UMA METODOLOGIA PARA O DESENVOLVIMENTO DE *WORDNETS* TERMINOLÓGICAS EM PORTUGUÊS DO BRASIL

***Ariani Di Felippo*[1] *& Gladis Maria de Barcellos Almeida*[2]**

RESUMO: Para o desenvolvimento de vários sistemas computacionais que processam língua natural (p. ex., sistemas de sumarização, sistemas de tradução automática etc.), os pesquisadores da área do Processamento Automático das Línguas Naturais (PLN) necessitam de certos recursos linguísticos (os *lingwares*), os quais desempenham papel central na arquitetura desses sistemas, p. ex.: as "bases de conhecimento lexical". Dada a necessidade crescente de se processar textos especializados, bases de conhecimento lexical especializado (ou terminológico) passaram a ser desenvolvidas para várias línguas, principalmente no formato *wordnet*. Ocorre que, embora exista um número razoável de *wordnets* terminológicas em diversas línguas, observa-se a 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*. Nesse cenário, está sendo desenvolvido o projeto TermiNet (do inglês, *terminological wordnet*), que objetiva: (i) a instanciação (ou seja, versão mais definida), para o desenvolvimento específico de

---

[1] Professora adjunta do Departamento de Letras (DL) da Universidade Federal de São Carlos (UFSCar) e pesquisadora do Núcleo Interinstitucional de Linguística Computacional (NILC) e do Grupo de Estudos e Pesquisas em Terminologia (GETerm).

[2] Professora associada do DL/UFSCar, fundadora do GETerm e pesquisadora vinculada ao NILC.

wordnets terminológicas, da metodologia genérica de pesquisa no PLN proposta por Dias-da-Silva (2006) e (ii) a sua aplicação na construção de uma base desse tipo em PB. Acredita-se que o TermiNet pode beneficiar não só o PLN, mas também a Terminologia/ Terminografia em PB, pois o formato *wordnet* é um modelo robusto e eficaz para a sistematização do conhecimento léxico-conceitual, fundamental também para o desenvolvimento de produtos terminográficos tradicionais. Neste artigo, em especial, apresenta-se o referido projeto, focalizando a primeira etapa das atividades, que é a instanciação da metodologia trifásica de Dias-da-Silva (2006).

PALAVRAS-CHAVE: Terminologia; PLN; Base de dados; Wordnet; TermiNet.

ABSTRACT: The development of computational systems capable of understanding and producing natural languages (e.g.: machine translation systems) requires some linguistic resources (lingwares), e.g. lexical knowledge databases. These resources are a crucial component of a wide variety of natural language processing (NLP) applications. Due to the increasing need to process specialized texts, domain-specific (or terminological) lexical databases have been built in many languages, especially in wordnet format. Despite the existence of a reasonable number of terminological wordnets in many languages, there is no clear and generic methodology to build them. For Brazilian Portuguese (BP), by the way, there is no domain-specific lexical database in the wordnet model. In this scenario, the TermiNet project has been developed. This project aims (i) to instantiate the generic NLP methodology proposed by Dias-da-Silva (2006) to develop terminological wordnets and (ii) to apply it to build a terminological wordnet in BP. In addition to the benefits to the NLP field, terminological wordnets may also contribute to the development of terminological/terminographic products as the organization of lexical-conceptual knowledge is an essential step in building such products. In this paper,

we offer an introduction to the TermiNet project focusing on the instantiation of the generic NLP methodology.

KEYWORDS: Terminology; NLP; Lexical database; Wordnet; TermiNet.

1. Introdução

Na área do Processamento Automático das Línguas Naturais (PLN), buscam-se desenvolver, em última instância, sistemas computacionais "capazes" de processar (interpretar/gerar) as línguas naturais, principalmente em meio escrito (Dias-da-Silva, 2006). Dentre eles, citam-se os sistemas de: tradução automática, correção ortográfica e gramatical, sumarização automática etc. (Mitkov, 2004). Quando baseados em conhecimento linguístico, tais sistemas podem apresentar uma arquitetura composta por três "bases de conhecimento estático": a gramatical, a conceitual e a lexical (Dias-da-Silva, 1996).

À base de conhecimento lexical (ou base lexical), em especial, cabe a tarefa de fornecer ao sistema uma coleção de unidades lexicais da língua que se está processando, juntamente com suas propriedades morfológicas, sintáticas, semânticas e pragmático-discursivas, dependendo da especificidade do sistema (Palmer, 2001; Hanks, 2004).

No caso do processamento semântico do inglês norte-americano, a Wordnet de Princeton (WN.Pr) (Fellbaum, 1998) é uma base lexical amplamente utilizada, principalmente por sua adequação científica e tecnológica (Morato *et al.,* 2004). Diante de sua reconhecida potencialidade tecnológica, a WN.Pr tem motivado a construção de bases lexicais no formato *wordnet* para inúmeras línguas. Atualmente, é possível encontrar *wordnets* para a maioria das línguas europeias, africanas e asiáticas. Em especial, a *wordnet* do português do Brasil (PB), a WordNet.Br (WN.Br) (Dias-da-Silva *et al.,* 2008), está em pleno desenvolvimento.

Nos últimos anos, dadas as aplicações reais para as quais os sistemas de PLN têm sido projetados, é premente que estes sejam "capazes" de processar textos técnicos ou especializados

(Jacquemin e Bourigault, 2004). Para tanto, faz-se necessário que as bases de conhecimento lexical sejam enriquecidas com "unidades terminológicas" (termos) associadas às suas respectivas propriedades.

Nesse sentido, é possível encontrar vários trabalhos, por exemplo, Magnini e Speranza (2001), Buitelaar e Sacaleanu (2002), Gangemi *et al.* (2003), Smith e Fellbaum (2004), Sagri *et al.* (2004), Bentivogli *et al.* (2004), Roventini e Marinelli (2004) e Poprat *et al.* (2008), que relatam a expansão das bases *wordnets* pelo acréscimo de conhecimento especializado. Isso se dá, especificamente, pela inclusão de "unidades terminológicas", ou seja, unidades lexicais da língua geral que se caracterizam por expressarem conhecimento especializado, produzido no âmbito das ciências e das técnicas (Cabré, 1999).

De modo geral, tal expansão é feita em duas etapas. Na primeira, sistematiza-se o conhecimento especializado de certo domínio no formato *wordnet* e, na segunda, integra-se esse conhecimento às bases de língua geral. Da sistematização realizada na primeira etapa, resultam bases lexicais autônomas, isto é, verdadeiras "*wordnets* terminológicas". Esse tipo de base pode ser exemplificado por:

(a) a JurWordnet (Sagri *et al.,* 2004) e a ArchiWordnet (Bentivogli *et al.,* 2004), responsáveis por enriquecer a *wordnet* do italiano com unidades terminológicas do domínio jurídico e da arquitetura, respectivamente.
(b) a Medical Wordnet (Smith e Fellbaum, 2004) e a BioWordnet (Poprat *et al.,* 2008), que ampliam a WN.Pr para os domínios da medicina e da biomedicina, respectivamente.

