O *CORPUS*.EaD NO PROJETO TERMINET: ESTRATÉGIAS DE CONSTRUÇÃO

Ariani Di Felippo (UFSCar)
Jackson Wilke da Cruz Souza (UFSCar)

No Processamento Automático das Línguas Naturais (PLN), um sistema que processa (interpreta e/ou gera) língua natural (p.ex.: tradutor automático) pode ou não ser linguisticamente motivado. No caso em que o sistema se baseia no processamento da língua em algum nível, certos recursos linguísticos são necessários. Dentre eles, citam-se as bases de dados lexicais, ou seja, espécies de dicionário armazenadas na memória do sistema de PLN. O desenvolvimento de recursos léxico-conceituais (ou seja, enriquecidos com informações semânticas) é um dos grandes desafios que os pesquisadores enfrentam. Essa dificuldade deve-se à grande quantidade, variedade e complexidade das informações léxico-semânticas. Um modelo de base lexical bastante difundido no PLN é o *wordnet,* que teve origem com a construção da base WordNet de Princeton (WN.Pr) (FELLBAUM, 1998). Em uma base do tipo *wordntet,* as unidades lexicais (palavras ou expressões) da língua em questão estão divididas em quatro categorias sintáticas: nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-sinônimas. Cada *synset* é construído de modo a representar um único conceito lexicalizado por suas unidades constituintes. Os *synsets* estão inter-relacionados pela relação léxico-semântica da antonímia e pelas relações semântico-conceituais da hiperonímia/hiponímia, holonímia/meronímia, acarretamento e causa.

Dada a necessidade crescente de se processar textos especializados, *wordnets* terminológicas passaram a ser desenvolvidas para várias línguas, p.ex.: JurWordnet (Sagri et al., 2004), ArchiWordnet (Bentivogli et al., 2004), Medical Wordnet (Smith, Fellbaum, 2004), BioWordnet (Poprat et al., 2008), etc. Tais recursos são comumente construídos com base em metodologias que consistem na aquisição manual do conhecimento léxico-conceitual armazenado em recursos estruturados, como dicionários, glossários, *thesaurus,* etc. Diante da escassez de recursos especializados que sejam estruturados e da demora na coleta manual do conhecimento descrito em tais recursos, observa-se que, embora exista um número razoável de *wordnets* terminológicas em diversas línguas, há carência de uma metodologia suficientemente clara e genérica que facilite e, sobretudo, estimule a criação dessas bases.

Diante disso, no projeto de dois anos denominado TermiNet, que teve início em Set./2009 (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 do conhecimento léxico-conceitual 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, em especial, 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 *wordnets* terminológicas proposta pelo TermiNet engloba uma metodologia de construção de *corpora.* Segundo essa metodologia, um *corpus* que servirá de fonte para o desenvolvimento de uma terminet deve ser construído de acordo com os seguintes passos: (a) projeto do *corpus*, (b) compilação dos textos que comporão o *corpus*, (c) pré-processamento (conversão, limpeza, nomeação e anotação) dos textos e (d) disponibilização do *corpus*. Para validar especificamente a metodologia de construção do *corpus,* construiu-se o *Corpus.*EaD, a partir do qual a WordNet.EaD será desenvolvida.

Neste trabalho, em especial, apresentam-se as estratégias adotadas para a realização das quatro tarefas (a-d) que compõem a metodologia de construção de *corpus* no projeto TermiNet. Tais estratégias são brevemente descritas na sequência.

Quanto ao projeto do *corpus,* que consiste na especificação da tipologia do mesmo, Di Felippo e Souza (2010) tomaram como base três grupos de critérios: (a) a definição do objeto *corpus*, (b) o tipo de recurso lexical a ser construído e (c) as decisões de projeto. Da discussão desses critérios, os autores delimitaram que o *corpus* nesse cenário precisa apresentar as seguintes características: (a) tamanho (representatividade e amostragem) → médio-grande; (b) balanceamento → por gênero; (c) modalidade →escrita (*vs* fala/áudio); (d) tipo textual → escrito (*vs* transcrições); (e) meio → jornais, livros, revistas, manuais e outros; (f) cobertura da língua → especializado; (g) gêneros → técnico-científico, científico de divulgação, informativo e instrucional; (h) quantidade de línguas → monolíngue; (i) anotação → anotado em nível morfossintático; (j) comunidade produtora → falantes nativos; (k) mutabilidade → aberto; (l) variações históricas → sincrônico (contemporâneo); (m) disponibilidade → disponível via *web.*

A fase de compilação dos textos, que engloba os processos de seleção das fontes e coleta dos textos armazenados nessas fontes, foi feita com base em uma delimitação do domínio de especialidade em questão, no caso, a EaD. Considerando-se a *web* como fonte principal no projeto TermiNet, a delimitação do domínio da EaD guiou a seleção de páginas da *web* e a procura e coleta de textos específicos armazenados em tais páginas. Vale ressaltar que, mesmo havendo inúmeras ferramentas computacionais que auxiliam a coleta em massa de textos na *web*, optou-se no TermiNet pelo método mais simples de seleção das fontes e coleta dos textos, caracterizado por: acesso às páginas desejadas e *download* dos textos em um computador local.

Na fase de pré-processamento, os textos devem ser convertidos para o formato txt, legível por máquina. No caso, optou-se pela conversão manual dos arquivos, posto que as ferramentas de conversão apresentaram várias limitações. O processo de conversão é importante porque, de acordo com a metodologia especificada no TermiNet, o conhecimento léxico-conceitual necessário à construção de uma terminet é extraído de forma semiautomática, ou seja, os textos são processados por ferramentas computacionais e os dados extraídos são revisados manualmente. Dados corrompidos pelo processo de conversão ou mesmo desnecessários para o projeto foram retirados dos textos na fase de limpeza. O pré-processamento engloba ainda a nomeação dos arquivos e a anotação estrutural de dados externos (ou seja, documentação do *corpus* em cabeçalho que inclui os metadados textuais como autoria, tipologia textual, etc.). No TermiNet, essas duas tarefas foram feitas por meio da ferramenta "Header Editor" disponível no "Portal de Córpus" do Projeto PLN-Br (http://www.nilc.icmc.usp.br:8180/portal/). Quanto à disponibilização, salienta-se que o *Corpus.*EaD ainda não está disponível na *web,* pois os pedidos de permissão de uso aos autores dos textos ainda não foram feitos.

Palavras-chave: construção de *corpus*; *wordnet* terminológica; recursos lexicais especializados; processamento automático das línguas naturais.