar com a ideia de pureza linguística, que está atrelada à imagem de um falante nativo idealizado" (Divino 2024: 166). Em substituição ao termo "corpus de aprendizes", contemporaneamente outros autores propõem a nomenclatura "corpus de L2" (Gablasova, Brezina e McEnery 2017; Staples e Fernández 2019), "corpus de língua estrangeira" (Mauranen 2011; Meunier 2011) ou "corpus de língua franca" (Mauranen 2011). No CorCel, optamos por utilizar "corpus de língua adicional", assumindo que o português seja uma língua adicionada ao repertório do examinando e que, por razões pessoais, educacionais ou profissionais, lhe interessa certificar a proficiência no uso dessa língua.

Embora corpora com textos produzidos por examinandos não sejam novidade para línguas como o inglês, até onde sabemos, o CorCel é o primeiro grande corpus com textos escritos por falantes de português como língua adicional em que os textos foram produzidos em resposta às mesmas tarefas e avaliados duas vezes por examinadores capacitados, seguindo o sistema de avaliação do Celpe-Bras. As análises possibilitadas por um corpus como esse podem mostrar características linguísticas recorrentes em diferentes níveis de proficiência e podem ser usadas para desenvolver ferramentas para auxiliar professores e estudantes de português como língua adicional.

3.1. Dados

O CorCel é constituído por textos de quatro edições do Celpe-Bras, compreendendo 16 tarefas inseridas em 6 esferas de atuação, relacionadas a 10 temáticas, que solicitam a produção de textos de 9 diferentes gêneros do discurso com 7 diferentes propósitos, a serem publicados em 8 diferentes suportes[10].

Como mencionado anteriormente, o projeto CorCel visa compilar mais de 70.000 textos produzidos em resposta às tarefas escritas do Celpe-Bras em quatro edições (2015-2, 2016-1, 2016-2 e 2017-1). Cada edição contém quatro tarefas e, para

[10] A descrição completa das tarefas que compõem o CorCel, bem como de todas as tarefas já aplicadas no Celpe-Bras, está disponível em www.ufrgs.br/grupoavalia.

cada tarefa, existem seis subcorpora de textos avaliados com notas 0, 1, 2, 3, 4 e 5 por dois avaliadores (além da nota da reavaliação, quando foi o caso). Até o momento, o corpus já digitado é composto por 15.315 textos digitados e revisados[11], com até 200 textos avaliados em cada nota (0 a 5) para cada tarefa (quatro por edição).[12]

3.2. Metadados

Conforme já mencionado, o CorCel se constitui como um corpus de português como língua adicional, não um corpus de aprendizes. Não há metadados disponíveis sobre os examinandos que produziram os textos, como é comum em corpora de aprendizes, visto que os textos foram recebidos sem identificação. Não há também identificação do posto aplicador onde o exame foi realizado nem tampouco informações sobre os avaliadores. A falta de informações que caracterizem os participantes do exame impede, portanto, a formulação de quaisquer hipóteses relacionadas a especificidades como o tempo de estudo da língua portuguesa, o número de vezes que o examinando realizou o exame ou a língua materna dos examinandos.

Há, no entanto, informações relativas às tarefas que geraram esses textos. A descrição de todas as tarefas aplicadas no Celpe-Bras, feita por Schoffen et al. (2018), identifica o material de insumo da tarefa (áudio, vídeo ou texto escrito), seu tema, a esfera de atividade na qual o texto solicitado está inserido, o(s) principal(is) propósito(s) comunicativo(s), os interlocutores, o gênero discursivo e o suporte em que esse texto seria publicado. O Quadro 1 mostra a análise realizada para as tarefas que compõem o CorCel.

Edição/ Tarefa	Temática	Esfera de Atuação	Propósito	Gênero do Discurso	Suporte
2015-2 T1	transporte	político-cidadã	incentivar	seção de guia	guia
2015-2 T2	educação	científico-educacional	incentivar	notícia	mural
2015-2 T3	políticas e cidadania	político-cidadã	solicitar	carta/e-mail	carta/e-mail

[11] Como o processo de digitação de todos os 70.000 textos seria bastante demorado, decidimos, inicialmente, compilar o corpus com um número menor de textos para iniciar as análises. O corpus será aumentado gradativamente à medida que mais textos forem digitados e revisados.
[12] Os textos já digitados do CorCel totalizam cerca de três milhões de palavras.

2015-2 T4	patrimônio cultural	político-cidadã	solicitar	carta aberta	jornal
2016-1 T1	educação	científico-educacional	incentivar	depoimento	site
2016-1 T2	ciência e tecnologia	profissional	indicar	carta/e-mail	carta/e-mail
2016-1 T3	mundo do trabalho	profissional	sugerir	relatório	relatório
2016-1 T4	estilos de vida	jornalística	posicionar-se	artigo de opinião	revista
2016-2 T1	patrimônio cultural	científico-educacional	divulgar	notícia	site
2016-2 T2	mundo do trabalho	profissional	orientar	carta/e-mail	carta/e-mail
2016-2 T3	políticas e cidadania	comunitária	incentivar	artigo	blog
2016-2 T4	consumo	jornalística	posicionar-se	carta do leitor	jornal
2017-1 T1	turismo	jornalística	divulgar	notícia	jornal
2017-1 T2	ciência e tecnologia	científico-educacional	sugerir	carta/e-mail	carta/e-mail
2017-1 T3	políticas e cidadania	político-cidadã	solicitar	carta/e-mail	carta/e-mail
2017-1 T4	estilos de vida	jornalística	posicionar-se	carta do leitor	jornal

Quadro 1: Metadados do CorCel referentes às tarefas
Fonte: elaborado pelas autoras com base nas informações disponíveis em www.ufrgs.br/grupoavalia

Além dos metadados relacionados às tarefas, o CorCel dispõe também de todas as notas recebidas por cada texto. Estão disponíveis também as notas obtidas pelo examinando que escreveu cada texto em cada uma das outras tarefas da edição, suas notas na parte oral do Celpe-Bras e o nível de certificação recebido. Essas informações permitem analisar os perfis de proficiência dos examinandos em cada tarefa e comparar e contrastar esses perfis entre as diferentes tarefas.

3.3. Desenho do corpus

A seleção dos 15.315 textos que compõem a primeira versão do CorCel seguiu critérios específicos para garantir a representatividade e a diversidade do corpus. Foram incluídos textos produzidos nas diferentes tarefas e edições do Celpe-Bras a que tivemos acesso, e representativos das diferentes notas atribuídas. Essas escolhas visaram contemplar tanto a variedade temática e de gêneros dos textos quanto a

amplitude de desempenhos linguísticos representados, de modo a permitir análises comparativas.

Para compilar o corpus, selecionamos inicialmente textos que receberam a mesma nota dos dois avaliadores, sem a necessidade de reavaliação por um terceiro avaliador. Com o objetivo de compilar um número semelhante de textos em cada subcorpora, 200 textos foram selecionados aleatoriamente de cada nota e de cada tarefa de cada edição. Nos casos em que havia menos de 200 textos com duas notas concordantes, o número foi completado com textos que haviam sido reavaliados, usando-se randomização para a seleção desses textos. Quando alguma das notas não apresentou um número de textos igual ou superior a 200, todos os textos classificados com essa nota foram incluídos no corpus, o que significa que, nesses casos, o subcorpus referido é composto por um número menor de textos. O corpus compilado até o momento é apresentado na Tabela 1. Cada coluna exibe o número de textos compilados em cada nota por tarefa e por edição. A coluna mais à direita mostra o número total de textos compilados por tarefa e por edição.

