Corpus TecEM: o processo de construção de um *corpus* de produções textuais em Língua Portuguesa elaboradas por alunos de Ensino Médio Integrado a cursos técnicos

Maitê Moraes Gil[1]
Julia Ferri Pinto[2]
Vitor Gouvêa[3]
Bruno Corrêa de Almeida[4]
Pedro de Andrade Santos[5]
Augusto Weiand[6]

Resumo

O presente artigo visa a apresentar o processo de construção do *Corpus* TecEM, uma ferramenta que disponibiliza textos escritos por alunos de cursos técnicos integrados ao Ensino Médio em suas aulas de Língua Portuguesa. A construção de um *corpus* com essas características se apresenta fundamental devido às potencialidades acadêmicas e ao desenvolvimento de práticas docentes que sua exploração pode embasar. Como base teórica, parte-se da compreensão da Linguística de *Corpus* como uma abordagem baseada em *corpus*, uma perspectiva para o estudo da linguagem. Para tanto, foi necessário: (a) levantamento bibliográfico; (b) contato com professores de Língua Portuguesa de Institutos Federais (IFs) e convite aos interessados em contribuir na coleta de textos; (c) coleta de textos escritos por alunos de cursos técnicos integrados ao Ensino Médio durante suas aulas de Língua Portuguesa; (d) compilação dos textos a partir dos critérios estabelecidos; (e) armazenamento do *corpus* TecEM em um banco de dados online, definindo sua estrutura de maneira alinhada aos critérios de compilação; (f) desenvolvimento e disponibilização da ferramenta. Em março de 2019, o *Corpus* TecEM possuía 226 textos (111,800 palavras) escritos por alunos dos IFs localizados no estado do Rio Grande do Sul e mais textos já foram coletados a partir de abril do mesmo ano. Com a construção deste *corpus*, entende-se que foi disponibilizada à comunidade acadêmica uma base de dados rica e criteriosa para futuras pesquisas, contribuindo tanto para o desenvolvimento teórico quanto para a proposição de novas práticas de ensino de Língua Portuguesa, de modo geral, e em contextos de formação tecnológica, em particular.

Palavras-Chave: *corpus* de Língua Portuguesa. Ensino de língua portuguesa. Formação tecnológica.

Abstract

This paper aims to describe the construction of TecEM *Corpus*, an online tool that makes publicly available texts written by high school and vocational course students during their Brazilian Portuguese (BP) classes. In addition to its potential for academic use, such a corpus can also support the development of teaching practices. As a theoretical basis, this project assumed the understanding of *Corpus* Linguistics as a *corpus*-based approach and a perspective for the study of language. These six steps were followed: (a) bibliographical survey; (b) contact with BP teachers from Federal Institutes and invitation to those interested in contributing

[1] Doutora em Letras. Professora do IFRS – campus Osório e investigadora de pós-doutoramento no Centro de Estudos Filosóficos e Humanísticos da Universidade Católica Portuguesa/Braga.
[2] Aluna do curso de Letras do IFRS – campus Osório.
[3] Aluno do curso de Letras do IFRS – campus Osório.
[4] Aluno do curso de Ciências da Computação na UFRGS.
[5] Aluno do curso de Ciências da Computação na UFSM.
[6] Doutorando em Informática na UFRGS.

to the *corpus*; (c) collection of texts written by high school and vocational course students during their BP classes; (d) compilation of texts based on established criteria; (e) storing the TecEM *corpus* in an online database, defining its structure aligned with the compilation criteria; (f) development and availability of the tool. In March, 2019, TecEM *Corpus* had 226 texts (111,800 words) written by students of Federal Institutes located in the Brazilian state of Rio Grande do Sul and more texts were already collected from April onward. With the construction of this *corpus*, we argue that a rich and careful database for future research was made available to the academic community, contributing both to theoretical development and to the proposition of new BP teaching practices, in general, and to technological education contexts, in particular.

Keywords: Brazilian Portuguese *corpus.* Brazilian Portuguese teaching. Technological education.

1 Primeiras palavras: o que é o *Corpus* TecEM?

A produção textual é hoje o cerne do ensino de língua portuguesa no Brasil. Nos anos 1980, as discussões do Círculo de Bakhtin chegaram ao país, sendo estudadas e tomadas como referência por alguns linguistas brasileiros, como é o caso de Carlos Alberto Faraco (2009) e João Wanderley Geraldi (1997). Desde o ponto de vista bakhtiniano (2006[1929]), a língua é uma interação entre sujeitos, que se constituem por meio dessa interação. Como aponta Marchioro (2010), com a amplitude das relações sociais, bem como com a grande variedade de textos, é de considerável importância que o ensino de Língua Portuguesa gire em torno do texto e busque maior eficácia na ampliação das competências linguísticas, textuais e comunicativas dos alunos. Nessa mesma época, como destaca Marchioro, "surge a expressão 'produção de textos'. E assim, muda-se o foco do ensino: a língua deixou de ser vista apenas como estrutura e voltou-se para o contexto de produção e recepção dos textos" (2010, p.11). Foi apenas nos Parâmetros Curriculares Nacionais do ano de 1998, no entanto, que o texto se consolidou, ao menos nos documentos norteadores, como centro do ensino da língua portuguesa no Brasil. Nesta mesma década, as teorias de Análise do Discurso, Sociolinguística e Linguística Textual começaram a ser utilizadas no processo aprendizagem da língua materna (MARCHIORO, 2010).