Embora exista um número razoável de *wordnets* terminológicas, observa-se a carência de uma metodologia suficientemente clara e genérica que facilite e estimule a criação de bases de conhecimento lexical especializado nesse formato.

Diante desse cenário, está sendo desenvolvido o Projeto TermiNet, o qual é descrito na próxima seção.

2. O projeto TermiNet

O Projeto TermiNet, financiado pela Agência de Amparo à Pesquisa do Estado de São Paulo (FAPESP)[3] e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[4], tem duração prevista de dois anos, sendo que as atividades tiveram início em setembro de 2009. O TermiNet objetiva, especificamente:

(a) instanciar a metodologia genérica de pesquisa no PLN elaborada por Dias-da-Silva (2006) para o desenvolvimento de *wordnets* terminológicas ou terminets (do inglês, *terminological wordnets*). A estratégia de pesquisa de Dias-da-Silva destaca-se por equacionar todo empreendimento no PLN em três fases (a linguística, a representacional e a implementacional) e, sobretudo, evidenciar a importância do conhecimento linguístico nesse tipo de pesquisa;
(b) aplicar a metodologia instanciada para a construção de uma terminet em PB, língua ainda carente de bases lexicais, sejam elas de língua geral ou terminológicas.

Dentre os resultados previstos no TermiNet, estão: (i) aquisição de um arcabouço teórico-metodológico para a construção de um tipo específico de recurso linguístico-computacional (ou seja, *wordnets* terminológicas); (ii) criação de um *corpus*[5] de um domínio especializado para o qual a primeira terminet será construída; (iii) construção de uma base de conhecimento lexical especializado no formato *wordnet*, ou seja, uma terminet, e (iv)

---

[3] Processo 2009/06262-1.

[4] CNPq 471871/2009-5.

[5] "Um corpus é um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise". (Sanchez, 1995, *apud* Berber Sardinha, 2000)

possibilidade de expansão da WN.Br por meio da inclusão de conhecimento especializado.

Para a sua realização, o TermiNet conta com uma equipe interdisciplinar, composta por linguistas e cientistas da computação e conta com os recursos dos laboratórios do Grupo de Estudos e Pesquisas em Terminologia (GETerm)[6] e Núcleo Interinstitucional de Linguística Computacional (NILC)[7]. Além das autoras deste texto, Ariani Di Felippo e Gladis M. de B. Almeida, que atual, respectivamente, como coordenadora e especialista em Terminologia, o projeto TermiNet está sendo desenvolvido com a colaboração dos pesquisadores descritos no Quadro 1.

Equipe		
Pesquisadores	**Função**	**Filiação**
Profa. Dra. Maria das Graças Volpe Nunes	Pesquisadora colaboradora	Instituto de Ciências Matemáticas e de Computação (ICMC)/USP
Profa. Dra. Sandra Maria Aluísio	Pesquisadora colaboradora	Instituto de Ciências Matemáticas e de Computação (ICMC)/USP
Prof. Dr. Thiago Alexandre Salgueiro Pardo	Pesquisador colaborador	Instituto de Ciências Matemáticas e de Computação (ICMC)/USP

Quadro 1: Equipe responsável pelo desenvolvimento do TermiNet.

A Profa. Dra. Maria das Graças Volpe Nunes é uma das fundadoras do NILC e atual coordenadora geral do laboratório. Sua vasta experiência na coordenação de projetos interdisciplinares na área do PLN está sendo de grande valia para a realização do TermiNet. A Profa. Dra. Sandra Maria Aluísio é uma das principais pesquisadoras do Brasil na área de Linguística de Corpus e, por isso, sua participação como colaboradora está sendo fundamental para a projeção e montagem de *corpus*. O Prof. Dr. Thiago A. Salgueiro Pardo tem demonstrado forte interesse pelo desenvolvimento e disponibilização de recursos lexicais computacionais para o processamento automático do PB em suas pesquisas mais recentes e, dessa forma, também tem contribuído para o desenvolvimento dos aspectos computacionais deste projeto.

---

[6] www.geterm.ufscar.br

[7] www.nilc.icmc.usp.br

A seguir, na seção 3, são apresentados os pressupostos teórico-metodológicos que fundamentam o desenvolvimento do TermiNet.

3. Os pressupostos teórico-metodológicos

Especificamente, descrevem-se a metodologia genérica de pesquisa no PLN proposta por Dias-da-Silva (2006) e o formato *wordnet* (Fellbaum, 1998).

3.1 A metodologia genérica de pesquisa no PLN

Para Dias-da-Silva (2006), os sistemas de PLN são vistos como "sistemas especialistas" (do inglês, *expert systems*) ou "sistemas baseados em conhecimento" (do inglês, *knowledge-based systems*). Segundo essa concepção, a construção de um sistema de PLN, ou parte dele, envolve uma "engenharia do conhecimento linguístico", a qual é equacionada em função das etapas previstas por Hayes-Roth (1990) para o desenvolvimento dos sistemas especialistas, a saber: "extração do solo" (isto é, explicitação dos conhecimentos e habilidades), "lapidação" (isto é, representação formal desses conhecimentos e habilidades) e "incrustação" (isto é, o programa de computador que codifica essa representação) (Dias-da-Silva, 1998).

Dias-da-Silva (2006), com base em Hayes-Roth, propõe uma metodologia que decompõe a construção de um sistema, ferramenta (p.ex.: um analisador sintático) ou recurso (p.ex.: as bases de conhecimento lexical) em um conjunto de atividades sucessivas e complementares, agrupadas, segundo sua natureza, em três domínios: o linguístico, o linguístico-computacional (ou representacional) e o implementacional. No domínio linguístico, as atividades ficam concentradas na investigação dos fatos da língua natural em diferentes dimensões (morfológica, sintática, semântico-conceitual e até mesmo pragmático-discursiva) de acordo com a especificidade do sistema, ferramenta ou recurso que se queira desenvolver. No domínio representacional, por sua vez, estudam-se modelos formais de representação para os conhecimentos reunidos no domínio linguístico que sejam tratá-

veis por computador. E, por fim, no domínio implementacional, as atividades ficam concentradas nas questões relativas à implementação do sistema de PLN.

Tal metodologia tem sido aplicada com sucesso na construção de recursos (p.ex.: Maziero *et al.,* 2008; Dias-da-Silva *et al.,* 2008; Di Felippo e Dias-da-Silva, 2008) de PLN em PB.

3.2. A Wordnet de Princeton e o formato *wordnet*

Em meados da década de 1980, os pesquisadores do Laboratório de Ciência Cognitiva da Universidade de Princeton (EUA), impulsionados por pressupostos psicolinguísticos sobre a organização do léxico mental, decidiram construir uma base lexical de língua geral em que as unidades lexicais não se organizariam alfabeticamente (ou seja, em função da forma), mas sim em função do seu significado (Miller e Fellbaum, 1991). Essa iniciativa deu origem, no início da década de 90, à WN.Pr.