Edição/Nota	0	1	2	3	4	5	Total
2015-2							**3.996**
T1	59	128	200	200	200	193	926
T2	82	200	200	200	200	200	1082
T3	33	189	200	200	200	138	960
T4	28	200	200	200	200	200	1.028
2016-1							**4.085**
T1	15	200	200	200	200	200	1015
T2	48	200	200	200	200	200	1048
T3	44	200	200	200	200	151	995
T4	93	200	200	200	200	134	1.027
2016-2							**3.704**
T1	21	188	200	200	200	159	968
T2	22	130	200	200	200	73	825
T3	30	143	200	200	200	200	973
T4	43	200	200	200	200	95	938
2017-1							**3.530**
T1	4	54	156	200	200	200	814
T2	19	119	200	200	200	200	938
T3	7	73	200	200	200	135	815
T4	8	155	200	200	200	200	963

Tabela 1 - Número de textos por tarefa e edição

Fonte: elaborado pelas autoras

3.4. Preparação dos dados

A prova do Celpe-Bras é realizada em papel, de forma manuscrita, e os textos foram fornecidos pelo Inep em arquivos digitalizados. Para cada examinando, identificado por um código numérico sequencial de 1 até o número total de examinandos inscritos em cada edição, foram disponibilizadas imagens digitalizadas de suas produções escritas e as respectivas notas. As notas correspondem à nota de cada uma das quatro tarefas da parte escrita (incluindo a nota de dois ou, em caso de discrepâncias, três avaliadores), bem como à nota final da parte escrita. Esses dados foram organizados em uma planilha do Excel específica para cada edição do exame, com links para as imagens digitalizadas dos textos manuscritos, arquivadas em formato .tif.

O processo de digitação dos textos segue as diretrizes desenvolvidas pelo grupo Avalia, segundo as quais o primeiro passo foi digitar os textos em arquivos .txt. A identificação do documento em que cada texto é digitado segue a ordem ano, edição, tarefa, participante e nota, sem espaço entre as informações, como no exemplo 20152t4p903n3. Esse código indica que o texto foi produzido em 2015, na segunda edição, em resposta à Tarefa 4, pelo participante número 903 e avaliado com nota 3. Após passar pelo processo de revisão, que consiste em comparar o conteúdo do texto digitado com a imagem original, um "r" é adicionado ao final do código, tornando-o 20152t4p903n3r.

Embora os textos tenham sido fornecidos sem qualquer informação sobre os examinandos, optamos por anonimizar todos os nomes próprios escritos nos textos que não estivessem presentes no texto de insumo, substituindo-os por "Bibiana Cambará"[13]. Para evitar que essa alteração de nome modificasse o número de palavras no corpus, as diretrizes incluíram um total de três nomes (Bibiana Arminda Flora) e três sobrenomes (Cambará Terra Amaral), a serem utilizados de acordo com o número de nomes e sobrenomes no texto original, sem acréscimos. Embora haja variação nos nomes, essa ordem é sempre respeitada para facilitar a contagem. Além disso, também existem instruções para a ocorrência de apelidos e siglas, buscando manter a digitação o mais próxima possível do texto original. Além de evitar que

[13] O nome foi escolhido em homenagem a uma importante personagem da literatura gaúcha, presente na obra O Tempo e o Vento, de Érico Veríssimo.

qualquer nome que o examinando possa ter escrito permaneça no documento por razões éticas e de privacidade, a padronização preserva informações importantes para a análise, permitindo, por exemplo, a verificação do número de textos assinados.

Os textos digitados seguem o layout do texto original da forma mais fiel possível, mantendo as entradas de parágrafos, a diferença entre letras maiúsculas e minúsculas e a sinalização de possíveis elementos gráficos, como desenhos. As rasuras não são consideradas para a versão digitada, que mantém apenas as palavras que não estão riscadas. Em casos de palavras ilegíveis, as diretrizes orientam a comparação das letras com as demais presentes no texto. Quando não é possível identificar, cada letra ilegível é substituída pelo símbolo §.

3.5. Normalização

Diferentemente de muitos corpora de aprendizes, que contam com sistemas detalhados de anotação de erros, nosso objetivo no projeto é fazer uma normalização ortográfica mínima, de forma a permitir a contagem de palavras e outras análises sem distorções. O estudo de Hanauer (2022) mostrou diferenças significativas em análises dos mesmos textos realizadas com e sem a padronização ortográfica realizada de acordo com a norma dicionarizada da língua portuguesa. A proporção entre *types* e *tokens* (*type-token ratio* ou TTR), uma medida amplamente utilizada para calcular diversidade lexical, diminui significativamente quando se analisa os textos normalizados visto que, com a normalização, diversas realizações ortográficas de uma mesma palavra (como “casa” e “caza” para “casa”) contam como apenas um *type*. Além disso, Divino (2021) também demostrou que as diferenças ortográficas dificultam a busca por trechos reproduzidos do texto de insumo, visto que as ferramentas utilizadas para a análise não conseguem recuperar *ngrams* quando as palavras não estão escritas da mesma forma Para o processo de normalização, foi desenvolvido um protocolo especial que tem por objetivo permitir análises mais precisas e também identificar a cópia de trechos dos textos de insumo, e não contar e classificar tais escritas como erros. A versão final do CorCel contará, então, com dois arquivos do mesmo texto: a versão original e a versão normalizada, permitindo assim aos pesquisadores realizarem análises qualitativas e quantitativas de acordo com seus interesses de pesquisa.

Entretanto, a realização manual deste processo, tendo em vista a quantidade de textos do corpus, é uma tarefa que demanda muito tempo e equipe especializada. Para facilitar este processo, foi desenvolvida, em parceria com o Instituto de Informática da UFRGS, uma ferramenta semiautomática de normalização, intitulada CorSpell (Schoffen et al. em preparação). A partir de uma interface intuitiva, os pesquisadores responsáveis pela tarefa de normalização podem selecionar textos (filtrando pela identificação, nota e usuário responsável) previamente inseridos em um banco de dados. A ferramenta utiliza dois dicionários para fazer a análise de cada token e sugerir formas alternativas para palavras não reconhecidas. O usuário pode ainda inserir uma outra palavra, caso as sugestões não contemplem a palavra-alvo, e também selecionar palavras que não foram identificadas com problemas de ortografia pelos dicionários existentes na ferramenta. Uma vez finalizada a normalização de um texto, a ferramenta pode exportar um arquivo em formato .txt que utiliza sintaxe XML para inserir a nova forma com o uso de *tags* (<>), permitindo seu uso em ferramentas de linguística de corpus como *AntConc* (Anthony 2024) e *Sketch Engine* (Kilgariff et al. 2004). Ao final do período de testes e implementação de todas as funcionalidades, a ferramenta será disponibilizada gratuitamente e poderá ser customizada de forma a utilizar outros dicionários e listas de palavras a depender das necessidades dos usuários.

Hanauer (em preparação) já utilizou a ferramenta para normalizar os corpora das diferentes notas da tarefa 1 da edição 2016-2. Como resultado, verificou-se que houve uma consistente diminuição no TTR nos subcorpora de todas as notas. No subcorpus de nota 5, o TTR passou de 12,83 para 11,50. No subcorpus de nota 2, de 17,03 para 13,22. A normalização vai permitir análises que reflitam de forma mais precisa o uso de recursos lexicais dos examinandos, bem como a recuperação de *n-grams* comuns a vários textos, o que viabiliza estudos sobre cópia e paráfrases nas diferentes notas.