Diante da centralidade do texto nas aulas de Língua Portuguesa, muitos professores gostariam de ter acesso às produções textuais dos alunos para práticas de ensino e pesquisa. No entanto, alguns fatores dificultam tal ação, dois deles são: o armazenamento dos textos e a usabilidade das informações. Onde armazenar as produções textuais? Como

encontrar nos textos elementos a serem analisados? Como identificar as dificuldades mais comuns? Essas são algumas questões que surgem juntamente à vontade de guardar os textos para trabalhos posteriores.

Em face a essas interrogações, a Linguística de *Corpus* se apresenta como uma alternativa, uma vez que, como destacam Tagnin e Fromm (2010), ela "há já algum tempo, dedica-se a esse tipo de pesquisa e numerosas coletâneas de textos em formato eletrônico foram compiladas para os mais variados objetivos" (s/p). Dentre os diferentes tipos possíveis de *corpus*, aqueles dedicados a textos de aprendizes são os relacionados a este estudo, embora o *corpus* aqui proposto seja mais adequadamente caracterizado como um "*corpus* de estudantes". Localizamos, no Brasil, ao menos dois *corpora* desse tipo: o CoMAprend – *Corpus* de Aprendizes (TAGNIN, 2006) e o BELT - *Brazilian English Leaner Corpus* (PACHECO, 2010). Ambos são coletâneas de textos em língua estrangeira elaborados por estudantes brasileiros. Não foram localizados, entretanto, *corpora* disponíveis *online* cujos textos sejam exclusivamente produções de alunos brasileiros de Ensino Médio em Língua Portuguesa.

A proposta deste estudo é ainda mais específica: a construção do *Corpus* TecEM, um *corpus* de textos escritos por alunos de cursos técnicos integrados ao Ensino Médio. Justifica-se o interesse em produções de alunos de cursos técnicos integrados ao Ensino Médio com o fato de este ser um espaço recente de estudos, o qual se fortaleceu com a expansão dos institutos federais. Além da compilação dos textos, buscamos a disponibilização *online* desse *corpus* e ferramentas de consulta às produções, a fim de que professores e/ou pesquisadores de qualquer região possam ter acesso a esse material de maneira funcional. Como ressaltam Aluísio e Almeida (2006), "entende-se que disponibilização de *corpus* compilado para futuras pesquisas é uma característica inerente ao *corpus*, [...] uma vez que se tem uma referência padrão de língua ou de variedade de língua que pode ser utilizada por outros pesquisadores" (p.158).

A construção e a disponibilização do *corpus* aqui apresentado se justificam, portanto, pela centralidade que as produções textuais têm nas aulas de língua portuguesa e pelas possibilidades de estudo e de aprimoramento da prática docente que elas oferecem a partir de um estudo sistemático. Desta maneira, este estudo apresenta tanto contribuições

de ordem técnica e científica, por meio da disponibilização de um *corpus* de aprendiz inexistente no contexto nacional, quanto – a longo prazo – de ordem social, visto que são possíveis impactos positivos nas práticas de ensino de língua materna a partir das investigações conduzidas no *corpus* aqui apresentado.

2 Bases Teóricas

A Linguística de *Corpus* (LdC) é uma abordagem empírica para o estudo da língua que se fortaleceu nas últimas décadas. Embora o primeiro grande *corpus* eletrônico linguístico tenha sido lançado na década de 1960, tal abordagem ganhou espaço no Brasil bem mais recentemente", com o avanço das tecnologias digitais. Para Berber Sardinha (2000), a "história da Linguística de *Corpus* está, portanto, intimamente ligada à disponibilidade de corpora eletrônicos" (p. 329).

Há uma questão frequente sobre o status da LdC: trata-se de uma disciplina ou de uma metodologia? Por um lado, o fato de o objeto da LdC não ser delimitado como em outras áreas aponta para o entendimento de que a LdC não é uma disciplina a exemplo da Psicolinguística ou da Sociolinguística, uma vez que a LdC se ocupa de fenômenos geralmente abordados por outras áreas, como léxico, sintaxe, semântica e morfologia. Por outro lado, não é possível compreender a LdC apenas como uma metodologia *instrumental*, visto que ela leva à produção de novos conhecimentos, sendo mais que um instrumental computacional. Diante disso, surge uma terceira alternativa: a compreensão da LdC como uma abordagem baseada em *corpus*, uma perspectiva para o estudo da linguagem (BERBER SARDINHA, 2004). A proposta deste estudo está alinhada a este último entendimento, pois, ao invés de investigar o que é teoricamente possível na língua, tem como foco a investigação do uso e da maneira como os estudantes/usuários utilizam os recursos de linguagem disponíveis (MOTTIN, 2012).