Na WN.Pr, as unidades lexicais (palavras ou expressões) do inglês norte-americano 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 (p.ex.: {car; auto; automobile; machine; motorcar}). Cada *synset* é, por definição, construído de modo a representar um único conceito lexicalizado por suas unidades constituintes. Assim, não é preciso explicitar o valor semântico de cada conjunto de sinônimos por meio de um rótulo conceitual. Os *synsets* estão inter-relacionados pela relação léxico-semântica da antonímia[8] e pelas relações semântico-conceituais da hiperonímia/ hiponímia, holonímia/ meronímia, acarretamento e causa.

A WN.Pr também registra outras informações, ditas adicionais, a saber: (a) para cada unidade lexical, há uma frase-exemplo para ilustrar o seu contexto de uso, p.ex.: para car, no *synset*

---

[8] A antonímia é uma relação entre unidades lexicais, ou seja, formas linguísticas. A relação de antonímia entre *synsets* (ou conceitos) indica, na verdade, uma oposição conceitual e não uma antonímia propriamente.

{car; auto; automobile; machine; motorcar}, há a frase-exemplo "he needs a car to get to work" ("ele necessita de um carro para ir trabalhar"); (b) para cada *synset*, há uma glosa que especifica informalmente o conceito por ele lexicalizado, p. ex.: para o *synset* {car; auto; automobile; machine; motorcar}, há a glosa "a motor vehicle with four wheels; usually propelled by an internal combustion engine" ("um veículo com quatro rodas; usualmente impulsionado por um motor de combustão interno"); (c) para cada *synset*, há também a especificação do tipo semântico expresso pelo conceito a ele subjacente; p. ex.: o *synset* {bicycle; bike; wheel; cycle} é do tipo semântico <noun.artifact>.

Como mencionado, na WN.Pr, as unidades lexicais estão organizadas em quatro categorias sintáticas. Cada uma delas constitui uma base lexical própria, em que os *synsets* estão organizados por relações semântico-conceituais específicas, responsáveis pela estruturação interna da base. O Quadro 2, baseado em Fellbaum (1998), resume o conjunto principal de relações em função das categorias sintáticas.

Relações	Categorias sintáticas	Exemplos
Antonímia (oposição conceitual)	Adj, Adv N, V	*mulher* é antônimo de *homem* [9] *claro* é antônimo de *escuro* *rapidamente* é antônimo de *lentamente* *descer* é antônimo de *subir*
Hiponímia/ Hiperonímia (subordinação)	N	*veículo* é hiperônimo de *carro* *carro* é hipônimo de *veículo*
Meronímia/ Holonímia (parte-todo)	N	*carro* é holônimo de *roda* *roda* é merônimo de *carro*
Troponímia (modo)	V	*sussurrar* é tropônimo de *falar*
Acarretamento	V	*correr* acarreta *deslocar-se*
Causa	V	*matar* causa *morrer*
Legenda: N= nome; V= verbo; Adj=adjetivo; Adv=advérbio		

Quadro 2: As relações semânticas da WN.Pr em função das categorias sintáticas.

[9] Na WN.Pr, o *synset* {man, adult male} é considerado antônimo (no caso, "oposto conceitual") do *synset* {woman, adult female}.

A seguir, na seção 4, apresenta-se a instanciação da metodologia genérica de pesquisa no PLN proposta por Dias-da-Silva (2006) para o desenvolvimento de *wordnets* terminológicas.

4. A instanciação da metodologia genérica de pesquisa no PLN

Com base na metodologia genérica de pesquisa no PLN e no formato *wordnet* para bases de dados lexicais, a instanciação da metodologia para a construção de uma terminet fica assim delimitada:

- Domínio linguístico: (i) delimitação do domínio de conhecimento especializado; (ii) delimitação das fontes e da estratégia de aquisição do conhecimento necessário à criação de uma *wordnet* (p. ex.: dicionários, taxonomias, *corpora* etc.), e (iii) delimitação e compilação do conhecimento léxico-conceitual, ou seja, das categorias sintáticas; das unidades lexicais, das relações lexicais de sinonímia e antonímia, das relações semântico-conceituais de hiperonímia/ hiponímia, holonímia/ meronímia, acarretamento e causa, das glosas e das frases-exemplo;
- Domínio representacional: representação do conhecimento delimitado no domínio linguístico em um formalismo que seja "computacionalmente tratável"; no caso de uma base wordnet, tal representação baseia-se na noção de *synset* e de matriz lexical;
- Domínio implementacional: transformação da representação do conhecimento linguístico em uma base lexical propriamente dita.

4.1. As tarefas do domínio linguístico e os meios para a sua realização

4.1.1. A delimitação do domínio especializado

Quando se planeja realizar um projeto terminológico, deve-se delimitar o domínio, evitando-se eleger como objeto da pesquisa uma área completa, pois em geral há desdobramentos em vários outros níveis cada vez mais específicos. Em razão disso,

cada uma das especificidades que compõem cada nível podem ser distintas no que se refere a abordagens teóricas, métodos, setores de aplicação etc. Para dar um exemplo mais próximo, imagine-se a dificuldade de sistematizar a terminologia da Linguística.

Segundo Almeida e Correia (2008), lidar com uma área como um todo pode revelar-se contraproducente por pelo menos duas razões:

(a) via de regra, as áreas se compõem de subáreas com distintas especificidades, o que evidentemente gera um universo muito grande de fontes de obtenção dos textos que deverão compor o *corpus*. Além disso, há grande possibilidade de dispersão, que pode resultar em um problema no momento da extração dos candidatos a termos, pois com essa profusão de especificidades temáticas, corre-se o risco de deixar de considerar termos relevantes para determinada especialidade e fazer o inverso com outra, ou seja, acabar privilegiando uma em detrimento de outra;
(b) torna-se necessário contar com uma assessoria especializada muito maior, o que dificulta o trabalho.

As autoras apontam alguns fatores que podem auxiliar na delimitação do domínio:

(a) interesse dos especialistas do domínio em ter sua terminologia sistematizada e organizada num produto terminológico (redes semânticas, glossário, dicionário, ontologia etc.);
(b) número de profissionais colaboradores com os quais se poderá contar;
(c) relevância de determinada especificidade do ponto de vista educacional, social, político, econômico, científico e/ou tecnológico para o país;
(d) facilidade de obtenção de textos já em formato digital para agilizar a compilação do *corpus*.

4.1.2. A delimitação das fontes para a compilação do conhecimento léxico-conceitual

Segundo os pressupostos gerais da Teoria Comunicativa da Terminologia (Cabré, 1999; 2003), os termos (isto é, os signos

que ocorrem como unidades terminológicas) e suas propriedades só podem ser identificados e descritos no seu ambiente natural de ocorrência, ou seja, nos discursos especializados. Dessa forma, esses princípios teóricos e metodológicos põem em evidência a importância do uso dos *corpora* (fontes não-estruturadas) em qualquer trabalho terminológico (Nascimento, 2003; Agbago, Barrière, 2005; Cabré et al., 2005; Almeida, 2006).