4. Estudos já realizados sobre o CorCel

Alguns dos textos que compõem o CorCel já foram analisados em estudos que utilizaram metodologias qualitativas e quantitativas (conforme resumido em Stumpf et al. 2025). Kunrath (2019) analisou, com o auxílio do software *Coh-Metrix*, 50 textos das tarefas 1 e 4 da edição 2016-1 para distinguir os que receberam nota 3 dos

que receberam nota 5, e propôs uma progressão de níveis baseada na recontextualização das informações do texto de insumo e no uso de recursos linguísticos e discursivos específicos. Mendel (2019) descreveu a integração das habilidades nas tarefas de leitura e escrita a partir da análise qualitativa de 236 textos de todas as notas das tarefas 3 e 4 da edição 2015-2. Os resultados indicam que os textos que recebem notas mais altas demonstram maior uso de conhecimentos e de repertório linguístico prévio do seu autor, além da recontextualização das informações do texto de insumo ficar mais completa à medida que as notas aumentam. Textos com notas mais altas também demonstram melhor adequação ao gênero e configuração mais consistente da relação de interlocução proposta na tarefa. Tais características vão sendo executadas nos textos de forma mais adequada à medida que o nível de proficiência aumenta, configurando assim textos com mais autoria, especialmente nos textos de nota 5. Já o trabalho de Sirianni (2020) (ver o artigo de Sirianni nesta edição) analisou qualitativamente 120 textos que receberam notas 1 e 2 nas quatro tarefas da edição 2016-2, a fim de descrever diferenças entre textos não certificados e textos de nível intermediário. Os resultados mostram que as principais diferenças entre essas notas são: a) compreensão do texto de insumo, b) configuração da relação de interlocução no gênero solicitado e c) número de inadequações linguísticas e discursivas apresentadas.

Alguns estudos já utilizaram subcorpora do CorCel para identificar índices lexicais relevantes para descrever as características dos textos que receberam notas diferentes em diferentes tarefas do Celpe-Bras (Divino 2021; 2024; Hanauer 2023; Soztrunik 2023; Raupp 2024; Xavier 2025). As análises, realizadas com o *Sketch Engine* e o teste de significância estatística *Log-Likelihood* (LL) (Rayson 2002), indicaram que os textos avaliados com nota 5 são consistentemente mais longos, apresentando número maior de palavras e sentenças em relação aos textos avaliados com nota 2, sugerindo que a extensão do texto é um fator importante para determinar o nível de proficiência. Também foi possível verificar que textos mais avançados apresentam maior diversidade lexical, ou seja, eles contêm mais palavras diferentes das do material de insumo, apesar de serem morfologicamente relacionadas (por exemplo, o uso do substantivo “conservação” no material de insumo e do verbo “conservar” nos textos) (Divino 2024; Raupp 2024; Xavier 2025). Os textos de nota 5 também apresentam maior número de conjunções em relação aos textos avaliados com nota 2 (Soztrusnik 2023), o que sugere, além de maior

conhecimento lexical da língua portuguesa, maior explicitação da articulação coesiva nos textos de níveis mais avançados. A lista de frequência de palavras demonstrou ainda que os textos de nota 5 também apresentam mais palavras típicas dos gêneros discursivos solicitados (Divino 2021; 2024; Hanauer 2023; Xavier 2025) e configuram a interlocução proposta na tarefa de forma mais direta.

Os dados também colocam em xeque a noção de que a cópia do material de insumo ocorre apenas nos níveis mais baixos, mostrando que longos trechos do texto de insumo e/ou do enunciado, contendo pelo menos seis itens em sequência, são encontrados nos corpora de todas as notas (Divino 2024; Divino e Schoffen, 2025). Raupp (2024) e Divino (2024) verificaram ainda que textos de notas mais altas tendem a parafrasear de forma mais completa as informações essenciais do texto de insumo, ao passo que textos avaliados com notas mais baixas trazem trechos mais fragmentados e insuficientes para cumprir plenamente a tarefa, resultado que reforça a análise de Sirianni (2020). Os resultados mostram também que uma tarefa pode ser levada a cabo de diferentes formas, visto que há termos relacionados ao cumprimento dos propósitos da tarefa que não estão presentes em 100% dos textos, mesmo nos de nota 5 (Divino 2024). Essas análises quantitativas também corroboraram as análises qualitativas realizadas anteriormente, mostrando uma maior incidência de estruturas características do gênero do discurso solicitado (Mendel, 2019), juntamente com termos mais adequados à relação de interlocução proposta (Sirianni, 2016).

A Tabela 2 apresenta os dados dos textos do CorCel que já foram utilizados nos estudos de Divino (2021, 2024), Hanauer (2023), Sostrusznik (2023), Raupp (2024), Xavier (2025) e Hanauer (em preparação). A tabela mostra o número de textos já analisados de cada nota em cada tarefa, seguido por *Types*, *Tokens* e TTR em cada subcorpus. Embora o número de textos analisados varie entre os estudos, o que altera o resultado do TTR, já que essa medida é dependente do tamanho do corpus, os resultados aqui apresentados são um exemplo dos dados que pretendemos disponibilizar de todo o corpus em um futuro próximo.

Edição	**Tarefa**	**Nota**	**N_textos**	**Types**	**Tokens**	**TTR**
2015-2	**T3**	5	138	3,355	31,480	10,66%
		2	200	4,313	34,123	12,64%

	T4[14]	5	237	5,663	47,616	11.91%
		4	477	7,247	88,235	8.21%
		3	715	8,838	123,033	7.19%
		2	628	8,537	99,457	8.58%
		1	211	4,004	30,290	13.35%
		0	25	826	2,284	-[15]
2016-2	**T1**	5	100	3087	24059	12,83
		4	100	2916	22750	12,82
		3	100	3164	21086	15,01
		2	100	2963	17402	17,03
		1	100	3058	14795	20,67
		0	24	1061	2413	-
	T3	5	200	4,823	45,848	10.51%
		2	200	4,083	31,479	12.97%
	T4	5	50	2,641	12,222	21.60%
		2	50	1,845	7,891	23.38%
2017-1	**T3**	5	100	2,876	23,748	12.11%
		2	100	2,660	16,593	16.03%
	T4	5	50	2,626	11,650	22.54%
		2	50	2,023	8,376	24.15%

Tabela 2 - Proporção entre Types e Tokens por tarefa e nota
Fonte: Elaborado pelas autoras.

Os dados apresentados na Tabela 2 foram todos calculados antes da normalização dos textos. A partir do uso da plataforma *CorSpell* para a normalização do corpus, será possível disponibilizar os resultados das descrições tanto nos textos originais quanto nos textos normalizados.

[14] No estudo de Divino (2024), foi utilizado o total de textos disponíveis, não apenas os selecionados para compor a primeira versão do corpus. Por esse motivo, o número de textos analisados é maior do que o apresentado na Tabela 1.
[15] Devido ao pequeno número de textos do subcorpus nota 0, o cálculo do TTR não foi realizado. Nesse cálculo, o tamanho do corpus tem uma interferência direta e o resultado não seria confiável.

5. Contribuições possíveis do CorCel para o exame Celpe-Bras e para a área de PLA

A análise do processo de compilação do CorCel e das características dos textos que o compõem permite discutir tanto os aspectos linguísticos quanto os recursos discursivos utilizados na produção escrita dos textos produzidos e avaliados com diferentes notas no exame Celpe-Bras. O corpus foi estruturado para possibilitar pesquisas que descrevam como examinandos de diferentes níveis de proficiência mobilizam recursos linguístico-discursivos ao realizar tarefas que exigem a configuração da interlocução para o cumprimento de propósito(s) comunicativo(s) determinados, revelando assim a construção da autoria no texto. Também será possível descrever como textos que receberam a mesma nota acionam recursos diferentes para realizar a tarefa, permitindo refletir sobre a diversidade de usos da língua dentro do mesmo nível de proficiência. Essas perspectivas de pesquisa reforçam a concepção de que não existe uma única possibilidade de usar a língua para configurar a interlocução dentro de determinado gênero do discurso, e que cada texto é único e deve ser avaliado em sua singularidade (Schoffen 2009; Inep 2020).

