Evento	Salão UFRGS 2017: SIC - XXIX SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2017
Local	Campus do Vale
Título	Enfoque estatístico inicial sobre complexidade textual: características do vocabulário no corpus TERMISUL em Português sobre Conservação, Restauração e Preservação de Documentos em Suporte Papel
Autor	VINÍCIUS ALCES MACHADO
Orientador	MARIA JOSE BOCORNY FINATTO

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE LETRAS

Enfoque estatístico inicial sobre complexidade textual: características do vocabulário no *corpus* TERMISUL em Português sobre Conservação, Restauração e Preservação de Documentos em Suporte Papel

AUTOR: Vinícius Alces Machado (UFRGS)
ORIENTADORA: Profa. Dra. Maria José Bocorny Finatto (UFRGS)

RESUMO: este trabalho apresenta uma pequena amostra do projeto em andamento sobre a complexidade textual com os *corpora* de idioma português reunidos pelo grupo TERMISUL para a criação de um glossário multilíngue da área de Conservação, Preservação e Restauração de Documentos em Suporte Papel (CPRDSP). Para um estudo-piloto, selecionaram-se os textos de três revistas acadêmicas que tratam sobre CPRDSP. O universo lexical desses textos soma 8.023 palavras diferentes de uma totalidade de 62.307. Com suporte da Linguística de Corpus e da Estatística Lexical, foi comparado o vocabulário desses textos com a lista de palavras do *Dicionário Ilustrado do Português*, de M. T. C. Biderman, obra que se propôs a repertoriar o universo das palavras do vocabulário mais básico do Português escrito. Foram verificados também os índices de variedade lexical e número médio de palavras por sentença, fatores que podem contribuir para a complexidade de um texto. As ferramentas utilizadas para gerar os dados foram o *AntConc*, *Flesch Calculator* e *Compare two lists*. Os resultados iniciais obtidos foram: na comparação com a lista de Biderman, que contém aproximadamente 5.700 verbetes, apenas 1.671 integram o universo lexical dos textos sobre CPRDSP, o que representa apenas 20% de palavras contempladas no estudo de Biderman. No entanto, para refinar esses resultados, entre as 6.352 palavras que são encontradas apenas no universo lexical dos textos, será preciso retirar aquelas que são formas flexionadas dos lemas presentes nas entradas do dicionário, assim como antropônimos e topônimos. O índice de variedade lexical dos textos da amostra tem média de 33%, o que é considerado esperável frente a outros tipos de textos. No entanto, quando a análise volta-se para a média de palavras por sentença, os resultados apontam um provável fator de complexidade dos textos. A menor média é 26,61, o que caracteriza uma escrita com várias sentenças longas que, em tese, tendem a exigir maior esforço do leitor para compreendê-las. A maior média é 42,06. Por sua vez, o Índice Flesch (IF) dos textos exibe valores abaixo de 40, o que confirma a tendência de alta complexidade, salientando-se que o valor máximo do IF em 100 corresponde a textos, em tese, muito fáceis. Assim, os resultados do estudo piloto fazem acreditar que as produções intectuais da área sobre CPRDSP, nesse tipo de texto, tenderiam, em tese, para uma alta complexidade de vocabulário, com sentenças bem extensas. Esperava-se o uso de vocabulário complexo, visto serem textos técnico-científicos com alta riqueza terminológica, mas este estudo piloto já permitiu ponderar sobre outros elementos também importantes para compor o quadro da complexidade dos textos do *corpus* TERMISUL, tendo-se em mente o futuro glossário, cujas definições precisarão ser acessíveis ao usuário da obra. Esse estudo ainda está em andamento e prentende aprofundar a análise desses resultados e incorporar mais dados de outros textos, além de artigos, avançando a observação com teses e dissertações sobre o tema em foco.