De acordo com Nascimento (2003), Barros (2004) e Cabré *et al.* (2005), a partir de *corpora*, pode-se fazer observações precisas sobre o real comportamento linguístico de gente real, proporcionando informações altamente confiáveis e isentas de opiniões e de julgamentos prévios sobre os fatos de uma língua. Por meio de *corpus*, é possível observar aspectos morfológicos, sintáticos, discursivos etc. relevantes para uma pesquisa linguística. É possível descobrir fatos novos na língua, não perceptíveis pela intuição.

Assim, para a construção de uma wordnet terminológica, os *corpora* constituem a principal fonte da qual o conhecimento léxico-conceitual deve ser extraído. Naturalmente, os recursos especializados ditos estruturados (p.ex.: dicionários, taxonomias, ontologias etc.), sejam eles impressos ou em formato eletrônico, também podem ser utilizados como fontes, caso existam e/ou estejam disponíveis.

Com base nos pressupostos da Linguística de Corpus, a construção do *corpus* deve seguir quatro três etapas: (a) projeção do *corpus*, que consiste na definição do tipo de *corpus* necessário à pesquisa; (b) compilação dos textos que comporão o *corpus*; (c) pré-processamento, que consiste nas tarefas de conversão, limpeza, nomeação e anotação dos textos compilados; (d) a aquisição das permissões de uso (caso 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 (Giouli e Peperidis, 2002).

Assim, para servir de base à construção de uma terminet, um *corpus* precisa, de início, apresentar certas características: (i) ser monolíngue; (ii) ser relativo a um domínio especializado e

proporcionar a descrição sincrônica do léxico temático desse domínio, e (iii) conter textos escritos, ou seja, textos da modalidade escrita da língua registrados em meio escrito (*vs* arquivos em formato de áudio), pois as bases *wordnets* são recursos para o tratamento computacional das línguas naturais registradas em tal meio.

Certas decisões de projeto também determinam propriedades específicas do *corpus*. No caso, as informações léxico-conceituais (ou seja, os termos e as relações léxico-semânticas e semântico-conceituais) necessárias à construção de uma *wordnet* terminológica são comumente obtidas por meio de métodos semiautomáticos de extração a partir de *corpora*. Alguns desses métodos baseiam-se no reconhecimento de padrões léxico-sintáticos, o que requer a anotação morfossintática do *corpus*. Outra decisão de projeto diz respeito à disponibilização. Como os *corpora* especializados são recursos extremamente úteis e de construção cara, é desejável que estes sejam disponibilizados 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 (Nascimento, 2003).

Com base nessa caracterização inicial, tem-se uma projeção parcial do tipo de *corpus* necessário à construção de uma terminet, a qual é apresentada no Quadro 3.

Critérios	**Características**
Modalidade	Escrito
Cobertura da língua	Especializado
Quantidade de línguas	Monolíngue
Anotação	Anotado (nível morfossintático)
Mutabilidade	Aberto
Variações históricas	Sincrônico
Disponibilidade	Disponível via *Web*

Quadro 3: Projeção inicial do *corpus*

Além dessa caracterização inicial, certos requisitos precisam ser atendidos para que uma coleção de textos possa ser denominada *corpus*, como autenticidade, representatividade, amostragem, balanceamento, diversidade e tamanho (Kennedy, 1998; Biber *et al.,* 1998; Renouf, 1998; Berber Sardinha, 2000, 2004 e Sinclair, 2005). Consequentemente, a etapa de projeção do *corpus* engloba a discussão de tais requisitos e a identificação de possíveis estratégias para atendê-los.

Para a coleta ou compilação dos textos, tem-se optado preferencialmente por material disponível na *web* devido ao custoso trabalho de digitalização de material impresso. Além disso, essa preferência justifica-se pelo fato de a *web* ser uma mina de dados linguísticos de riqueza e acessibilidade sem precedentes (Kilgarriff e Grefenstette, 2003). Para tal coleta, duas abordagens são comumente aplicadas: a manual, que consiste na seleção manual de páginas e documentos na *web,* e a automática, que consiste na utilização de certas ferramentas computacionais que captam automaticamente material *on line*. Exemplos paradigmáticos de tais ferramenta são o BootCaT (do inglês, *Bootstrapping Corpora and Terms*) (Baroni e Bernadini, 2004), um extrator automático de *corpus* (e de termos), e o Corpógrafo (Sarmento *et al.,* 2004), um ambiente Web que possibilita a compilação automática e investigação de *corpora* especializados.

Após a compilação, o *corpus* precisa ser preparado para que possa receber um tratamento ou processamento computacional. A preparação ou pré-processamento engloba os processos de (i) conversão manual e/ou automática dos textos nos formatos doc, pdf e html para o formato txt, (ii) limpeza manual dos dados corrompidos pela conversão; (iii) nomeação padronizada dos arquivos, anotação estrutural dos textos e geração de cabeçalho. Os processos descritos em (iii) são comumente realizados por uma ferramenta computacional denominada "editor de cabeçalho". Para o pré-processamento do *corpus* em PB, algumas ferramentas estão disponíveis. Para as etapas de nomeação padronizada dos arquivos, anotação estrutural dos textos e geração de cabeçalho, tem-se o editor de cabeçalho do projeto Lácio-Web (Aluísio *et al.,* 2004).

Finalmente, o *corpus* precisa passar por um processo de anotação morfossintática para que os métodos de extração de conhecimento léxico-conceitual possam ser aplicados. O processo de anotação morfossintática, que consiste em atribuir etiquetas de classes gramaticais (do inglês, *part-of-speech tags*) aos elementos dos textos, também recebe o nome de "etiquetação" (do inglês, *tagging*). O método mais eficiente de anotação é o semiautomático, que consiste na utilização de uma ferramenta computacional denominada "etiquetador" (do inglês, *tagger)* e na posterior revisão por humanos dos dados gerados pelo etiquetador. A anotação morfossintática do *corpus*, em particular, é essencial para a aplicação (i) das abordagens linguística e híbrida de extração de termos e (ii) do método linguístico (baseado em padrões léxico-sintáticos) de extração das relações lexicais e semântico-conceituais.

Para a anotação morfossintática de *corpora* em PB, tem-se o pacote de etiquetadores composto pelo MXPOST (Ratnaparkhi, 1996), TreeTagger (Schmid, 1994) e BRILL (Brill, 1995), além do etiquetador do *parser* PALAVRAS (Bick, 2000).