As tarefas do exame solicitam a produção de diferentes gêneros discursivos – como cartas, relatórios, artigos e cartas do leitor. Essa diversidade reforça o potencial do corpus para fomentar estudos que analisem o uso de diferentes recursos linguístico-discursivos em diferentes gêneros, permitindo investigações sobre como os examinandos interpretam os propósitos comunicativos das tarefas, de que modo ajustam o estilo, o vocabulário e a estrutura textual e como constroem a autoria nos textos. Além disso, o corpus possibilita estudos comparativos sobre o uso de recursos linguísticos (lexicais, morfossintáticos e discursivos) em diferentes níveis de proficiência (e também dentro do mesmo nível). Esses dados podem fundamentar descrições mais precisas do desenvolvimento da proficiência escrita em português como língua adicional e oferecer subsídios empíricos para a revisão de parâmetros de avaliação e descritores de proficiência.

Outra contribuição relevante do CorCel para o ensino e a formação de professores é permitir a análise, a partir de textos autênticos, das relações entre tarefa, gênero e desempenho dos examinandos. Assim, entendemos que os resultados das análises do CorCel têm potencial para serem utilizados em cursos de formação de

professores de português como língua adicional, em pesquisas sobre ensino e avaliação de escrita e na elaboração de materiais didáticos baseados em textos autênticos de diferentes níveis de proficiência.

O CorCel tem potencial para contribuir ainda para a validação empírica do construto do Celpe-Bras, pois oferece evidências observáveis de como o desempenho dos examinandos se configura nos diferentes níveis de proficiência e de que forma as tarefas integradas de compreensão e produção efetivamente avaliam as habilidades comunicativas pretendidas. Essas análises fortalecem a compreensão do exame como instrumento de avaliação coerente com sua fundamentação teórica, e consolidam o vínculo entre pesquisa, ensino e avaliação.

6. Perspectivas para o futuro

A compilação do CorCel amplia possibilidades de pesquisa, que certamente contribuirão significativamente para o aprimoramento da descrição dos níveis de proficiência do Celpe-Bras, além de impactar o ensino de PLA e a formação de professores e pesquisadores da área. Utilizando os dados do CorCel, está em desenvolvimento uma plataforma automática de avaliação de textos, a ser disponibilizada gratuitamente no domínio da UFRGS. Essa plataforma tem por objetivo auxiliar professores e estudantes na preparação para o Celpe-Bras, realizando avaliação automatizada de textos, provendo feedback e sugestões imediatas e personalizadas, identificando aspectos do texto a serem melhorados e sugerindo estratégias de estudo adaptativas, com base no desempenho individual do autor do texto. Em uma segunda fase de implementação, será possível desenvolver na plataforma outras funcionalidades, como gerar relatórios detalhados dos usos da língua para estudantes e professores, criar exercícios interativos e atividades que respondam de maneira dinâmica às áreas de dificuldade e aos progressos observados.

Além disso, pretende-se integrar à plataforma ferramentas de análise que permitam consultas aos dados do CorCel realizadas com apoio de inteligência artificial, tornando a exploração de dados acessível mesmo para quem não domina ferramentas complexas de Linguística de Corpus. Ferramentas tradicionais de análise de corpus têm grande potencial pedagógico, especialmente em abordagens como a aprendizagem guiada por dados (*data-driven learning*), mas ainda enfrentam barreiras de usabilidade para professores e alunos. A plataforma contará com

funcionalidades específicas que possibilitarão a visualização de linhas de concordância, listas de palavras e *n-grams*, entre outras, feitas em linguagem natural. Somando todas essas funcionalidades, a plataforma busca contribuir para o acesso democratizado a dados empíricos que podem fomentar pesquisas em aprendizagem e avaliação de proficiência de português como língua adicional.

O CorCel abre um vasto leque de possibilidades para o avanço das pesquisas também Processamento de Linguagem Natural (PLN) em português. Considerando que o português ainda é uma língua com recursos limitados para tratamento automático (Penteado e Perez 2023), o CorCel representa uma base empírica inédita para o desenvolvimento e o aprimoramento de modelos de linguagem de grande escala (LLMs), possibilitando a criação de ferramentas de avaliação automatizada de textos (*Automated Essay Scoring* – AES) e de avaliação automatizada de escrita (*Automated Writing Evaluation* – AWE) ajustadas às especificidades do português como língua adicional e às características do Celpe-Bras.

Além de subsidiar pesquisas que objetivem investigar a validade de construto do Celpe-Bras, futuras pesquisas podem explorar o potencial do CorCel para o ajuste fino de modelos de linguagem com vistas à classificação de textos por nível de proficiência, à extração de características linguísticas relevantes e à identificação de padrões discursivos e lexicais característicos de cada nível. Tal exploração está em consonância com estudos recentes (Yancey et al. 2023; Mizumoto e Eguchi 2023; Pack et al. 2024), que destacam a importância de investigações interdisciplinares envolvendo linguística, computação e educação para melhorar a acurácia e a validade dos modelos preditivos. O CorCel, por conter textos autênticos de diferentes gêneros, propósitos comunicativos e contextos de produção, oferece uma base ideal para esses ajustes, contribuindo para o desenvolvimento de sistemas mais robustos e sensíveis à diversidade textual do português como língua adicional.

Outra vertente promissora de estudos possíveis refere-se à integração de ferramentas de inteligência artificial (IA) generativa e *feedback* automatizado em ambientes de preparação para o Celpe-Bras. A partir do CorCel, é possível treinar modelos capazes de oferecer orientações imediatas e personalizadas aos estudantes, simulando a avaliação realizada no exame e auxiliando o desenvolvimento da proficiência escrita. Estudos internacionais (Braz e Chenoll 2024; Wang 2022; Ayotunde, Jamil e Cavus 2023) evidenciam que sistemas de feedback automatizado baseados em IA podem melhorar o desempenho dos aprendizes, ao permitir a

correção autônoma de inadequações linguísticas e discursivas e o monitoramento do progresso individual. Essas pesquisas podem ainda contribuir para o avanço da personalização do ensino de português como língua adicional, alinhando-se às demandas pedagógicas de individualização das práticas de ensino.

O CorCel também se constitui como um espaço fértil para o debate sobre questões éticas, metodológicas e de validade no uso de IA em contextos avaliativos. Conforme exposto por Hao et al. (2024), a precisão, a explicabilidade, a generalização e a falta de consistência dos modelos podem afetar o quanto os resultados dos testes avaliados com IA generativa, e LLMs em particular, realmente medem o que se propõem, limitando seu uso em avaliações de alto impacto. Estudos recentes (Pack et al. 2024; Kohnke, Moorhouse e Zou 2023) alertam para a necessidade de mitigar vieses algorítmicos e em relação aos dados de treino, garantir a transparência dos modelos e preservar a natureza interpretativa e contextual da avaliação humana. Assim, além de servir como base para o aprimoramento técnico de modelos de linguagem, o CorCel pode fomentar reflexões críticas sobre a interface entre tecnologia, linguagem e avaliação, fortalecendo o papel das pesquisas acadêmicas para a produção de conhecimento ético, socialmente engajado e tecnologicamente inovador.

7. Considerações finais

O desenvolvimento do CorCel representa um avanço significativo tanto para os estudos sobre o Celpe-Bras quanto para os campos de ensino e avaliação de português como língua adicional, Linguística de Corpus e de Processamento da Linguagem Natural. O corpus oferece uma base empírica inédita para análises sobre a escrita em contextos avaliativos autênticos e sobre o modo como falantes de diferentes níveis de proficiência mobilizam a língua portuguesa para agir discursivamente.

