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

Vinícius Alces Machado (Bolsista de IC, UFRGS)
Orientadora: Prof. Drª. M. J. B. Finatto

Contextualização

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).

Os textos selecionados para esta análise são onze artigos que tratam de CPRDSP publicados em três periódicos:

	Types	Tokens
ABC	3.016	13.639
Acervo	4.832	24.673
Ágora	4.005	23.995
Todas	8.023	62.307

Objetivo

- Observar a complexidade textual (CT) presente nos textos que compõe o *corpus* do idioma português do grupo TERMISUL para a criação de um glossário multilíngue CPRDSP a fim de produzir definições acessíveis a especializados e leigos no assunto.

Metodologia

Com suporte da Linguística de Corpus e da Estatística Lexical:

1. Verificação dos índices de variedade lexical e número médio de palavras por sentença, pois esses são fatores que podem contribuir para a complexidade de um texto.
2. Comparação do vocabulário desses textos com a lista de palavras do Dicionário Ilustrado do Português (2005) de M. T. C. Biderman.
3. As ferramentas utilizadas para gerar os dados foram o *AntConc*, *Flesch Calculator* e *Compare two lists*.

Sobre o dicionário de M. T. C. Biderman:

Baseado em um *corpus* específico, essa obra buscou representar um universo vocabular compatível com leitores em início de escolarização. Compostas por textos de vários genêros discursivos, as 5 milhões de ocorrências foram comparadas com materiais didáticos da 1ª a 4ª série, permitindo delimitar um vocabulário elementar do português brasileiro. Nesta análise sobre CT, o dicionário de Biderman é usado como referência de vocabulário de baixa complexidade.

Resultados

	ABC	Acervo	Ágora
TTR	34,65	32,10	28,30
Palavras/ sentença	30,44	35,32	35,03
IF	29,18	25,36	28,94

- A média da variedade lexical é 33%.
- A média do número de palavras por sentença aponta um provável fator de complexidade dos textos.
- Os textos apresentam valores abaixo de 40 no Índice Flesch (IF), salientando-se que o valor máximo do IF em 100 corresponde a textos, em tese, muito fáceis.
- Na comparação com a lista de verbetes do dicionário de Biderman, entre as 8.023 formas:

  - 62% são palavras simples – nesta contagem, contabilizam-se as entradas do dicionário e suas lematizações, assim como antropônimos e topônimos.
  - 38% são palavras complexas – não possuem equivalente na obra de Biderman, que é usada nesta análise como referência de vocabulário de baixa complexidade.

A complexidade lexical, de forma isolada, não é suficiente para verificar a CT de um texto. Há outros traços que precisam ser observados nesse tipo de análise, como a frequência e o tamanho das palavras, o tamanho das sentença e o número de palavras por sentença.

Exemplo de sentença possivelmente complexa:

Esse diretor também se apresentava com um certo conhecimento técnico da matéria arquivística, defendendo, pela primeira vez, aquilo que mais tarde será denominado “descarte”, ou seja, a inutilização de documentos sem importância e a necessidade de se remeter a relação deles ao governo, pedindo autorização para serem vendidos ou inutilizados, “providência esta que, como na Europa, deverá ser repetida depois de certo trato de tempo, a fim de não tomarem espaço inutilmente”

Considerações finais

- Os resultados deste estudo piloto fazem acreditar que as produções intectuais da área sobre CPRDSP tendem, em tese, para uma alta complexidade de vocabulário, com sentenças muito extensas.
- O uso de vocabulário complexo era esperado, visto serem textos técnico-científicos com alta riqueza terminológica, mas este estudo 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 aos usuários da obra.

Referências: ANTHONY, L. (2014). *AntConc (Version 3.4.4)* [Programa de Computador]. Tóquio, Japão: Universidade de Waseda. Disponível em: http://www.laurenceanthony.net/. Acessado em: 13 set. 2017. | BIDERMAN. M.T.C. *A face quantitativa da linguagem: um dicionário de frequências do português do Brasil*. São Paulo: Alfa, 1998. p. 161-181. | BIDERMAN, M. T. C. *Dicionário Ilustrado de Português*. São Paulo: Ática, 2005. | CAMARGO, M. T. A. *Estatística Linguística*. [S.l.: s.n.], 1967. | COMPARE TWO LISTS. Whitehead BaRC public tools. 2013. Disponível em: <http://jura.wi.mit.edu/bioc/tools/compare.php>. Acesso em: 13 set. 2017. | FINATTO, M. J. B. ; EVERS, A. ; PASQUALINI, B. Desenhos do vocabulário no Português popular escrito: pesquisa, ensino, produtos lexicográficos e terminológicos. In: MURAKAWA, Clotilde de Almeida Azevedo; NADIN, Odair Luiz (Org.). *Terminologia: uma ciência interdisciplinar*. 1ed. São Paulo: Cultura Acadêmica, 2013, p. 255-278. | FULGÊNCIO, Lúcia; LIBERATO, Yara. *Como facilitar a leitura*. 6. ed. São Paulo: Contexto, 1992. | WILKENS, Rodrigo. Flesch Calculator for Portuguese. Porto Alegre: UFRGS, inédito, 2016.