Com a consolidação da LdC, o termo *corpus*, inicialmente utilizado para designar um conjunto de dados sobre um determinado tema, adquiriu um novo sentido, visto que o próprio conceito de *corpus* precisou ser redelimitado. Compreende-se que "os textos em um *corpus* são selecionados de acordo com critérios explícitos a fim de serem utilizados

como uma amostra representativa de determinada língua ou de um subconjunto daquela língua" (BOWKER; PEARSON, 2002, p.10). O *corpus* é, portanto, um artefato produzido para pesquisa. Para Berber Sardinha (2004), deve-se considerar, ao menos, os seis pontos a seguir para a definição de um *corpus* alinhada à LdC:

> (a) A origem: Os dados devem ser autênticos
> (b) O propósito: O *corpus* deve ter a finalidade de ser um objeto de estudo linguístico
> (c) A composição: O conteúdo do *corpus* deve ser criteriosamente escolhido
> (d) A formatação: Os dados do *corpus* devem ser legíveis por computador
> (e) A representatividade: O *corpus* deve ser representativo de uma língua ou variedade
> (f) A extensão: O *corpus* deve ser vasto para ser representativo. (p. 18-19)

Associados a esses pontos, o autor destaca quatro pré-requisitos para a formação de um *corpus* computadorizado: (i) os textos que compõem o *corpus* devem ser autênticos e em linguagem natural (não podem ser produzidos especificamente para comporem o *corpus* e nem em linguagem artificial); (ii) os textos precisam ser produzidos por falantes nativos (a exceção de corpora de aprendizes de língua estrangeira); (iii) o conteúdo do *corpus* deve ser criteriosamente escolhido; e (iv) o *corpus* deve conter um conjunto de textos representativos de uma variedade linguística. A questão da representatividade, como destaca Sarmento (2010), envolve ainda conhecer o "todo" que, no caso da linguagem, não é conhecido. Deve-se, por isso, "tentar dividir esse todo estimado em partes" (p. 90). Conforme ilustra a autora, um *corpus* de "linguagem jornalística", por exemplo, deve incluir diferentes tipos de jornais, textos das diferentes seções, um número aproximado de palavras em cada categoria, etc. Quando esses aspectos são considerados, a combinação do uso de ferramentas computacionais com os dados de *corpora* pode gerar resultados quantitativos e qualitativos confiáveis com potencial para revelar fenômenos desconhecidos sobre a língua.

Em suma, na concepção da LdC, um *corpus* é "uma grande coleção de textos autênticos, que foram reunidos em forma eletrônica de acordo com um conjunto específico de critérios" (BOWKER; PEARSON, 2002, p.20). Tais critérios dependem da natureza e da proposta de cada projeto. O objetivo principal de um *corpus* é "servir como referência do que é típico na língua, sendo assim utilizado em pesquisas linguísticas. Através do

distanciamento de exemplos artificiais, o uso da LdC confere plausibilidade às pesquisas linguísticas de natureza quantitativa e qualitativa de descrição da língua" (MOTTIN, 2012, p.18).

Segundo Kennedy (1998), existem quatro concentrações principais nos estudos em LdC, são elas: (i) compilação de *corpus*; (ii) desenvolvimento de ferramentas; (iii) descrição da linguagem; e (iv) aplicação de corpora (para o ensino de línguas, tradução, etc.). A primeira parte do nosso estudo, a qual é descrita neste artigo, enquadra-se na primeira linha e, as etapas subsequentes, na quarta. A fim de que se possa realizar um trabalho de investigação mais detalhado acerca da produção escrita em língua materna de alunos de cursos Técnicos Integrados ao Ensino Médio, faz-se necessária a compilação de um *corpus* de textos específicos desses estudantes.

Como são muitos os objetos relacionados à LdC, há uma nomenclatura extensa empregada na definição do conteúdo e do propósito dos *corpora*. Os principais tipos, no entanto, podem ser agrupados nos seguintes critérios: modo (falado x escrito), tempo (sincrônico x diacrônico; contemporâneo x histórico), seleção (de amostragem x monitor; dinâmico x estático; equilibrado), conteúdo (especializado x regional x multilíngue), autoria (de aprendiz x de língua nativa) e finalidade (de estudo x de referência x de treinamento).

Considerando os critérios mencionados, o *corpus* desenvolvido por este estudo e aqui apresentado atende à tipologia: escrito, sincrônico, contemporâneo, de amostragem, estático, equilibrado, especializado, de língua nativa e de estudo.

Conforme destacado anteriormente, um *corpus* é um repositório de textos digitais. Desta forma, a fim de que seu conteúdo seja acessado, é necessário que haja recursos ou ferramentas específicas. Os *corpora* maiores e mais consolidados geralmente possuem seus próprios recursos ou ferramentas de acesso, já outros *corpora* necessitam ser armazenados e acessados através de programas específicos para a descrição linguística, como o *WordSmith Tools* ou o *Sketch Engine*. Em ambas as formas de acesso, há três principais recursos utilizados em investigações linguísticas: concordâncias, lista de frequência de palavras e lista de colocações. O concordanciador busca, no *corpus*, uma palavra ou um sintagma específico, apresentando todas as ocorrências daquela palavra ou daquele sintagma com as palavras que os antecedem ou seguem. A lista de frequência se trata de

uma lista de todas as formas ou vocábulos (*types)* presentes em um *corpus*, assim como o número de ocorrências de cada forma/vocábulo (*tokens*). Por fim, a lista de colocações refere-se à tendência com que as palavras co-ocorrem com outras, apresentando colocações, padrões e fraseologias. Para o *Corpus* TecEM, como ficará melhor descrito na metodologia, propõe-se, em um primeiro momento, a sua disponibilização *on-line* a usuários em geral tanto para consulta à base de dados quanto para acesso às diferentes categorias do *corpus*. Em um segundo momento, planeja-se aprimorar o site do *corpus* TecEM, a fim de disponibilizar nele mesmo os três recursos apresentados. Atualmente, o *corpus* já está disponível *online* (https://web.osorio.ifrs.edu.br/pesquisa/corpustecem/) e o concordanciador está em fase de desenvolvimento.