Ao reunir textos autênticos produzidos e avaliados no Celpe-Bras, o CorCel se diferencia de outros corpora disponíveis por compilar textos classificados segundo níveis de desempenho reconhecidos institucionalmente. Isso permite não apenas a realização de estudos linguísticos e discursivos, mas também a validação empírica dos descritores de proficiência e dos parâmetros de avaliação utilizados no exame. Ainda que os textos passem por uma normalização ortográfica mínima, guiada por

um protocolo elaborado pela equipe para prevenir os problemas mais comuns de análise automática, os dados originais são mantidos, permitindo pesquisas que comparem as duas possibilidades ou que queiram trabalhar apenas com os textos não normalizados. Mesmo automatizada com a ferramenta desenvolvida para este fim, a normalização é um processo que altera o texto ao propor uma outra forma para substituir palavras que não são compreendidas, o que depende fortemente do repertório da pessoa responsável pela tarefa, tanto na compreensão de formas não-padrão quanto na sugestão de outra forma próxima, porém reconhecida e validada por determinada comunidade linguística.

Do ponto de vista teórico, o CorCel reforça a concepção de linguagem como ação social mediada por gêneros do discurso, mostrando que o desempenho linguístico não pode ser dissociado das práticas comunicativas que o constituem. É importante ressaltar que as diferenças entre o desempenho materializado nos textos avaliados com diferentes notas não são tratadas como déficits, mas como indícios do desenvolvimento progressivo da proficiência em língua portuguesa, o que se alinha à concepção de linguagem subjacente ao exame, que reconhece a interdependência entre a forma dos recursos linguísticos e seu uso situado em contexto. Essa perspectiva aproxima a análise linguística da realidade de uso e contribui para repensar o ensino e a avaliação de português como língua adicional a partir de uma abordagem mais situada e interacional.

No plano metodológico, o CorCel aponta para a relevância de integrar princípios da linguística de corpus e outras tecnologias, como IA generativa, à pesquisa em avaliação e ensino de línguas. A possibilidade de explorar dados autênticos e sistematizados amplia o diálogo interdisciplinar e oferece uma base sólida para o desenvolvimento de práticas pedagógicas baseadas em evidências.

Ao possibilitar análises qualitativas e quantitativas sobre textos de diferentes níveis de proficiência, o CorCel contribui para a democratização do conhecimento e para a consolidação de uma cultura de pesquisa baseada em dados empíricos. Além disso, espera-se que o CorCel estimule novas pesquisas sobre o desenvolvimento da proficiência escrita em português, a relação entre a configuração da interlocução nos diferentes níveis de proficiência e como essa configuração ocorre nos diferentes gêneros discursivos, as implicações pedagógicas da preparação para o exame Celpe-Bras, entre várias outras temáticas. Essas possibilidades reforçam o compromisso do Grupo Avalia com a produção de conhecimento socialmente

relevante e comprometida com a educação de professores e de estudantes de português em contextos multilíngues e multiculturais.

Agradecimentos

Parte das reflexões apresentadas neste artigo é fruto do projeto de pesquisa intitulado "Desenvolvimento de ferramentas de inteligência artificial a partir de modelos de linguagem de grande escala para a descrição de proficiência linguística e elaboração de recursos pedagógicos em português como língua adicional", financiado pelo CNPq (Chamada CNPq/MCTI/FNDCT No 22/2024).

Referências bibliográficas

ALDERSON, J. Charles; WALL, Dianne. Does Washback Exist? *Applied Linguistics*, v. 14, n. 2, p. 115-129, 1 Jun. 1993. DOI: http://dx.doi.org/10.1093/applin/14.2.115. Disponível em: https://academic.oup.com/applij/article-abstract/14/2/115/224706. Acesso em: 4 jul. 2025.

ALDERSON, J. Charles. Do corpora have a role in language assessment? In: THOMAS, J.; SHORT, M. (ed.). *Using corpora for language research:* Studies in the honour of Geoffrey Leech. London: Longman, p. 248-259, 1996.

ANTHONY, Lawrence. *AntConc* versão 4.1.1 [software]. Tokyo: Waseda University, 2020. Disponível em: https://www.laurenceanthony.net/software/. Acesso em: 10 nov. 2025.

AYOTUNDE, Oke O.; JAMIL, Dashty I.; CAVUS, Nadire. The impact of artificial intelligence in foreign language learning using learning management systems: A systematic literature review. *Information Technologies and Learning Tools*, v. 95, n. 3, p. 215, 2023.

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

BANERJEE, Jayanti; FRANCESCHINA, Florencia; SMITH, Anne M. Documenting Features of Written Language Production Typical at Different IELTS Score Band Levels. British Council, Cambridge English Language Assessment and IDP: IELTS Australia. *IELTS Research Reports Online Series*, v. 7, 2007. Disponível em: https://ielts.org/researchers/our-research/research-reports/documenting-features-of-written-language-production-typical-at-different-ielts-band-score-levels. Acesso em: 4 jul. 2025.

BARKAOUI, Khaled. What Changes and What Doesn't? An Examination of Changes in the Linguistic Characteristics of IELTS Repeaters' Writing Task 2 Scripts. *IELTS Research Report Series*, n. 3, p. 1-55, 2016.

BARKER, Fiona; SALAMOURA, Angeliki; SAVILLE, Nick. Learner Corpora and Language Testing. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny. (org.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University, 2015. p. 511-534.

BIBER, Douglas; GRAY, Bethany. Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT® Test: A Lexico‐Grammatical Analysis. *ETS Research Report Series*, i-128, 2013.

BORTOLINI, Letícia S. *Os conceitos de uso de língua, identidade e aprendizagem subjacentes ao material didático para o ensino de português em Letícia (Colômbia)*. Trabalho de Conclusão de Curso (Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, 2006.

BRAZ, Ana; CHENOLL, Antonio. O processo de avaliação num contexto online na era da Inteligência Artificial: um duplo desafio. *Re@D - Revista de Educação A Distância e Elearning,* [S.L.], p. 1-20, 8 jul. 2024. http://dx.doi.org/10.34627/REDVOL7ISS1E202412.

CHAPELLE, Carol A.; PLAKANS, Lia. Assessment and Testing: Overview. In: CHAPELLE, Carol. A (Ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell/Wiley, 2013. p. 240-244.

CHAPELLE, Carol. A. Conceptions of Validity. In: FULCHER, G.; DAVIDSON, F. (ed.). *The Routledge Handbook of Language Testing*. London: Routledge, 2012. p. 21-33.

CHENG, Liying; SULTANA, Nasreen. Washback: Looking Backward and Forward. In: FULCHER, Glenn; HARDING, Luke. (ed.). *The Routledge Handbook of Language Testing*. 2nd ed. London: Routledge, 2022. Chap. 8. p. 136-152.

COSTA, Éverton V. da. *Práticas de formação de professores de português língua adicional em um instituto cultural brasileiro no exterior*. 2013. 162 f. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

CRESPO, Maria C. R. M.; ROCHA, Maria L. S. J.; STURZENEKER, Mariana L.; SERRAS, Felipe R.; MELLO, Guilherme L. de; COSTA, Aline S.; PALMA, Mayara F.; MESQUITA, Renata M.; GUETS, Raquel de P.; SILVA, Mariana M; da; FINGER, Marcelo; SOUSA, Maria Clara P. de; NAMIUTI, Cristiane; MONTE, Vanessa M. do. Carolina: a general corpus of contemporary Brazilian Portuguese with provenance, typology and versioning information. *arXiv preprint arXiv:2303.16098*, 2023.

