DESCRIÇÃO MORFOLÓGICA DOS TERMOS DA EDUCAÇÃO A DISTÂNCIA NO PROJETO TERMINET

Gianoti, Ana C.[1](IC); Di-Felippo, Ariani[1](O)
aninhagianoti@gmail.com
[1]*Departamento de Letras, Universidade Federal de São Carlos*

No Processamento Automático das Línguas Naturais (PLN), os sistemas computacionais (p.ex.: tradutores automáticos) podem ser construídos segundo o paradigma simbólico e, nesse caso, necessitam de conhecimento variado sobre a língua em processamento. Tal conhecimento é comumente codificado em gramáticas e bases de dados lexicais. Para o desenvolvimento das bases lexicais de língua geral ou terminológicas, o modelo *wordnet*, que teve origem com a construção da base WordNet de Princeton (FELLBAUM, 1998) para o inglês americano, é bastante difundido. A construção de *wordnets* terminológicas, em especial, baseia-se comumente na extração manual do conhecimento léxico-conceitual a partir de recursos estruturados (p.ex.: dicionários, glossários). Diante especialmente da escassez de recursos especializados estruturados, especificou-se, no projeto TermiNet (2009-2011) (FAPESP 2009/06262-1/ CNPq 471871/2009-5), uma metodologia que se caracteriza pela extração semiautomática do conhecimento a partir de recursos não-estruturados (ou seja, *corpus*), a qual estabelece que a construção de uma *wordnet* terminológica deve ser feita em três fases: a linguística, a representacional e a implementacional (DI-FELIPPO, ALMEIDA, 2010). Tal metodologia está sendo validada com a construção de uma *wordnet* do domínio da Educação a Distância (EaD) em português do Brasil (PB), a WordNet.EaD. Segundo a metodologia do TermiNet, as etapas realizadas no domínio linguístico para a construção da WordNet.EaD foram: (i) delimitação e construção do *corpus* e (ii) extração do conhecimento léxico-conceitual (no caso, as unidades lexicais e as relações de sinonímia e hiponímia). A pesquisa de iniciação científica/tecnológica do qual este trabalho faz parte teve como primeiro objetivo extrair os candidatos a termos do *corpus* com base no paradigma linguístico (CABRÉ *et al*., 2001). Os candidatos, uma vez validados, são inseridos na WordNet.EaD. A extração em questão foi feita por meio da ferramenta ExATOlp (Extrator Automático de Termos para Ontologias em Língua Portuguesa) (LOPES et al., 2009), que identifica os candidatos em *corpora* anotados em nível morfossintático pela identificação da nuclearidade em sintagmas nominais (GIANOTI, 2011). Após os processos de limpeza e validação dos candidatos, partiu-se para o segundo objetivo do projeto: a descrição morfológica dos termos validados. Apesar de a descrição morfológica não estar diretamente ligada à construção de uma *wordnet,* salienta-se que esta é relevante para, por exemplo, refinar os extratores automáticos de termos do PB por meio da inserção de conhecimento sobre a estrutura morfológica dos termos. Para a descrição em questão, somente as lexias simples, ou seja, termos formados apenas por um único radical, foram considerados. Tais termos também são denominados "unigramas". Dentre todas as lexias simples validadas, optou-se por descrever as 200 mais frequentes, utilizando como fundamentação teórica para a identificação dos padrões morfológicos os trabalhos de Alves (2004) e Rocha (2008). Neste trabalho, descrever-se-á o método de análise das lexias quanto aos processos morfológicos identificados no domínio da EaD.

PIBITI/CNPq