XXV ENCONTRO DE JOVENS PESQUISADORES
VII MOSTRA ACADÊMICA DE INOVAÇÃO E TECNOLOGIA
De 17 a 19 de outubro de 2017
Campus-Sede da UCS • Caxias do Sul
UCS UNIVERSIDADE DE CAXIAS DO SUL

LINGUÍSTICA DE *CORPUS:* A LEMATIZAÇÃO DO CORPUS JORNALÍSTICO DE DATA-UV

Tainá Copetti Bernardi (PIBIC-CNPq), Heloisa Pedroso de Moraes Feltes (Orientador(a))

Esta comunicação tem o objetivo de descrever os processos de lematização e limpeza do *corpus*
 jornalístico compilado para o projeto DATA – UV, o qual é constituído por matérias de quatro jornais brasileiros dos estados do Rio Grande do Sul e São Paulo. Os jornais selecionados para a construção do *corpus* são a Folha de São Paulo, o Estado de São Paulo, Zero Hora e Pioneiro, de 1º/01 a 31/12 de 2014. Com esse *corpus* visa-se à análise das representações da violência urbana na mídia impressa brasileira. O processo de lematização, que envolve o agrupamento de formas com diferentes variações de desinência e de sufixação de gênero e grau em uma só entrada no *corpus*, foi realizado em parceria com a Universidade Tecnológica Federal do Paraná. Para tal processo, foi utilizado *The Parser System Palavras,* que é parte integrante do projeto *Visual Interactive Syntax Learning do Institute of Language and Communication da University of Southern Denmark.* Entre algumas das variações disponíveis, a que melhor atendeu a demanda do *corpus* em estudo foi a opção *flat* do sistema. Com o resultado obtido foram criadas duas versões de saídas, uma contendo os balizadores e outra não. Para a automatização do processo foi utilizado um script PHP tanto para o uso das expressões regulares quanto para a consulta automática no sistema Palavras. Após a lematização, com o uso do *software* Text Crawler, foi possível verificar e corrigir algumas palavras que não foram devidamente alteradas pelo processo de lematização. O *corpus* não lematizado possui 1.778.282 e o lematizado possui 1.960.626 palavras. Após todos os procedimentos, o que foi possível conferir, até o momento, é que esse processo viabiliza uma análise mais direta dos termos buscados. Dessa maneira, tanto a busca de palavras-chave quanto a análise de colocados revelou dados diferentes daqueles com o *corpus* não lematizado, os quais permitem novas formas de análise com resultados mais abrangentes com o uso dos *softwares* AntConC e GraphColl. O *corpus* não lematizado é utilizado ainda na análise qualitativa de excertos de textos.

Palavras-chave: violência urbana, corpus, lematização

Apoio: UCS, CNPq