PROJETANDO O *CORPUS* PARA A CONSTRUÇÃO DE UMA *WORDNET* TERMINOLÓGICA

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

Para o desenvolvimento de vários sistemas computacionais que processam línguas naturais (p.ex.: sistemas de tradução automática, sistemas de sumarização, etc.), os pesquisadores da área do Processamento Automático das Línguas Naturais (PLN) necessitam de certos recursos linguísticos (os chamados *lingwares*), os quais desempenham papel central na arquitetura desses sistemas. Dentre esses recursos, destacamos, por exemplo, as “bases de dados lexicais” (BDLs). As BDLs são entendidas como grandes repositórios de informação lexical, um “mega arquivo”.

Dada a necessidade crescente de se processar textos especializados, BDLs terminológicas passaram a ser desenvolvidas para várias línguas, principalmente no formato *wordnet*. Embora exista um número razoável de *wordnets* terminológicas em diversas línguas, há carência de uma metodologia suficientemente clara que facilite e, sobretudo, estimule a criação dessas bases. Para o português do Brasil (PB), aliás, não há bases de conhecimento especializado no formato *wordnet*.

Assim, o projeto Terminet, que se encontra em estágio inicial, objetiva (i) instanciar uma metodologia genérica de pesquisa em PLN para o desenvolvimento de *wordnets* terminológicas (ou terminets) e (ii) aplicar tal metodologia instanciada na construção de uma base desse tipo em PB.

A metodologia a ser instanciada destaca-se por equacionar todo empreendimento no PLN em três fases (a linguística, a representacional e a implementacional) e por evidenciar a importância do conhecimento linguístico nesse tipo de empreendimento. Quanto às fases, ressaltamos que, na fase linguística, as atividades ficam concentradas na investigação dos fatos da(s) língua(s) em diferentes dimensões (morfológica, sintática, semântico-conceitual e pragmático-discursivo). Na fase representacional, por sua vez, estudamos os modelos formais de representação para os conhecimentos reunidos no domínio linguístico que sejam computacionalmente tratáveis. Em outras palavras, buscamos identificar o modelo formal mais adequado para a representação do conhecimento descrito na fase anterior. E, por fim, no domínio implementacional, as atividades ficam concentradas nas questões relativas à implementação do sistema de PLN.

Ao instanciar a metodologia de pesquisa em PLN para a construção de *wordnets* terminológicas, algumas das tarefas relativas ao domínio linguístico são: (a) projetar o *corpus*, (b) compilar os textos que compõem o *corpus*, (c) pré-processar (conversão, limpeza, nomeação e anotação) o *corpus* e (d) adquirir permissões de uso (caso o *corpus* seja disponibilizado na Web).

A tarefa de projetar o *corpus*, em especial, consiste na definição do tipo de *corpus* necessário à pesquisa, pois um *corpus* deve ser projetado em função da pesquisa para a qual ele está sendo construído.

Assim, para servir de base à construção de uma terminet, um *corpus* precisa, de início, apresentar certas características: (i) o *corpus* deve ser monolíngue; (ii) os textos devem ser relativos a um domínio especializado e proporcionar a descrição sincrônica do léxico temático desse domínio, e (iii) os textos devem ser do tipo escrito, ou seja, de língua falada ou escrita registrada em meio escrito, pois as bases *wordnets* são recursos para o tratamento computacional das línguas naturais registradas em tal meio.

Além dessa caracterização inicial, certos requisitos precisam ser atendidos para que uma coleção de textos possa ser denominada *corpus*. Tais requisitos podem ser: autenticidade, representatividade, amostragem, balanceamento, diversidade e tamanho. Tendo em vista a construção de terminets, tais requisitos podem ser assim satisfeitos: (a) a autenticidade pode ser garantida pela coleta de textos em comunicações “especializadas” genuínas e de fontes confiáveis, (b) a representatividade e a amostragem pela construção de um *corpus* médio (de 250 mil a 1 milhão de palavras), (c) a diversidade pela coleta de textos dos gêneros técnico-científico, científico de divulgação e instrucional, veiculados por livros, revistas, jornais, etc., e (d) o balanceamento pela distribuição dos componentes do *corpus* (gêneros, textos, etc.) em quantidades semelhantes.

Certas decisões de projeto também determinam propriedades específicas do *corpus*. No caso, as informações léxico-semânticas necessárias à construção de uma *wordnet*, terminológica ou não, são comumente obtidas por meio de métodos semi-automáticos de extração a partir de *corpora*. Tais métodos caracterizam-se pela aplicação de ferramentas computacionais de extração e subsequente verificação dos dados obtidos por linguistas. Tais ferramentas baseiam-se principalmente no reconhecimento de padrões léxico-sintáticos, o que só pode ser feito em um *corpus* que apresenta anotações no nível morfossintático.

Outra decisão de projeto diz respeito à disponibilização. Como os *corpora* especializados são recursos extremamente úteis e de construção cara, é recomendável a sua disponibilização via *Web*, tanto para pesquisadores do PLN quanto da Terminologia.

Além disso, um *corpus* para pesquisas terminológicas deve ser aberto, permitindo a inclusão e exclusão de textos para acompanhar as rápidas alterações que se registram nas terminologias pertencentes a certos domínios científicos e técnicos.

Com base na discussão de critérios como os descritos neste resumo e em outros, objetivamos apresentar o projeto do *corpus* que será utilizado no projeto Terminet.

Dessa forma, ao final deste trabalho, teremos delineado as principais características ou requisitos que um *corpus* precisa satisfazer para subsidiar a construção de *wordnets* terminológicas ou terminets em PB. Tal caracterização é de extrema relevância para o projeto Terminet, pois compreende a fase inicial do processo de construção de uma base terminet, mas também pode contribuir, de modo geral, para as discussões a respeito da construção de *corpora* especializados em PB.

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