Ao abordar a utilização de *corpora* em diferentes linhas da Linguística, Sarmento (2010) destaca áreas em que a aplicação da LdC tem se mostrado fértil, são elas: estudos do léxico e lexicografia, estudos gramaticais, variação e análise de gênero, estudos de tradução e ensino e aprendizagem de línguas. Diante dos conceitos apresentados até aqui, o *Corpus* TecEM é uma tentativa de preencher uma lacuna até então existente na área: a inexistência de um *corpus* de produções textuais autênticas em Língua Portuguesa elaboradas por alunos brasileiros de cursos técnicos integrados ao Ensino Médio.

3 Processos Metodológicos

A fim de que construir um *corpus* que atenda aos requisitos metodológicos necessários para que sua consistência seja atestada, foi necessária a obtenção de três objetivos específicos, a saber: (i) estabelecer os critérios para a organização dos dados linguísticos textuais coletados; (ii) organizar e rotular os textos do *corpus* de acordo com os critérios estabelecidos; e (iii) disponibilizar o *corpus* elaborado para futuras pesquisas relacionadas ao ensino de Língua Portuguesa, de modo geral, e ao ensino de língua em contextos educacionais tecnológicos, em particular.

Nosso estudo, como descrito nas seções anteriores, está inserido no quadro teórico da LdC, a qual, conforme Berber Sardinha (2004), "trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem como sistema

probabilístico; encaixa-se no que pode ser chamado de Linguística Empírica" (p. 30). Por abordagem empirista entendemos a doutrina filosófica segundo a qual o conhecimento se origina da experiência. Como sintetiza Berber Sardinha (2000), em Linguística, assumir tal abordagem significa "dar primazia aos dados provenientes da observação da linguagem, em geral reunidos sob a forma de um *corpus*. O empirismo se coloca em oposição ao racionalismo, segundo o qual, em linhas gerais, o conhecimento provém de princípios, estabelecidos *a priori*." (p. 350). Já a compreensão da linguagem como sistema probabilístico considera que, embora muitas combinações e características linguísticas sejam possíveis, nem todas são prováveis de ocorrer em determinados contextos.

A partir disso, os procedimentos adotados para a criação do *Corpus* TecEM foram desenvolvidos da seguinte forma, com vistas ao cumprimento dos objetivos delimitados:

> Passo 1: Levantamento bibliográfico, a fim de garantir o atendimento aos quatro pré-requisitos para a formação de um *corpus* computadorizado apresentados na seção anterior, a saber: autenticidade, produção por falantes nativos, seleção criteriosa de textos e representatividade (Biber, 1993);
> Passo 2: Contato com professores de Língua Portuguesa de Institutos Federais, em um primeiro momento, localizados no Rio Grande do Sul, para apresentação da proposta da pesquisa e convite aos interessados em contribuir na coleta de textos;
> Passo 3: Coleta de textos escritos por alunos de cursos técnicos integrados ao Ensino Médio durante suas aulas de Língua Portuguesa ao longo do ano (mediante assinatura de Termo de Consentimento Livre e Esclarecido pelos alunos e/ou responsáveis – para alunos menores de idade) com o auxílio dos professores interessados em colaborar com a construção do TecEM;
> Passo 4: Compilação dos textos a partir de critérios estabelecidos nos passos iniciais deste projeto;
> Passo 5: Armazenamento do *corpus* TecEM em um banco de dados *online*, definindo sua estrutura de maneira alinhada aos critérios de compilação;
> Passo 6: Levantamento de requisitos para o desenvolvimento das interfaces da ferramenta para acesso ao *corpus*;
> Passo 7: Desenvolvimento e disponibilização da ferramenta.

É importante destacar que a identidade dos autores dos textos foi preservada de todas as formas e que todos foram voluntários na pesquisa, tendo assinado um Termo de Consentimento Livre e Esclarecido, o qual apresenta informações sobre estudo e explicita o uso estritamente acadêmico e científico dos dados. Os alunos-autores estavam cientes, portanto, (i) das etapas de coleta dos textos; (ii) do uso a ser feito do *corpus*; (iii) de que podem desistir ou cancelar sua participação na pesquisa no momento em que desejarem.

Por fim, é necessário informar que este projeto foi aprovado pelo Comitê de Ética em Pesquisa da instituição em que está inserido[7].

3.1 Planejamento do *corpus*: apresentação dos critérios para a organização dos dados linguísticos textuais coletados

A representatividade é um dos princípios norteadores importantes da LdC e para a construção de um *Corpus.* Ao elaborarmos as diretrizes do *Corpus* TecEM, foi estabelecido como critério inicial para coleta dos textos que o compõem a sua autoria: alunos de cursos técnicos integrados ao Ensino Médio de Institutos Federais de Educação, Ciência e Tecnologia. Cabe ressaltar que textos de qualquer gênero textual e extensão podem ser incluídos no *corpus*, assim como as produções podem ser de autoria de alunos de qualquer curso técnico integrado ao Ensino Médio e ano. Tais informações estão vinculadas aos textos, dessa forma, ao realizar buscas, professores e pesquisadores podem filtrar os dados de acordo com seus interesses. No entanto, os critérios apresentados a seguir foram necessários, a fim de atender a logística para a construção do *corpus,* assim como para buscar representatividade dentro do contexto estabelecido.

Para tanto, a inclusão dos textos seguiu os critérios específicos abaixo, na ordem em que estão apresentados.

