CORPUS.EAD: CONSTRUÇÃO E DESCRIÇÃO

Souza, J. W. C.[1](IC); Di-Felippo, A.[1](O)
jackcruzsouza@gmail.com
[1]*Departamento de Letras, Universidade Federal de São Carlos*

No Processamento Automático das Línguas Naturais (PLN), um sistema que processa língua natural (p.ex.: tradutor automático) pode ser linguisticamente motivado e, nesse caso, ele necessita de uma base de dados léxico-conceituais. Para o desenvolvimento dessas bases, o modelo *wordnet* é bastante difundido, o qual teve origem com a construção da base WordNet de Princeton. Dada a necessidade crescente de se processar textos especializados, *wordnets* terminológicas passaram a ser desenvolvidas para várias línguas. Tais recursos são comumente construídos com base em metodologias que consistem na aquisição manual do conhecimento léxico-conceitual armazenado em recursos estruturados (p.ex.: dicionários, glossários). Diante da escassez de recursos especializados que sejam estruturados e da demora na coleta manual do conhecimento, observa-se a carência de uma metodologia suficientemente clara e genérica que facilite e estimule a criação dessas bases. No projeto TermiNet (2009-2011) (FAPESP 2009/06262-1/ CNPq 471871/2009-5), especificou-se uma metodologia para a construção de *wordnets* terminológicas (ou *terminets*) que se caracteriza pela extração semiautomática de conhecimento a partir de recursos não-estruturados. Diante da escassez de recursos léxico-computacionais em português do Brasil, tal metodologia está sendo validada com a construção de uma *terminet* do domínio da Educação a Distância (EaD), a WordNet.EaD. Os recursos não-estruturados nada mais são do que os *corpora* textuais que, a depender do domínio especializado sob sistematização, precisam ser construídos. Assim, a metodologia para o desenvolvimento de *terminets* engloba uma metodologia de construção de *corpora,* a qual foi elaborada em função de princípios da Linguística de *Corpus.* Segundo essa metodologia, um *corpus* deve ser construído de acordo com os seguintes passos: (a) projeção do *corpus*, que consiste na definição do tipo do *corpus* necessário à pesquisa; (b) seleção das fontes e compilação dos textos que comporão o *corpus*, (c) pré-processamento do *corpus,* que consiste na preparação do mesmo para o tratamento computacional e engloba os processos de conversão, limpeza, nomeação, armazenamento e anotação dos textos; e (d) disponibilização do *corpus*. Para validar essa metodologia, construiu-se, no âmbito do projeto de IC "Construção do *corpus* para o desenvolvimento de uma *wordnet* terminológica em português do Brasil", o *Corpus.*EaD. Neste trabalho, em especial, apresentam-se as estratégias adotadas para a realização dos quatro passos (a-d) que compõem a metodologia. Além disso, descreve-se o *Corpus.*EaD em função do número total de ocorrências e do número de textos por gênero. De modo geral, os resultados do referido projeto de IC são: (a) aquisição de um arcabouço teórico-metodológico para a construção de *corpora* com vistas à construção de *terminets*; (b) construção do *Corpus.*EaD, que poderá servir de fonte para outros projetos terminológicos em língua portuguesa; (c) disponibilização do *Corpus.*EaD na *web*; e (d) contribuição para o desenvolvimento de sistemas de PLN linguisticamente motivados.

CNPq