4.1.3. A delimitação e compilação do conhecimento léxico-conceitual

Tomando-se como base a WN.Pr, uma base lexical no formato *wordnet* define-se por armazenar as unidades da língua (palavras ou expressões) organizadas em função da sinonímia e de certas relações conceituais. Assim, na metodologia de construção de uma terminet, estão previstas as seguintes tarefas na etapa de delimitação e compilação do conhecimento léxico-conceitual: (i) delimitação das categorias sintáticas, (ii) compilação dos termos, (iii) identificação da sinonímia e a montagem dos *synsets* e (iv) delimitação e identificação das relações internas às terminets.

a) A delimitação das categorias sintáticas

Como mencionado, na WN.Pr as unidades lexicais estão organizadas em quatro categorias sintáticas: verbos, nomes, adjetivos e advérbios. Tendo em vista a proeminência das uni-

dades da categoria dos nomes na organização das terminologias, ou seja, dos conjuntos de termos das áreas especializadas, restringe-se a construção de uma terminet a tal categoria. Em outras palavras, uma terminet armazenará, em princípio, apenas unidades terminológicas da categoria dos nomes.

b) A compilação dos termos ou unidades terminológicas

Apesar de sua centralidade nas pesquisas terminológicas, a noção de "termo" ainda não é totalmente clara, tanto do ponto de vista linguístico quanto computacional. De acordo com duas obras que têm regulamentado a pesquisa terminológica em vários países, a saber: *Terminology work – Vocabulary – Part 1: Theory and application*, ISO 1087, e *Vocabulaire systématique de la terminologie*, termo é definido como sendo a "designação de um conceito numa língua de especialidade por meio de uma expressão linguística." (trad. nossa) e uma "unidade significante constituída de uma palavra (termo simples) ou de mais de uma palavra (termo complexo) e que designa um conceito de maneira unívoca no interior de um domínio de especialidade." (trad. nossa), respectivamente.

Por essas definições, percebe-se que o aspecto formal é o critério levado em conta, já que ambas se utilizam de unidades léxicas tais como *expressão linguística* e *unidade significante constituída de uma palavra ou de várias palavras*. Se o critério formal fosse suficiente, não haveria equívocos na identificação de termos em *corpus*, pois de imediato seria possível reconhecer marcas formais, principalmente no que concerne aos níveis morfológico e lexical. Essa facilidade se observa quando se está diante de uma formação marcadamente técnico-científica, como as que utilizam morfemas greco-latinos, posto que o nível morfológico já é suficiente para indicar que se trata de um termo e não de uma palavra. Infelizmente isso não é possível com a grande maioria dos termos originários da língua geral, termos esses que não têm marcas formais para facilitar a sua recolha em textos especializados, como por exemplo: *forno, secador, peneira, biscoito,* unidades da terminologia de Revestimento Cerâmico. Isso ilustra as dificuldades em identificar termo utilizando critérios estritamente formais.

Que critérios devem ser levados em conta para distinguir um termo de uma palavra, já que a partir de uma perspectiva linguística todos são igualmente signos da língua natural? Não existe, pois, um conjunto de termos isolados constituindo uma língua marginal à língua geral; o que há são signos da língua natural que se realizam ora como palavras, ora como termos, dependendo da temática, dos usuários, da situação comunicativa (Cabré, 1999; 2003). O que distingue, portanto, termo de palavra são critérios pragmáticos. Em outras palavras: quem diz o quê? Para quem? Em que situação? Se termo é assim concebido, então a sua identificação deve sempre ser feita nos contextos de uso; isso implica necessariamente a elaboração de um *corpus*, de maneira que seja possível observar os termos *in vivo* (Bessé, 1997). Em razão disso, a extração de candidatos a termos, mesmo sendo automática, nunca é uma tarefa fácil.

A extração automática de termos (EAT) diz respeito ao processo de obtenção computacional (isto é, por meio de uma ferramenta computacional denominada "extrator de termos"), a partir de *corpus*, de um conjunto de unidades terminológicas. No caso, essas unidades comporão os *synsets* da terminet. Na literatura, existem três abordagens de extração (Cabré *et al.,* 2001; Jacquemin e Bourigault, 2004; Pazienza *et al.,* 2005; Bernhard, 2006): (i) abordagem linguística; (ii) abordagem estatística; (iii) abordagem híbrida.

A abordagem linguística busca identificar os candidatos a termos por meio da aplicação automática de "filtros linguísticos" (Pazienza *et al.,* 2005; Bernhard, 2006). Dentre esses filtros, destacam-se os padrões morfossintáticos (p.ex.: [n-n] e [adj-n]), obtidos em *corpora* morfossintaticamente etiquetados, e os padrões léxico-sintáticos do tipo "é um tipo de", "caracterizado como" etc. Com base nesses filtros, um extrator de termos é capaz de identificar e extrair os candidatos a termo (Cabré *et al.,* 2001). A identificação dos filtros linguísticos, que são dependentes do domínio (e até mesmo de gênero), requer uma análise prévia (manual) do *corpus*.

A abordagem estatística baseia-se na aplicação de medidas estatísticas como frequência, informação mútua, *log-likelihood ratio* e coeficiente Dice. Tais medidas podem ser apli-

cadas por meio da utilização do pacote estatístico NSP (do inglês, *N-gram Statistics Package*), que realiza a análise de n-gramas (ou seja, sequência de elementos do texto). Por fim, na abordagem híbrida, o processo de extração é feito em duas etapas. Na primeira, extraem-se os candidatos por meio da aplicação de filtros linguísticos, resultando em uma lista de candidatos. Na segunda etapa, aplica-se uma métrica estatística (ou mais) à lista obtida na primeira etapa com o objetivo de ranquear os membros da lista.

Para a extração de candidatos a termo, a utilização de ferramentas como o BootCat (Baroni e Bernadini, 2004) e o Corpógrafo (Sarmento *et al.,* 2005) também deve ser considerada.

c) A identificação da sinonímia e a montagem dos *synsets* preliminares

A relação léxico-semântica de sinonímia pode ser automaticamente extraída do *corpus* por meio de abordagens estatísticas ou linguísticas.

Os trabalhos que utilizam a abordagem estatística assumem, com base em Harris (1968), que, quanto maior a similaridade distribucional entre as unidades lexicais, maior é a probabilidade de essas unidades serem sinônimas. Nessa linha, citam-se, por exemplo, os trabalhos de Church e Hanks (1990) e Lin (1998), aplicados a textos em inglês. Embora o método estatístico seja bastante robusto, pois não necessita da análise manual dos dados obtidos, ele somente funciona quando aplicado a *corpora* realmente extensos (10 milhões de palavras).

Os trabalhos que utilizam a abordagem linguística baseiam-se na identificação dos padrões léxico-sintáticos (p. ex.: Hearst, 1992; 1998), também denominados "marcadores relacionais" (Condamines, 2002). Especificamente, buscam-se identificar, nesses trabalhos, os vários padrões sintáticos e lexicais por meio dos quais certas relações semânticas são superficialmente expressas na língua (Suárez e Cabré, 2002; Nenadic *et al.,* 2004) etc.

Para a identificação da relação de sinonímia, em especial, são poucos os trabalhos que buscam identificar tais padrões. Dentre eles, citam-se, por exemplo, os de Feliu e Cabré (2002), Agbago e Barrière (2005) e Mitilelu (2006).