1. Coletar textos produzidos em Institutos Federais localizados no Rio Grande do Sul: IFRS, IFSul, IFFar. Essa decisão atende, em especial, uma questão de logística, necessária para possibilitar o desenvolvimento do projeto.

2. Buscar a textos oriundos de *campi* dos três IFs citados, a fim de que haja uma representatividade igualitária dos Institutos Federais localizados no estado

3. Observar o agrupamento dos COREDEs, fóruns regionais de discussão sobre estratégias, políticas e ações que visam ao desenvolvimento regional, em nove Regiões Funcionais, buscando textos oriundos de cada uma delas, a fim de que haja uma representatividade igualitária das diferentes regiões do estado.

[7] Certificado de apresentação para apreciação ética (CAAE) número 65469917.7.0000.8024 na Plataforma Brasil.

Entretanto, é importante destacar que, como o *corpus* se encontra em constante construção, nem todas as categorias já estão representadas de maneira igualitária. O objetivo, contudo, é - a longo prazo - ter representatividade de IFs localizados em todas as regiões do Brasil. A seguir, segue o detalhamento das análises que justificaram o estabelecimento dos critérios apresentados.

No Rio Grande do Sul, a rede federal de ensino tecnológico é dividida em três instituições: Instituto Federal Farroupilha, Instituto Federal Sul-rio-grandense e Instituto Federal do Rio Grande do Sul. Cada Instituição é formada por diversos Campi, que estão distribuídos pelo Rio Grande do Sul. O fato de o recorte dos COREDES em nove Regiões Funcionais ter sido estabelecido com base em estudos anteriores feitos nas regiões e considerar critérios de homogeneidade econômica, ambiental e social (RAMBO; VIANNA, 2018) levou-nos a adotar tal organização como guia para a escolha dos *campi* para coleta inicial. Diante disso, buscamos selecionar os *campi* dos três diferentes IFs do Rio Grande do Sul, a fim de que cada IF fosse contemplado com três campi. Como dito anteriormente, nem todas as produções já estão inseridas no *corpus*, uma vez que - para tanto - é necessário o engajamento de docentes dessas diferentes regiões. Entretanto, a coleta dos textos tem sido contínua, e a representatividade está próxima de ser atingida. A seguir, estão apresentadas as Regiões Funcionais do estado e os *campi* selecionados no primeiro momento de coleta. Textos oriundos do IFRS e IFFar já estão inseridos no *corpus*, os textos do IFSul ainda não foram disponibilizados. Como informado anteriormente, não há restrição ou justificativa quanto aos gêneros textuais inseridos, desde que atendam ao critério mais geral, isto é, terem sido elaborados por alunos de cursos técnicos integrados ao Ensino Médio em suas aulas de Língua Portuguesa. Como o *corpus* está em constante atualização, não se justifica a inclusão, neste texto, de um quadro quantitativo dos dados, mas os leitores podem obter tais informações a partir da seleção de critérios na tela de busca do *site* do *corpus*.

Região Funcional	Corede(s)	Campus
1	Centro Sul, Metropolitano Delta do Jacuí, Paranhana	IFRS - Campus Canoas

	Encosta da Serra, Vale do Caí e Vale do Rio dos Sinos	
2	Vale do Taquari e Vale do Rio Pardo	IFSul - Campus Venâncio Aires
3	Serra, Hortênsias e Campos de Cima da Serra	IFRS – Campus Bento Gonçalves
4	Litoral	IFRS - Campus Osório
5	Sul	IFSul – Campus Pelotas
6	Campanha e Fronteira Oeste	IFSul – Campus Santana do Livramento
7	Celeiro, Missões, Fronteira Noroeste e Noroeste Colonial	IFFar - Campus Santo Augusto
8	Alto Jacuí, Central, Jacuí Centro e Vale do Jaguari	IFFar - Campus Júlio de Castilhos
9	Alto da Serra do Botucaraí, Médio Alto Uruguai, Nordeste, Norte, Produção e Rio da Várzea	IFFar - Campus Frederico Westphalen

Quadro 1. Lista de Regiões Funcionais e *campi*

4 Descrição da ferramenta *online* do *corpus*

O desenvolvimento da ferramenta *online* para disponibilização do e consulta ao *corpus* mobilizou um trabalho conjunto de uma equipe multidisciplinar das áreas de Letras e Informática. Nas reuniões iniciais, foram discutidos os requisitos necessários para a criação do banco de dados, os quais foram norteados pelos critérios explicitados na seção anterior, e, então, deu-se início à fase de programação em si.

Hoje, a ferramenta apresenta três telas: tela inicial, tela de cadastro e tela de consulta. Cada uma delas será detalhada a seguir, a fim de que as suas funcionalidades fiquem claras.

4.1 Tela Inicial

Ao acessar a página do *Corpus* TecEM (https://web.osorio.ifrs.edu.br/pesquisa/*corpus*tecem/), o usuário encontra uma página de apresentação do projeto. Na parte central da tela, há três seções informativas disponíveis: o

que é?, Histórico e Critérios para composição do *Corpus*. O usuário pode selecionar a seção que deseja ler e terá acesso a textos informativos sobre cada item. Na primeira, há uma breve apresentação dos objetivos do *corpus*; na segunda, são apresentados os passos para construção do *corpus*, assim como a equipe envolvida em seu desenvolvimento; e, por fim, na terceira, são reproduzidos os critérios de compilação.

