SEQUÊNCIAS SEMÂNTICAS: UM ESTUDO BASEADO EM CORPUS

Tania Maria Granja Shepherd (UERJ)
tania.shepherd@gmail.com

O estudo da produção de um computador a partir de um corpus digital mostra a repetição de itens lexicais individuais, a repetição destes itens lexicais atrelados a outros itens, e o que é mais extraordinário, assim como descreve Hunston (2010), repetições padronizadas inerentes a gêneros textuais ou discursos específicos. O presente trabalho explora estas repetições padronizadas e está na interface da Linguística de Corpus e do estudo da fraseologia de discursos específicos em língua portuguesa.

Primeiramente, o trabalho mostra como extrair e trabalhar com termos de busca, emulando as abordagens de Gledhill (2000), Charles (2004) e Groom (2007), descritas em Hunston (2010). Essas abordagens servem como ponto de partida para verificar as marcas fraseológicas distintas dos gêneros diversos que compõem o Banco de Português (hospedado na PUC-SP), contendo cerca de um bilhão de palavras.São extraídas as sequências semânticas mais frequentes de alguns gêneros que compõem o Banco de Português e são feitas comparações das fraseologias que permeiam estes gêneros. Por fim são feitas algumas considerações acerca da validade de se basear um pesquisa sobre fraseologia a partir do que Hunston chama ´pequenas palavras´, ou seja, as preposições e alguns advérbios.