Para a montagem efetiva dos *synsets* preliminares, deve-se considerar o teste da substituição e a noção de "sinonímia contextual". Segundo a noção de sinonímia contextual, "duas unidades lexicais são sinônimas em um contexto C, se a substituição de uma pela outra em C não altera o valor de verdade de denotado por C" (Cruse, 2004; Miller e Fellbaum, 1991). Caso isso ocorra, tais unidades constituem um *synset*. Descrições mais precisas sobre esse e outros testes podem ser encontradas em Vossen (2002).

A validação dos *synsets* preliminares, assim como dos termos candidatos que os constituem, deve ser feita por um ou mais especialistas do domínio. Somente após a validação dos termos e dos *synsets*, as relações semântico-conceituais internas a uma terminet devem ser identificadas, bem como as glosas e as frases-exemplo.

d) A delimitação e identificação das relações internas às terminets

Essa etapa consiste na identificação no *corpus* das relações semântico-conceituais responsáveis pela estruturação interna da base. Tendo em vista que as unidades terminológicas a serem armazenadas em uma terminet pertencem à categoria dos nomes, as relações semântico-conceituais restringem-se à hiponímia e à meronímia.

Para a identificação e extração da relação da hiponímia, em particular, vários trabalhos (p.ex.: Cederberg e Widdows, 2003; Morin e Jacquemin, 2004 e Mititelu, 2006) têm aplicado a abordagem linguística que se baseia na identificação dos padrões léxico-sintáticos. Hearst (1992) identificou seis pistas textuais para a identificação da relação de hiponímia em textos de língua inglesa. Dentre elas, cita-se, por exemplo: {NP0 such as NP1}, que, em português, pode ser traduzida para {SN0 tais como | como SN1 (SN2,...)} (p.ex.: bactérias *como* a salmonella e a shighella).

Para a identificação e extração da relação da meronímia, padrões léxico-sintáticos também têm sido utilizados. Nos trabalhos de Feliu e Cabré (2002) e Agbago e Barrière (2005), por exemplo, algumas pistas textuais, como *is composed of* (é com-

posto de) e *is a part of* (é parte de), são apresentadas como resultado da análise de textos em espanhol e em inglês, respectivamente. O Corpógrafo, aliás, fornece recursos para que as relações possam ser extraídas por padrões definidos pelos seus próprios usuários.

As relações semântico-conceituais obtidas de *corpus* também podem ser verificadas pela aplicação de testes de substituição, como elucidado por Cruse (2004) e Vossen (2002), e validadas pelos especialistas de domínio.

Ressalta-se, por fim, que a hiponímia e a meronímia são relações organizadas hierarquicamente. Para a organização da hierarquia de conceitos, dois métodos podem ser adotados: (i) o *top-down*, em que se identificam os conceitos genéricos e, em seguida, os conceitos específicos e (ii) o *bottom-up*, em que se identificam os conceitos específicos e, em seguida, os conceitos genéricos (Uschold e Gruninger, 1996).

Além disso, a organização dos conceitos pode ser feita por meio de uma hierarquia única ou múltipla. Na primeira, a organização hierárquica dos conceitos expressos por nomes é feita a partir de um único conceito genérico do tipo {entidade}, a partir do qual os conceitos mais específicos devem ser organizados. Na segunda estratégia, adotada, aliás, na construção as WN.Pr (Fellbaum, 1998), os conceitos organizam-se a partir de um conjunto de conceitos (menos) genéricos, sendo que cada um dos conceitos do conjunto inicia uma hierarquia própria. A esses conceitos (menos) genéricos, é dada a denominação "iniciadores únicos" (do inglês, *unique beginners*).

A organização dos conceitos segundo o método *top-down* e a noção de hierarquia múltipla pode ser beneficiada com a elaboração de um "mapa conceitual" do domínio cujo conhecimento se quer sistematizar. O mapa conceitual constitui uma organização semântica da área-objeto, semelhante ao que se entende por árvore de domínio; a diferença é que os conceitos/termos estão ali armazenados em seus respectivos campos semânticos. Ressalte-se que, além dos conceitos, devem também estar representadas no mapa as relações entre eles. Em uma pesquisa terminológica, o mapa conceitual é fundamental para: (i) possibilitar uma abordagem mais sistemática de um campo de espe-

cialidade; (ii) circunscrever a pesquisa, já que todas as ramificações da área-objeto, com seus campos, foram previamente consideradas; (iii) delimitar o conjunto terminológico; (iv) determinar a pertinência dos termos, pois separando cada grupo de termos pertencente a um determinado campo, poder-se-á apontar quais termos são relevantes para o trabalho e quais não são; (v) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas; (vi) definir as unidades terminológicas de maneira sistemática e, finalmente; (vii) controlar a rede de remissivas (Almeida, 2000; Almeida *et al.,* 2007). Esse mapa, elaborado por terminólogos e especialistas do domínio, fornece uma visão geral da área-objeto (Almeida, 2006), podendo, assim, proporcionar o que Fellbaum (1998) denominou "iniciadores únicos".

e) A seleção das frases-exemplo e elaboração das glosas

As frases-exemplo, que fornecem os contextos de uso mínimos para cada unidade de um *synset*, são comumente extraídas de *corpus* por um concordanciador, ou seja, uma ferramenta que lista na tela todas as ocorrências de uma palavra ou expressão no *corpus,* acompanhadas do texto ao seu redor (o cotexto). A ferramenta Concord, que integra o pacote de ferramentas de análise de *corpus* WordSmith Tools (Scott, 1999), é um exemplo de concordanciador. As glosas, apesar de definições informais dos conceitos subjacentes aos *synsets*, devem ser elaboradas com base no contexto temático em que as unidades terminológicas do *synset* estão inseridas. Assim, a busca pelos contextos explicativos e/ou defintitórios é relevante para a elaboração das glosas.

4.2. As tarefas do domínio representacional e os meios para a sua realização

O formato *wordnet* fundamenta-se em três construtos formais (Fellbaum, 1998):

(i) *o método diferencial*: método segundo o qual os conceitos são ativados na mente por meio de formas lexicais sinônimas, elimi-

nando a necessidade de determinar o valor semântico das unidades;

(ii) *os synsets*: conjuntos de formas lexicais determinados pela relação de pertença e munidos de dois tipos de ponteiros, os que especificam relações lexicais (entre formas linguísticas) e os que especificam relações entre conceitos (*synsets*); por exemplo, o ponteiro '!→' entre as unidades *wet* ("molhado") e *dry* ("seco") em wet!→dry indica a relação lexical de antonímia e o ponteiro '@→ entre os *synsets* {jeep, landrover} ("jipe") e {car, auto, automobile, machine, motorcar} ("carro") em {jeep, landrover }@→{car, auto, automobile, machine, motorcar} indica a relação conceitual de hiponímia ("é um tipo de").