À esquerda, na mesma tela, estão os acessos para as outras duas. Com um "usuário" e "senha", os administradores do *corpus* podem fazer o login para ir à tela de cadastro. Já os demais pesquisadores precisam apenas selecionar a opção "buscas no *corpus*" para acessar a tela de buscas. Abaixo, está uma captura de tela da tela inicial.

Imagem 1. Tela Inicial
Fonte: site *Corpus* TecEM

4.2 Tela de Cadastro

A tela de cadastro é de acesso restrito aos administradores do *corpus*. Nela, são inseridos os dados relacionados aos autores e aos textos. Cada texto é inserido manual e individualmente, uma vez que há informações específicas vinculadas a cada produção.

Sobre a autoria do texto, são inseridos os dados relativos à instituição do aluno, ao curso e ao ano em que estava quando realizou a produção. Sobre o texto, indica-se a proposta apresentada pelo professor (ao enviar os textos, os professores enviam também a descrição das orientações que os alunos receberam), o período de produção, se ele foi inicialmente manuscrito (e então digitado pela equipe do *corpus*) ou se já foi entregue

digitado pelo aluno e o gênero textual. Após inserir os metadados de cada produção, é necessário selecionar o arquivo .txt para inclusão do texto ao *corpus*. A seguir, está a imagem do *layout* da tela de cadastro.

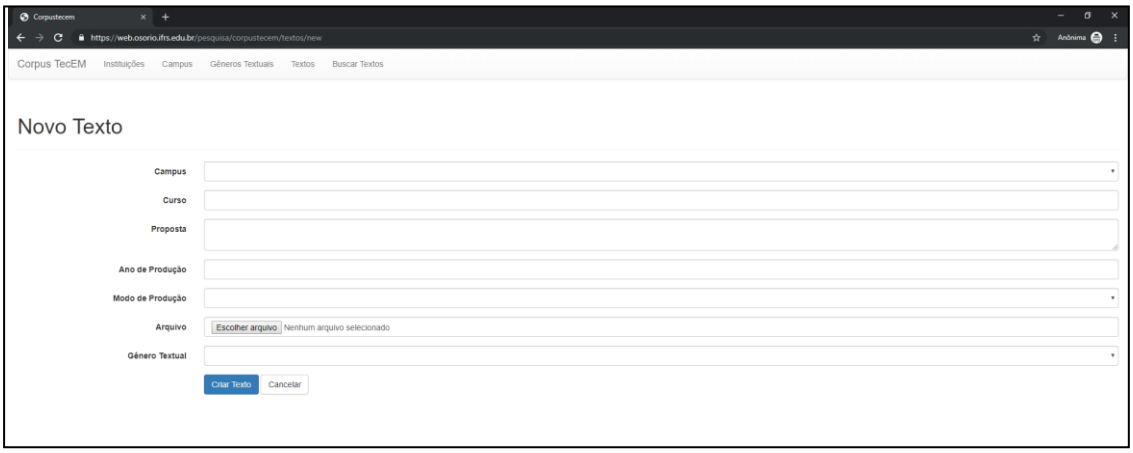

Imagem 2. Tela de Cadastro
Fonte: site *Corpus* TecEM

4.3 Tela de Buscas

Por fim, na tela de buscas, o pesquisador irá selecionar os critérios que estão disponíveis para a seleção de textos, que são: instituição; campus; curso; o intervalo das datas de produção e os gêneros textuais disponíveis. Ao selecionar os critérios, será baixado no computador do usuário uma pasta zipada com os textos que atendem a seleção realizada. Caso o pesquisador não queira utilizar nenhum dos critérios, todos os arquivos serão baixados também em uma pasta zipada. Em ambos os casos, os textos são disponibilizados no formato .txt.

A fim de auxiliar os usuários na busca, acima da tela de buscas, foram inseridas as seguintes orientações:

> Caros professores e pesquisadores, o *Corpus* TecEM foi desenvolvido para ser uma fonte de estudo, de pesquisas e de planejamento de aulas de Língua Portuguesa. Ficamos felizes em compartilhar esses dados textuais com vocês! Para baixar os textos, é possível: (i) selecionar os critérios listados ao lado, a fim de restringir a amostra; ou (ii) deixá-los (todos ou alguns) em branco. Quando o critério estiver em branco, todos os textos serão incluídos na seleção. Após este passo, deve-se clicar em "fazer download". Será, então, gerada uma pasta zipada com todos os textos que

> atendem aos seus critérios de seleção. Como esse material, diferentes investigações podem ser feitas a partir do uso de ferramentas da Linguística de *Corpus*, assim como materiais didáticos podem ser elaborados. (site do *Corpus* TecEM)

Como dito anteriormente, o concordanciador no site do *Corpus* ainda está em desenvolvimento. Apesar de aprimorar as ferramentas disponíveis no site próprio, essa ausência não exclui a opção de o professor ou pesquisador explorar o *corpus*, uma vez que é possível utilizar ferramentas que estão disponíveis online e que, além do concordanciador, possuem outras funcionalidades para investigação, como o *Sketch Engine*[8] e *Wordsmith*[9]. A seguir, estão as imagens da tela de buscas, assim como um exemplo de como os usuários terão acesso aos textos que constituem o *corpus.*