CUMMING, Alister; KANTOR, Robert; BABA, Kyoko; ERDOSY, Usman; EOUANZOUI, Keanre; JAMES, Mark. Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, [S.L.], v. 10, n. 1, p. 5-43, jan. 2005.

CUSHING, Sara T. Corpus Linguistics and Language Testing. In: FULCHER, Glenn; HARDING, Luke (ed.). *The Routledge Handbook of Language Testing*. 2. ed. London: Routledge, 2022. Chap. 32. p. 345-360.

CUSHING, Sara T. Corpus linguistics in language testing research. *Language Testing*, [S.L.], v. 34, n. 4, p. 441-449, 19 set. 2017. http://dx.doi.org/10.1177/0265532217713044.

DAVIES, Mark; FERREIRA, Michael. Corpus do Português. [S. l.], 2006. Disponível em: https://www.corpusdoportugues.org/. Acesso em: 29 ago. 2025.

DIVINO, Luiza S.; SCHOFFEN, Juliana 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 Linguistica Aplicada*, [S. l.], v. 24, n. 2, p. DOI: 10.26512/rhla.v24i2.56513. Disponível em: https://periodicos.unb.br/index.php/horizontesla/article/view/56513. Acesso em: 11 nov. 2025.

DIVINO, Luiza S. *Contribuições da linguística de corpus para a definição de níveis de proficiência escrita no exame Celpe-Bras*. 2024. Dissertação (Mestrado em Letras) - Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024.

DIVINO, Luiza S. *Índices lexicais de análise para a caracterização dos níveis intermediário e avançado superior no exame Celpe-Bras*: uma pesquisa guiada por corpus. 2021. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

DORIGON, Thomas. *O Celpe-Bras como Instrumento de Política Linguística:* um Mediador entre Propósitos e Materializações. Dissertação (Mestrado em Linguística Aplicada) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

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

FINATTO, Maria José B.; REBECHI, Rozane; SARMENTO, Simone; BOCORNY, Ana Eliza P. *Linguística de Corpus:* perspectivas. Porto Alegre: Instituto de Letras - UFRGS, 2018.

FORTES, Melissa S. *Uma compreensão etnometodológica do trabalho de fazer ser membro na fala-em-interação de entrevista de proficiência oral em português como língua adicional*. 2009. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

FULCHER, Glenn; HARDING, Luke. Epilogue: Language Testing: Where Are We Heading? In: FULCHER, Glenn; HARDING, Luke (ed.). *The Routledge Handbook of Language Testing*. 2. ed. London: Routledge, 2022. Chap. 32. p. 345-360.

FULCHER, Glenn. Does Thick Description Lead to Smart Tests? A Data-Based Approach to Rating Scale Construction. *Language Testing*, v. 13, n. 2, p. 208-238, 1996.

GABLASOVA, Dana; BREZINA, Vaclac; MCENERY, Tony. Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information. *Language Learning*, v. 67, n. 1, p. 130-154, 2017. DOI: http://dx.doi.org/10.1111/lang.12226.

GABLASOVA, Dana. Corpora for Second Language Assessments. In: WINKE, Paula; BRUNFAUT, Tineke. *The Routledge Handbook of Second Language Acquisition and Language Testing*. London: Routledge, 2021. p. 45-53.

GEBRIL, Atta; PLAKANS, Lia. Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, v. 10, n. 1, p. 9-27, 2013.

GOMES, Maíra S. *A complexidade de tarefas de leitura e produção escrita no exame Celpe-Bras*. 2009. 109p. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, 2009.

GOULART, Larissa. Formulaic sequences and writing development in Portuguese as a Second Language. *Spanish And Portuguese Review*, [S. L.], v. 6, n. 1, p. 79-107, 2020.

GREEN, Anthony; FULCHER, Glenn. Test Design Cycle. In: WINKE, Paula; BRUNFAUT, Tineke. (ed.) *The Routledge handbook of second language acquisition and language testing*. New York: Routledge, 2020. p. 69-77.

GUO, Liang. *Product and Process in Toefl iBT Independent and Integrated Writing Tasks:* A Validation Study. Tese (Doutorado em Filosofia) – Georgia State University, Atlanta, 2011.

HAO, Jiangang; VON DAVIER, Alina A.; YANEVA, Victoria; LOTTRIDGE, Susan; VON DAVIER, Matthias; HARRIS, Deborah. J. Transforming Assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, v. 43, n. 2, p. 16–29, jun. 2024.

HANAUER, Isadora 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. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

HANAUER, Isadora D. *Influência das inadequações ortográficas em análise de tarefa escrita do Celpe-Bras guiada por corpus*. XXXIV Salão de Iniciação Científica, UFRGS. 2022. Disponível em: https://youtu.be/C6iomOTWRwM. Acesso em: 4 jul. 2025.

HANAUER, Isadora D. *Tarefas de compreensão oral para produção escrita do Celpe-Bras:* uma análise de produção de examinandos guiada por corpus. [manuscrito em preparação]

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA INEP. *Documento-base do exame Celpe-Bras*. Brasília, DF: Instituto

Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2020. 130 p. Disponível em: https://www.ufrgs.br/acervocelpebras/wp-content/uploads/2021/12/Documento-base-do-exame-Celpe-Bras-2020.pdf. Acesso em: 5 jul. 2025.

JENKINS, Jennifer. ELF at the Gate: The Position of English as a Lingua Franca. *Humanising Language Teaching*, Oxford, v. 7, n. 2, Mar. 2005. Disponível em: http://old.hltmag.co.uk/mar05/idea.htm#C2. Acesso em: 4 jul. 2025.

KENNEDY, Christopher; THORP, Dilys. A Corpus-Based Investigation of Linguistic Responses to an IELTS Academic Writing Task. In: TAYLOR, Lynda; FALVEY, Peter. (ed.). *IELTS Collected Papers*: Research in Speaking and Writing Assessment. Studies in Language Testing, v. 19. Cambridge: Cambridge University Press, 2007. p. 316-378.

KILGARRIFF, Adam; RYCHLÝ, Pavel; SMRŽ, Pavel; TUGWELL, David. The Sketch Engine. *In*: EURALEX INTERNATIONAL CONGRESS, 11, 2004, Université de Bretagne-Sud. *Proceedings* [...]. Leiden: Euralex, 2004. p. 105-116. Disponível em: https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf. Acesso em: 4 jul. 2025.

KOHNKE, Lucas.; MOORHOUSE, Benjamin L.; ZOU, Di. ChatGPT for Language Teaching and Learning. *RELC Journal*, v. 54, n. 2, p. 537-550, 2023. DOI: https://doi.org/10.1177/00336882231162868.

KUHN, Tanara Z. A *design proposal of an online corpus-driven dictionary of Portuguese for University Students*. 2017. 421 f. Tese (Doutorado em Letras). Universidade de Lisboa, Lisboa, 2017.

KUNRATH, Simone P. *Os descritores gerais e a progressão dos níveis de proficiência do exame Celpe-Bras*. 2019. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

LEAL, Sidney E.; DURAN, Magali S.; SCARTON, Caroline E.; HARTMANN, Nathan S.; ALUÍSIO, Sandra M. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, v. 58, n. 1, p. 73-110, 2024.

LI, Ye. *A preparação de candidatos chineses para o Exame Celpe-Bras*: aprendendo o que significa "uso da linguagem". 2009. 130p. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

LIMA, Ronaldo A. *Representações do Brasil em textos do Exame para obtenção do Certificado de Língua Portuguesa para Estrangeiros*. 2008. 139f. Tese (Doutorado em Letras) – Universidade Federal Fluminense, Niterói, 2008.