(iii) *a noção de matriz lexical*: construto em cuja base a relação entre forma (unidade da língua) e conceito é estabelecida (Quadro 4) e segundo o qual uma base wordnet é construída. Segundo esse construto, cada unidade da língua (F) é descrita numa coluna e cada conceito lexicalizado (C)) é apresentado numa linha da matriz. O preenchimento de uma célula da matriz (p. ex.: F4*S2) implica que a unidade naquela coluna (F4) representa o conceito naquela linha (C2) e, por isso, essa unidade compõe o *synset* que codifica o conceito em questão (no caso, {frump;dog}). Se há duas células preenchidas na mesma coluna, a unidade em questão é polissêmica (p.ex.: F1) e, se há duas células na mesma linha (F1*C2 e F4*C2), as unidades são sinônimas (F1 e F4).

Conceitos lexicalizados (*Synsets*)	**FORMAS/ UNIDADES LEXICAIS**			
	F1 ***dog***	**F2** ***domestic dog***	**F3** ***Canis familiaris***	**F4** **frump**
C1 **{dog; domestic dog; Canis familiaris}**	**F1*C1**	**F2*C1**	**F3*C1**	
C2 **{frump;dog}**	**F1*C2**			**F4*C2**

Quadro 4: Ilustração na noção de matriz lexical

A montagem das bases *wordnets*, sejam elas de língua geral ou terminológicas, é comumente feita por meio de um processo "assistido por computador", ou seja, pela utilização de uma

ferramenta computacional que se fundamenta nos três construtos descritos. Tal ferramenta remete a pesquisa às atividades do domínio implementacional.

4.3. As tarefas do domínio implementacional e os meios para a sua realização

Nesse domínio, duas tarefas são previstas. Abordaremos ambas separadamente.

4.3.1. A especificação de uma ferramenta computacional ou editor

Essa tarefa, eminentemente computacional, consiste na seleção de uma ferramenta computacional para a montagem da terminet. Essa ferramenta deve desempenhar duas funções distintas: (i) a de editor, possibilitando ao linguista a inserção do conhecimento léxico-conceitual previsto pelo formato *wordnet*, e (ii) a de sistema de gerenciamento de dados, pela qual a ferramenta armazena o conhecimento léxico-conceitual no formato *wordnet*, gerando uma base do tipo relacional.

No Projeto TermiNet, investigar-se-á a possibilidade de utilização da ferramenta denominada VisDic (Horák *et al,* 2004). Essa ferramenta, originalmente proposta no âmbito do projeto de construção da rede multilíngue BalkaNet, é um software munido de uma interface gráfica que permite especificamente a montagem de bases no formato *wordnet*. A principal vantagem do VisDic reside na utilização da linguagem de marcação XML[10]. Uma vez nesse formato, uma rede *wordnet* pode ser exportada e utilizada em várias aplicações, por exemplo, pelos sistemas de PLN. Caso necessário, uma ferramenta desse tipo poderá ser desenvolvida no âmbito do projeto.

---

[10] XML (do inglês, *eXtensible Markup Language*) é uma linguagem padronizada de marcação capaz de descrever diversos tipos de dados; seu propósito principal é a facilidade de compartilhamento de informações através da *web*.

4.3.2. A inserção das informações no editor

Essa tarefa concentra-se na: inserção dos termos, montagem concreta dos *synsets*, especificação das relações semântico conceituais e inserção das frases-exemplo e das glosas. Em outras palavras, essa fase consiste efetivamente na construção concreta da base.

5. Considerações finais

De um modo geral, acredita-se que o projeto TermiNet fornece uma metodologia suficientemente clara e genérica para a construção de bases terminológicas no formato *wordnet.* Essa metodologia, no entanto, precisa ser validada, o que será feito, ainda no âmbito do projeto, por meio da construção de uma terminet em PB. A base terminet resultante da validação da metodologia poderá beneficiar não só o PLN, mas a própria construção de produtos terminológicos/terminográficos "tradicionais", pois o equacionamento ou sistematização do conhecimento léxico-conceitual é etapa fundamental na construção desses produtos.

O projeto TermiNet também prevê, como tarefa adicional do domínio implementacional, a possibilidade de avaliação de uma base terminet, que pode ser por meio da abordagem intrínseca ou extrínseca.

No PLN, a avaliação intrínseca avalia o desempenho do sistema computacional pela verificação da qualidade dos dados que produz. Para tanto, são usadas métricas calculadas automaticamente ou julgamentos subjetivos, realizados por leitores humanos. A avaliação extrínseca verifica a adequação do sistema ao seu uso em tarefas específicas; por essa razão, ela é comumente chamada de validação. Uma terminet, entendida como parte de um sistema de PLN, pode ser avaliada pelas mesmas abordagens aplicadas à avaliação dos próprios sistemas de PLN.

Especificamente, a avaliação intrínseca de uma terminet pode ser entendida como a própria validação do conhecimento léxico-conceitual feita pelos especialistas ao longo da construção da base. Já a avaliação extrínseca ou validação pode ser

feita pela utilização da base em alguma aplicação de PLN, como recuperação de informação ou outra.

Por fim, ressalte-se que, no âmbito do projeto TermiNet, os recursos (*corpus* e terminet) construídos serão disponibilizados na *web*, pois a visibilidade das línguas no mundo depende crucialmente do peso das suas tecnologias linguísticas, em particular das de livre acesso na *Web*.

Agradecimento

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro.

Referências bibliográficas

AGBAGO, A.; BARRIÈRE, C. (2005) Corpus construction for Terminology. *Proceedings of the Corpus Linguistics Conference*, Birmingham, pp. 14-17.

ALMEIDA, G. B. A. (2000) *Teoria Comunicativa da Terminologia: uma aplicação*. Tese de Doutorado em Linguística e Língua Portuguesa. São Paulo/Araraquara: Universidade Estadual Paulista (UNESP).

ALMEIDA, G. M. B. (2006) A Teoria Comunicativa da Terminologia e a sua prática. *Alfa*, vol. 50, pp. 81-97.

ALMEIDA, G. M. B.; ALUÍSIO, S. M.; OLIVEIRA, L. H. M. (2007) O método em Terminologia: revendo alguns procedimentos. In: ISQUERDO, A. N.; ALVES, I. M. (org.) *Ciências do léxico: lexicologia, lexicografia, terminologia*. Campo Grande/São Paulo: Editora da UFMS/ Humanitas, 1ª ed., vol. III, pp. 409-420.

ALMEIDA, G. M. B.; CORREIA, M. (2008) Terminologia e corpus: relações, métodos e recursos. In: TAGNIN, S. E. O.; VALE, O. A. (org.) *Avanços da Lingüística de Corpus no Brasil*. São Paulo: Humanitas, 1ª ed., vol. 1, pp. 63-93.

ALUISIO, S.; PINHEIRO, G. M.; MANFRIM, A. M. P.; OLIVEIRA, L. H. M. de; GENOVES Jr., L. C.; TAGNIN, S. E. O. (2004) The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. *Proceedings of the 4th International conference on language resources and evaluation* (LREC). Portugal: Lisboa, pp. 1779-1782.