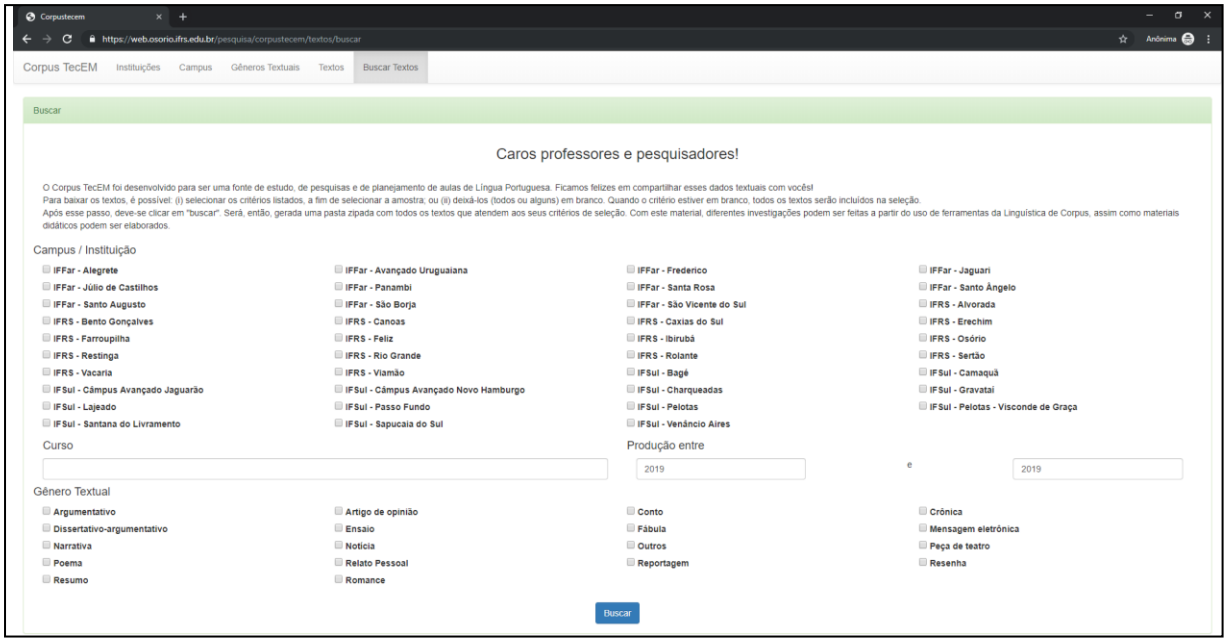

Imagem 3. Tela de Buscas
Fonte: site *Corpus* TecEM

[8] Disponível em: https://www.sketchengine.eu/.
[9] Disponível em: https://wordsmith.org/

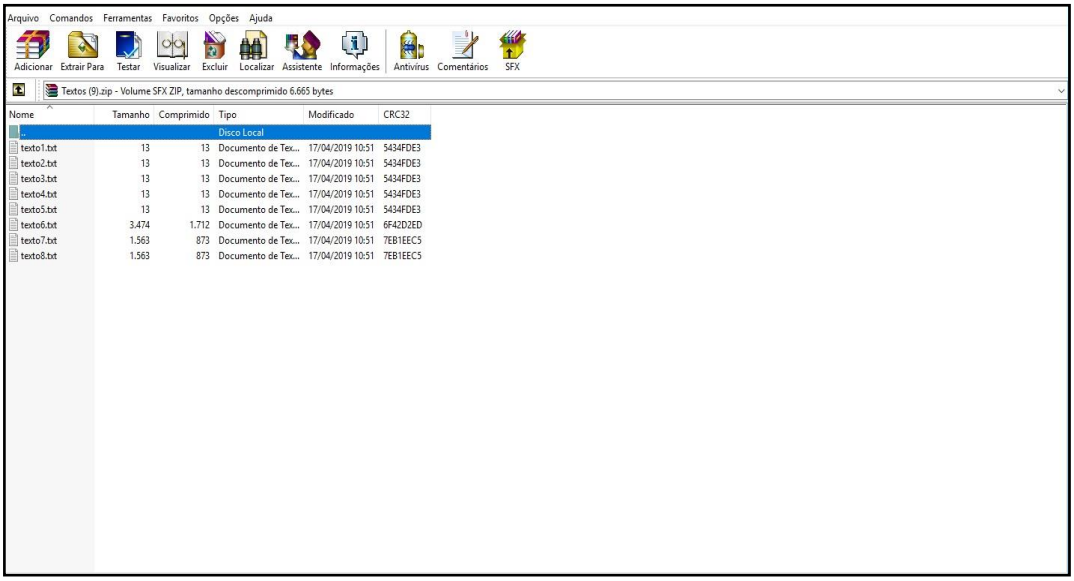

Imagem 4. Exemplo de acesso aos textos
Fonte: site *Corpus* TecEM

5 Considerações Finais

O projeto de construir um *corpus* com características únicas no cenário nacional tem se mostrado desafiador. A maior virtude deste projeto é também sua maior dificuldade: a especificidade do contexto de coleta dos textos. Se, por um lado, o resultado deste esforço é um vasto território de exploração para o ensino de língua portuguesa no Ensino Médio e, especificamente, no contexto de formação tecnológica; por outro, o caminho para o acesso aos textos tem se mostrado árduo. A maior dificuldade encontrada até o momento é o contato com professores de diferentes IFs (dificuldade que parece aumentar proporcionalmente ao aumento da distância geográfica entre os *campi*), pois, por vezes, os seus contatos não estão disponíveis nos sites dos seus *campi* ou, então, após a tentativa de contato, não recebemos um retorno positivo ou negativo sobre a contribuição para o *corpus*. Outra dificuldade tem sido a assinatura dos termos de consentimento para a disponibilização anônima dos textos, uma vez que muitos estudantes são menores de idade e é preciso do consentimento dos responsáveis para poder utilizar o texto na plataforma. Apesar dessas questões organizacionais, em março de 2019, o *Corpus* TecEM possuía 226 textos (111,800 palavras) escritos por alunos dos IFs localizados no Rio Grande do Sul e

mais textos já foram coletados a partir de abril do mesmo ano. Desta forma, a representatividade do *corpus* segue em uma crescente.