LINGUATECA. *CETENFolha*: Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo. Lisboa: Linguateca, s.d. Disponível em: https://www.linguateca.pt/cetenfolha/index_info.html. Acesso em: 10 nov. 2025.

MARTINS, Alexandre F. *A la recherche d'un lieu épistémologique brésilien en Portugais Langue Additionnelle:* regards croisés sur les discours académiques dans une perspective décoloniale. 2022. Tese (Doutorado em École Doctorale 58 Langues, Littératures, Cultures et Civilisations) - Université Paul-Valéry, 2022.

MARTINS, C.; FERREIRA, T.; SITOE, M.; ABRANTES, C.; JANSSEN, M.; FERNANDES, A.; SILVA, A.; LOPES, I.; PEREIRA, I.; SANTOS, J. *Corpus de produções escritas de aprendentes de PL2 (PEAPL2)*: subcorpus Português Língua Estrangeira. Coimbra: CELGA-ILTEC, 2019.

MATTE, Marine L.; GOULART, Larissa. Pacotes Lexicais e níveis de proficiência em Português como Segunda Língua: Uma investigação da função de pacotes lexicais. *Letras de Hoje*, [S. l.], v. 55, n. 4, p. e38377, 2020. DOI: 10.15448/1984-7726.2020.4.38377. Disponível em: https://revistaseletronicas.pucrs.br/fale/article/view/38377. Acesso em: 11 nov. 2025.

MATTE, Marine L.; STUMPF, Elisa. M. A corpus-based study of reporting verbs in academic Portuguese. *Research in Corpus Linguistics*, [S. l.], v. 10, n. 2, p. 46-69, 2022. http://dx.doi.org/10.32714/ricl.10.02.04.

MAURANEN, Anna. Learners and Users – Who Do We Want Corpus Data From? *In*: MEUNIER, Fanny; DE COCK, Sylvie; GILQUIN, Gaëtanelle; PAQUOT, Magali. *.A Taste for Corpora*: In Honour of Sylviane Granger. Amsterdam: John Benjamins, 2011. p. 155-171.

MENDEL, Kaiane. *Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras*. 2019. 177p. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

MENDEL, Kaiane. A avaliação integrada de leitura e escrita no exame CELPE-BRAS. 2017. 62p. Trabalho de Conclusão de Curso (Licenciatura em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017.

MENDES, Amália; ANTUNES, Sandra; JANSSEN, Maarten; GONÇALVES, Anabela. The COPLE2 Corpus: A Learner Corpus for Portuguese. In: *LREC*, 10., 2016. *Proceedings* [...]. Portorož: European Language Resources Association (ELRA) May 2016. p. 3207-3214. Disponível em: https://aclanthology.org/L16-1511/. Acesso em: 4 July 2025.

MEUNIER, Fanny. Corpus Linguistics and Second/Foreign Language Learning: Exploring Multiple Paths. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 11, n. 2, p. 459-477, 2011. Disponível em: https://www.scielo.br/j/rbla/a/BLJ6xy89SLRH7KNYSzJTPjS/?lang=en. Acesso em: 4 July 2025.

MINISTÉRIO DAS RELAÇÕES EXTERIORES. *Proposta curricular para o ensino de português nas unidades da rede de ensino do Itamaraty em países de língua oficial espanhola*. Brasília, DF: FUNAG, 2020a.

MINISTÉRIO DAS RELAÇÕES EXTERIORES. *Proposta curricular para o ensino de português nas unidades da rede de ensino do Itamaraty em países de língua oficial portuguesa*. Brasília, DF: FUNAG, 2020b.

MITTELSTADT, Daniela D. *Orientações curriculares e pedagógicas para o nível avançado de português como língua adicional*. 2013. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

MIZUMOTO, Atsushi, EGUCHI, Masaki. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics*, v. 2, n. 2, 100050, 2023. DOI: https://doi.org/10.1016/j.rmal.2023.100050.

NAGASAWA, Ellen Y. *Elaboração e análise de sequência didática de leitura e produção textual para preparação ao Exame Celpe-Bras*. 2016. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

NAGASAWA, Ellen Y. *Português como língua adicional para fins específicos*: preparação ao exame Celpe-Bras. 2018. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

NAGASAWA, Ellen Y. O *conteúdo de insumo em tarefas que integram leitura e escrita no Celpe-Bras*: uma abordagem informada por corpus. 2023. 302p. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

NAGASAWA, Ellen Y.; SCHLATTER, Margarete. A complexidade de enunciados de tarefas de produção escrita no Celpe-Bras e no Enem. *Tradterm*, São Paulo, v. 37, n. 1, p. 330–363, 2021. DOI: 10.11606/issn.2317-9511.v37p330-363. Disponível em: https://revistas.usp.br/tradterm/article/view/169414. Acesso em: 10 nov. 2025.

OHLWEILER, Beatriz Maria D. *Criação de um jornal na sala de aula de português língua estrangeira*. 2006. Dissertação (Mestrado em Letras – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

PACK, Austin; BARRETT, Alex; ESCALANTE, Juan. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, v. 6, p. 100234, 2024.

PAQUOT, Magali. Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, v. 15, n. 1, p. 29-43, 2018.

PAQUOT, Magali. The Phraseological Dimension in Interlanguage Complexity Research. *Second Language Research*, California, v. 35, n. 1, p. 121-145, Jan. 2019. DOI: https://doi.org/10.1177/0267658317694221.

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

RAUPP, Amanda. *Características lexicais das produções escritas do exame CELPE-BRAS na tarefa 3 de 2016-2*: uma pesquisa guiada por corpus. 2024. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2024.

RAYSON, Paul E. *Matrix*: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison. 2002. Tese (Doutorado em Ciência da Computação) – Universidade de Lancaster, Lancaster, 2002.

READ, John; NATION, Paul. An Investigation of the Lexical Dimension of the IELTS Speaking Test. *IELTS Research Reports*, v. 6, p. 207-231, 2006.

SARDINHA, Tony B.; MOREIRA FILHO, José. L.; ALAMBERT, Eliane. CORPUS Brasileiro: Corpus of Brazilian Portuguese. Sketch Engine, Pontifícia Universidade Católica de São Paulo, Apr. 2010. Disponível em: https://www.sketchengine.eu/corpus-brasileiro/. Acesso em: 29 ago. 2025.

SARDINHA, Tony B. *Linguística de Corpus*. Barueri, SP: Manole, 2004.

SARDINHA, Tony B. Text types in Brazilian Portuguese: A multidimensional perspective. *Corpora*, v. 12, n. 3, p. 483-515, 2017.

SARMENTO, Simone; IBAÑOS, Ana M. T.; MOTTIN, Lívia. P.; BERBER SARDINHA, T. *Pesquisa e perspectivas em linguística de corpus*. São Paulo: Mercado de Letras, 2014.

SCARAMUCCI, Matilde V. R. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. *Trabalhos em Linguística Aplicada*, v. 43, n. 2, p. 203-226, 2004. DOI: http://dx.doi.org/10.1590/s0103-18132004000200002. Disponível em: https://www.scielo.br/j/tla/a/PpG6gN9pbJVhVbhWB6JWZkw/. Acesso em: 4 jul. 2025.

SCARAMUCCI, Matilde V. R. Validade e consequências sociais das avaliações em contextos de ensino de línguas. *Linguarum Arena*, Porto, v. 2, p. 103-120, 2011. Disponível em: https://ler.letras.up.pt/uploads/ficheiros/9836.pdf. Acesso em: 4 jul. 2025.

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. do *et al.* (org.). *A redação no contexto do vestibular 2005*: a avaliação em perspectiva. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2005. p. 11-35.

