UNIVERSIDADE E COMUNIDADE EM CONEXÃO

XXXV SALÃO de INICIAÇÃO CIENTÍFICA

6 a 10 de novembro

Evento	Salão UFRGS 2023: SIC - XXXV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2023
Local	Campus Centro - UFRGS
Título	O processo de análise manual de dados fraseológicos no projeto Internacionalização da Produção Acadêmica com Corpus e Tecnologia (InPACT- CNPq)
Autor	GABRIELA CARVALHO ESCOBAR
Orientador	ANA ELIZA PEREIRA BOCORNY

Nos últimos anos, a produção científica brasileira cresceu significativamente, conquistando reconhecimento global. No entanto, esse crescimento não se refletiu proporcionalmente em seu impacto, principalmente devido aos desafios associados à publicação em inglês. Para enfrentar esse problema, a Aprendizagem Direcionada por Dados (*data-driven learning*) emerge como uma solução, aproveitando dados linguísticos extraídos de corpora para desenvolver recursos pedagógicos. Nesse contexto, o projeto "Internacionalização da Produção Acadêmica com Corpus e Tecnologia" (InPACT), vinculado ao CNPQ, propõe a criação de um recurso pedagógico de acesso aberto. Esse recurso será construído com base em padrões linguísticos extraídos de artigos de revistas internacionais de alto impacto nas áreas de Humanidades, Ciências Sociais Aplicadas e Linguística, Letras e Artes. O projeto começou com a compilação de um corpus utilizando a ferramenta AntCorGen (Anthony, 2022), selecionando 2000 textos de cada seção de artigos de pesquisa em 16 disciplinas de humanidades. Posteriormente, os corpora foram ajustados para conter uma amostra representativa de 1 milhão de palavras cada. Em seguida, foram identificados e extraídos pacotes lexicais (*lexical bundles*) com a ferramenta Sketch Engine (Kilgariff et al., 2014), estabelecendo critérios de extensão de 5 a 6 palavras, frequência mínima de 10 por milhão de palavras e dispersão em 5 textos distintos. A análise dos dados começou com a classificação manual dos pacotes lexicais, alinhando-os com a estrutura retórica das seções dos artigos de pesquisa. Posteriormente, a ferramenta ChatGPT foi utilizada para agrupar os pacotes lexicais com base em sua similaridade e função retórica. Uma próxima etapa visa automatizar ainda mais o processo, buscando padrões de estruturas lexicais (*lexical frames*) nos dados extraídos. O recurso pedagógico em desenvolvimento tem o potencial de se tornar um importante assistente na escrita acadêmica em inglês na área de humanidades, especialmente após a automação adicional do processo.