Os próximos passos do grupo são: (i) aumentar o número dos textos para ofertar uma maior gama de dados aos pesquisadores que optarem fazer pesquisas em produções textuais de alunos do Ensino Médio integrado a cursos técnicos oriundos dos IFs e (ii) disponibilizar o *corpus* em outras ferramentas de buscas, a fim de difundi-lo e, consequentemente, aumentar o interesse e a possibilidade de pesquisas nesta área de investigação. Para isso, contamos com a disponibilidade e o interesse de professores de língua portuguesa que lecionam no Ensino Médio em contextos de formação tecnológica. Todos aqueles que possuem interesse em contribuir com o *Corpus* TecEM podem entrar em contato com a equipe do projeto por meio do e-mail disponibilizado no site do *corpus*.

Ao fim da primeira etapa deste projeto, portanto, entende-se que foi disponibilizada à comunidade acadêmica uma base de dados rica e criteriosa para futuras pesquisas, contribuindo tanto para o desenvolvimento teórico quanto para a proposição de novas práticas de ensino de Língua Portuguesa, de modo geral, e em contextos de formação tecnológica, em particular.

Referências

ALUÍSIO, S; ALMEIDA, G. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Leopoldo, vol. 4, n. 3, p. 156-178, set/dez 2006.

BAKHTIN, M. *Estética da criação verbal.* São Paulo: Martins Fontes: 2006[1929].

BERBER SARDINHA. T. Linguística de *Corpus*: histórico e problemática. *D.E.L.T.A.*, São Paulo, v. 16, n.2, p. 323-367, 2000.

_____. *Linguística de Corpus*. São Paulo: Manole, 2004.

BIBER, Douglas. Representativeness in *Corpus* Design. *Literary and Linguistic Computing*, Oxford, v.8, n.4, p. 243-257, 1993.

BOWKER, L; PEARSON, J. *Working with Specialized Language*: a practical guide to using corpora. Londres: Routledge, 2002.

FARACO, C. A. *Linguagem e Diálogo*: ideias linguísticas do círculo de Bakhtin. São Paulo: Parábola, 2009.

FUNDAÇÃO DE ECONOMIA E ESTATÍSTICA DO RS. *Coredes.* Disponível em: <https://www.fee.rs.gov.br/perfil-socioeconomico/coredes/>. Acesso em 17 de abril de 2019.

GERALDI, J.W. *Portos de Passagem.* 4ª ed. São Paulo: Martins Fontes, 1997.Kennedy (1998),

MARCHIORO, M. Z. A análise linguística e o texto dissertativo-argumentativo: um olhar sobre o ensino de língua portuguesa. *Uniletras*. Ponta Grossa, v. 32, n. 1, p.9-30, jan./jun. 2010.

MOTTIN, L. *Análise da produção metafórica no Brazilian English Learner Corpus*. Dissertação (mestrado em Letras). Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

PACHECO, A. *A aquisição de morfemas em inglês como L2:* uma análise dos padrões evolutivos através do BELC (*Brazilian English Learner Corpus*). Tese (doutorado em Teoria e Análise Linguística). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

RAMBO, A. G. ; VIANNA, G. M. . Mecanismos de governança e escalas do desenvolvimento: considerações sobre o Colegiado Territorial e o Conselho Regional de Desenvolvimento no Litoral Norte Gaúcho. In: IV Seminário de desenvolvimento regional, estado e sociedade: democracia e desigualdades regionais, 2018, Palmas. *Anais do IV Seminário de desenvolvimento regional, estado e sociedade: democracia e desigualdades regionais*, 2018. p. 422-436.

RIO GRANDE DO SUL. *Constituição estadual de 1989*. Disponível em: <http://www2.al.rs.gov.br/dal/LinkClick.aspx?fileticket=9p-X_3esaNg%3D&tabid=3683&mid=5358> Acesso em 17 de abril de 2019,

_____. *Regiões Funcionais*. Disponível em: <https://planejamento.rs.gov.br/28-regioes>. Acesso em 17 de abril de 2019.

SARMENTO, S. Linguística de *Corpus*: histórico, metodologia, campos de aplicação. *Revista Trama*. Vol. 6, n. 12, p. 87-107, 2010.

TAGNIN, S. A multilingual learner *corpus* in Brazil. In: WILSON. A; ARCHER, D; RAYSON, P. (Org.). *Corpus linguistics around the world*. Amsterdam - New York: Rodopi, p.195-202, 2006.

_____; FROMM, G. CoMaprend – a experiência da construção de um *corpus* de aprendizes para estudo. *Domínios de Lingu@gem*, Uberlândia, v.2, n.2, 2008. Não paginado.

Data de submissão: 18/04/2019. Data de aprovação: 23/05/2019