SCHLATTER, Margarete; SCARAMUCCI, Matilde V. R.; PRATI, Silvia; ACUÑA, Leonor. Celpe-Bras e Celu: impactos da construção de parâmetros comuns de avaliação de proficiência em português e em espanhol. In: FONTANA, Monica. G. Z.

O português do Brasil como língua transnacional. Campinas: Editora Rg, 2009. p. 95-122.

SCHOFFEN, Juliana R.; STUMPF, Elisa M.; AMARAL, Deise; DIVINO, Luiza S.; HANAUER, Isadora D.; LISBOA, Isabel; RAUPP, Amanda; XAVIER, Brenda. Compilation and Tagging of a Corpus with Celpe-Bras Texts. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 16., 2024, Santiago de Compostela. *Proceedings* [...]. [S. L.]: Association for Computational Linguistics, 2024. p. 627-632.

SCHOFFEN, Juliana R.; BALREIRA, Dennis G.; STUMPF, Elisa M., GOULART, Larissa; KUHN, Tanara Z.; PAZZINATO, Gabriel R.; HANAUER, Isadora D.; SILVA, José Henrique S.; DIVINO, Luiza S.; MATTE, Marine L. *CorSpell: introducing a semiautomatic tool for spelling normalization in Brazilian Portuguese*. [manuscrito em preparação]

SCHOFFEN, Juliana R.; MARTINS, Alexandre. F. Políticas linguísticas e definição de parâmetros para o ensino de português como língua adicional: perspectivas portuguesa e brasileira. *ReVEL*, Novo Hamburgo, v. 14, n. 26, p. 271-306, Mar. 2016.

SCHOFFEN, Juliana R.; SCHLATTER, Margarete; KUNRATH, Simone P.; NAGASAWA, Ellen Y.; SIRIANNI, Gabrielle R.; MENDEL, Kaiane; TRUYLLIO, Luana R.; DIVINO, Luiza S. *Estudo descritivo das tarefas da Parte Escrita do exame Celpe-Bras*: edições de 1998 a 2017. Porto Alegre: Instituto de Letras, UFRGS, 2018. Disponível em: https://lume.ufrgs.br/handle/10183/195625. Acesso em: 4 jul 2025.

SCHOFFEN, Juliana R. *Avaliação de proficiência oral em língua estrangeira*: descrição dos níveis de candidatos falantes de espanhol no exame Celpe-Bras. 2003. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

SCHOFFEN, Juliana R. *Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no Exame Celpe-Bras*. 2009. 192p. Tese (Doutorado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SHOHAMY, Elana. *Language Policy*: Hidden Agendas and New Approaches. London; New York: Routledge, 2006.

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

SILVA, Asaf. D. C.; DELGADO, Heloísa O. K.; FINATTO, Maria José B. Acessibilidade textual e terminológica para o português brasileiro: pesquisa, estratégias e orientações de [re]escrita simplificada. *Moara – Revista Eletrônica do Programa de Pós-Graduação em Letras,* [S.L.], n. 58, p. 322, 31 jul. 2021. http://dx.doi.org/10.18542/moara.v0i58.10903.

SIRIANNI, Gabrielle 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. 76p. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

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

SOMMER-FARIAS, Bruna; NOVIKOV, Aleksey; PICORAL, Adriana; CENTANIN-BERTHO, Mariana; STAPLES, Shelley. A multilingual learner corpus for less commonly taught languages. *International Journal of Learner Corpus Research*, v. 8, n. 2, p. 261-282, 2022.

SOSTRUZNIK, Julia L. *O uso de conjunções em produções escritas no exame Celpe-Bras*: um estudo baseado em corpus. 2023. 76p. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

SOUZA NETO, Maurício J. *CELPE-BRAS e CAPLE: A proficiência em português como língua não materna em paralaxe*. 2018. 123 f. Dissertação (Mestrado em Língua e Cultura) – Programa de Pós-Graduação em Língua e Cultura, Instituto de Letras, Universidade Federal da Bahia, Salvador, 2018.

STAPLES, Shelley; BIBER, Douglas; REPPEN, Randi. Using corpus‐based register analysis to explore the authenticity of high‐stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal*, v. 102, n. 2, p. 310-332, 2018.

STAPLES, Shelley; FERNÁNDEZ, Julieta. Corpus Linguistics Approaches to L2 Pragmatics Research. In: TAGUCHI, Naoko. (ed.). *The Routledge Handbook of Second Language Acquisition and Pragmatics*. London: Routledge, 2019. p. 241-254.

STUMPF, Elisa M.; SCHOFFEN, Juliana R.; DIVINO, Luiza S.; HANAUER, Isadora D.; RAUPP, Amanda; XAVIER, Brenda. Corpus-driven lexical analyses of CorCel: a comparative analysis of preliminary findings of written proficiency in Portuguese as an additional language. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 2025, Fortaleza/CE. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 673-681. DOI: https://doi.org/10.5753/stil.2025.37870.

TAYLOR, Lynda; BARKER, Fiona. Using Corpora for Language Assessment. *In*: HORNBERGER, Nancy. H. (ed.). *Encyclopaedia of Language and Education*: Language Testing and Assessment, vol. 7. New York: Springer, 2008. p. 241-254.

TAYLOR, Lynda. Washback and Impact. *ELT Journal*, v. 59, n. 2, p. 154-155, 1 Apr. 2005. Oxford University Press (OUP). DOI: http://dx.doi.org/10.1093/eltj/cci030.
The Routledge handbook of second language acquisition and language testing. New

TOSATTI, Natália M. *O desempenho de estudantes de Países Africanos de Língua Oficial Portuguesa no Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras)*. 2021. 260p. Tese (Doutorado em Letras) – Universidade Federal de Minas Gerais, Belo Horizonte, 2021.

VICENTINI, Mônica P. *As dimensões do construto compreensão oral para produção escrita no Exame Celpe-Bras*: percepções, processos, estratégias e desempenhos em examinandos. 2022. 241p. Tese (Doutorado em Letras) – Universidade Estadual de Campinas, Campinas, 2022. Disponível em: https://repositorio.unicamp.br/acervo/detalhe/1253389. Acesso em: 4 July 2025.

WALL, Dianne. Washback. *In*: FULCHER, GlennG.; DAVIDSON, Fred. *The Routledge Handbook of Language Testing*. London: Routledge, 2012. p. 79-92.

WALTERS, F. Scott. Ethics and Fairness. *In*: FULCHER, Glenn; HARDING, Luke (ed.). *The Routledge Handbook of Language Testing*. 2. ed. London: Routledge, 2022. p. 345-360.

WANG, Zhijie. Computer-assisted EFL writing and evaluations based on artificial intelligence: a case from a college reading and writing course. *Library Hi Tech*, v. 40, n. 1, p. 80-97, 2022.

WEIGLE, Sara C. *Assessing Writing*. Cambridge: Cambridge University Press, 2002. 282p.

XAVIER, Brenda R. *Entre o material de insumo e o repertório:* caracterização dos textos produzidos em resposta à tarefa 3 da edição 2015/2 do exame Celpe-Bras à luz da linguística de corpus. 2025. 93p. Trabalho de Conclusão de Curso (Licenciatura em Letras) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2025.

YAN, Qiaorong. *De práticas sociais a gêneros do discurso*: uma proposta para o ensino de português para falantes de outras línguas. 2008. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

YANCEY, Kevin P.; LAFLAIR, Geoffrey; VERARDI, Anthony; BURSTEIN, Jill. Rating short L2 essays on the CEFR scale with GPT-4. In: 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA 2023). p. 576-584. Disponível em: https://aclanthology.org/2023.bea-1.49.pdf York: Routledge, 2020.