XXVIII Congresso {virtual} de Iniciação Científica da Unicamp 2020

Compilação de um corpus anônimo de textos e excertos produzidos no curso ProFIS

O objetivo desta Iniciação Científica foi compilar um corpus de materiais escritos produzidos no primeiro semestre de 2013 do curso ProFIS, durante a disciplina de Escrita e Leitura Acadêmica I e, a partir dessa compilação, produzir a sistematização estruturada de dados qualificados para análises linguísticas de interesse para os pesquisadores do IEL, parte integrante do ProCorp (ProFIS Corpus), seguindo as direções dos demais trabalhos desenvolvidos pelo Grupo de Pesquisa CNPq *Práticas de escrita e de reflexão sobre a escrita em diferentes mídias*.

Como definido por Sinclair (2005, apud Aluísio e Almeida, 2006, p. 157)1, para a Linguística de Corpus, um corpus é uma coleção de textos em formato eletrônico, selecionados de acordo com critérios externos para representar, na medida do possível, uma língua ou variedade de língua como fonte de dados para pesquisa linguística. Dado que o ProCorp se propõe a documentar o processo de aquisição de escrita acadêmica, é necessário que os elementos de variação linguística, textual e de gênero discursivo permaneçam em sua forma original. Por isso, na transcrição, os textos foram mantidos da forma como foram escritos.

Todos os arquivos foram anonimizados no processo de transcrição, preservando, assim, a identidade dos autores, seguindo critérios e parâmetros aprovados pelo Comitê de Ética em Pesquisa (CEP) da UNICAMP (Parecer Conep no. 2.214.618), segundo os quais os nomes dos estudantes, datas, nomes de lugares e referências a colegas, bem como características ortográficas, foram apagados. Um termo de consentimento também foi assinado pelos participantes.

Tais elementos não interferem nos processos de sistematização e análise dos textos e fragmentos que compõem o corpus, já que os métodos analíticos utilizados neste projeto, embora também fundamentados em análises qualitativas da interação, pressupõem a observação de tendências discursivas e interativas genéricas com base em padrões quantitativamente identificáveis.

Os arquivos digitalizados são nomeados de acordo com a convenção previamente estabelecida para o ProCorp, que é a identificação formada pelo código do sujeito (S + número sequencial) + sexo (M ou F) + idade absoluta + ano e semestre da produção + código da atividade.

O processo de transcrição de originais manuscritos criou arquivos em TXT próprios para sistematização e leitura por ferramentas computacionais a partir dos documentos escaneados. A nomenclatura dos arquivos de transcrição seguiu a mesma convenção dos arquivos digitalizados.

Para o processo de transcrição também foram convencionadas as etiquetas <riscado>, <acrescido> e <incompreensivel> para trechos rasurados, acrescidos ao texto e incompreensíveis no original manuscrito, respectivamente.

QUADRO DE DESCRIÇÃO GERAL DOS RESULTADOS

Materiais incorporados ao corpus no período de agosto de 2019 a agosto de 2020		
Atividade	**Quantidade de arquivos**	**Formato original**
Atividade Diagnóstica (D2)	89	Manuscrito
Atividade 1 (A1)	68	Manuscrito
Atividade 2 (A2)	58	Manuscrito
Atividade 3 (A3)	60	Manuscrito
Atividade 4 (A4)	121	Manuscrito
Prova 1 (P1)	141	Manuscrito
Prova 2 (P2)	130	Manuscrito e digitado
Prova 3 (P3)	130	Manuscrito
Exame (E)	33	Manuscrito

Sendo o ProCorp o primeiro conjunto de dados de aquisição de gêneros acadêmicos em português brasileiro por ingressantes de um programa de inclusão da Unicamp, ele pode trazer grandes contribuições para as pesquisas sobre práticas de letramento por aprendizes com o mesmo perfil em diferentes cursos e programas institucionais. Nesse sentido, é relevante o fato desse corpus compreender uma grande diversidade de materiais produzidos por sujeitos vindos de contextos de aprendizagem também diversificados, em razão das diferenças sociais e das disparidades que caracterizam o sistema educacional brasileiro.

O projeto por mim desenvolvido, que permitiu a inserção de um total de 830 documentos, consistiu, portanto, numa etapa relevante para que o ProCorp possa começar a ser explorado pelos pesquisadores o mais breve possível.