BARONI, M.; BERNARDINI, S. (2004) BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of the 4th International conference on language resources and evaluation* (LREC). Portugal: Lisboa, pp. 1313-1316.

BARROS, L. A. (2004) *Curso básico de Terminologia*. São Paulo: EDUSP.

BENTIVOGLI, L.; BOCCO, A.; PIANTA, E. (2004) ArchiWordnet: integrating Wordnet with domain-specific knowledge. *Proceedings of the 2nd International Global Wordnet Conference*. Brno: Masaryk University, pp. 39-47. Disponível em: <http://www.fi.muni.cz/gwc2004/proc/101.pdf>. Acesso em 16 de julho de 2010.

BERNHARD, D. (2006) Multilingual term extraction from domain-specific corpora using orphological Structure. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (ACL). Itália: Trento, pp. 171-174.

BIBER, D.; CONRAD, S.; REPPEN, R. (1998) *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

BICK, E. (2000) *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. PhD Thesis. Aarhus University.

BRILL, E. (1995) Transformation-based error-driven learning of natural language: a case study in part of speech tagging. *Computational Linguistics*, vol. 21, pp. 543-565.

BUITELAAR, P.; SACALEANU, B. (2002) Extending synsets with medical terms. *Proceedings of the 1st International Global Wordnet Conference*. India: Mysore, pp. 1-6.

CABRÉ, M. T. (1999) *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Linguística Aplicada.

____ (2003) Theories of terminology: their description, prescription and explanation. *Terminology*, vol. 9(2), pp. 163-200.

CABRÉ, M. T.; ESTOPÀ, R.; PALATRESI, J. V. (2001) Automatic term detection: a review of current systems. In: BOURIGAULT, D. et al. (eds.) *Recent Advances in Computational Terminology*. Amsterdam & Philadelphia: John Benjamins Publishing Co., pp. 53-87.

CABRÉ, M. T.; COMDAMINES, A.; IBEKWE-SANJUAN, F. (eds.) (2005) Application-driven terminology engineering. *Terminology*, vol. 11(2), pp. 1-19.

CEDERBERG, S.; WIDDOWS, D. (2003) Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. *Processings of the 11th Conference on Computational Natural Language Learning*. Canadá: Edmonton, pp. 111-118.

CHURCH, K. W; HANKS, P. (1990) Word association norms, mutual information and lexicography. *Computational Linguistics*, vol. 16(1), pp. 22–29.

CONDAMINES, A. (2002) Corpus analysis and conceptual relation patterns. *Terminology*, vol. 8(1), pp. 141-162.

CRUSE, A. (2004) *Meaning in language: an introduction to semantics and pragmatics*. Oxford: Oxford University Press.

BESSÉ, B. de (1997) Terminological Definitions. *Handbook of Terminology Management*. Amsterdam: John Benjamins, pp. 63-74.

DIAS-DA-SILVA, B. C. (1998) Bridging the gap between linguistic theory and natural language processing. *Proceedings of the 16th International Congress of Linguistics*. Paris: France, pp. 1-10.

____ (2006) O estudo linguístico-computacional da linguagem. *Letras de Hoje,* vol. 41(2), pp. 103-138.

DIAS-DA-SILVA, B. C.; DI FELIPPO, A.; NUNES, M. G. V. (2008) The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC). Marrocos: Marrakech, pp. 335-342.

DI-FELIPPO, A.; DIAS-DA-SILVA, B. C. (2008) REBECA: uma base de dados léxico-conceituais bilíngue inglês-português. *Proceedings of the 4th Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence* (WTDIA). Bahia: Salvador, pp. 1-10.

FELLBAUM, C. (ed.) (1998) *Wordnet*: *an electronic lexical database*. Ca, MA: MIT Press.

FELIU, J.; CABRÉ, M. T. (2002) Conceptual relations in specialized texts: new typology and an extraction system proposal. *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering*. França: Nancy, pp. 45-49.

GANGEMI, A.; SAGRI, M. T.; TISCORNIA, D. (2003) Jur-wordnet, a source of metadata for content description in legal information. *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*. Prague, pp. 1-6.

GIOULI, V.; PIPERIDIS, S. (2002) *Corpora and HLT*: *current trends in corpus processing and annotation*. Bulgaria: Institute for Language and Speech Processing. Disponível em: <http://www.larflast.bas.bg/balric/eng_files/corpora1.php>. Acesso em 16 de julho de 2010.

HANKS, P. (2004) Lexicography. In: MITKOV, R. (ed.). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press, pp. 48-69.

HARRIS, Z. S. (1968) *Mathematical Structures of Language*. New York: John Wiley & Sons.

HAYES-ROTH, F. (1990) Expert systems. In: SHAPIRO, E. (ed.). *Encyclopedia of artificial intelligence*. New York: Wiley, pp. 287-298.

HEARST, M. (1992) Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, pp. 539-545.

____ (1998) Automated discovery of Wordnet relations. In: FELLBAUM, C. (ed.) *Wordnet: an electronic lexical database*. Cambridge, MA: MIT Press, pp. 131-152.

HORAK, A.; SMRZ, P. (2004) VisDic: Wordnet browsing and editing tool. *Proceedings of the 2nd International Global Wordnet Conference*. Brno: Masaryk University, pp. 136-141.

ISO 1087 (2000) *Terminology work – Vocabulary – Part 1: Theory and application*, Genebra (ISO/1087-1)

JACQUEMIN, C.; BOURIGAULT, D. (2004) Term extraction and automatic indexing. In: MITKOV, R. (ed.) *Handbook of Computational Linguistics*. Oxford University Press, pp. 599-615.

KENNEDY, G. (1998) *An introduction to corpus linguistics*. London: Longman.

KILGARRIFF, A.; GREFENSTETTE, G. (2003) Introduction to the special issue on the Web as Corpus. *Computational Linguistics*, vol. 29, p. 333347.

LIN, D. (1998) Automatic retrieval and clustering of similar words. In: *Proceedings of the Joint International Conference on Computational Linguistics*. Canadá: Montréal, pp. 768-773.

MAGNINI, B.; SPERANZA, M. (2001) Integrating generic and specialized wordnets. *Proceedings of the 2nd Conference on Recent Advances in Natural Language Processing*. Bulgária: Tzigov Chark, pp. 149-153.

SARMENTO, L.; MAIA, B.; SANTOS, D. (2004) The Corpógrafo: a Web-based environment for corpora research. *Proceedings of the 4th In-*

ternational conference on language resources and evaluation (LREC), Lisboa, pp. 449-52.

MAZIERO, E. G. et al. (2008) A base de dados lexical e a interface web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil. *Proceedings of the 6th Workshop in Information and Human Language Technology*. Vila Velha-ES, pp. 390-392.

MILLER, C.; FELLBAUM, C. (1991) Semantic networks of English. *Cognition*, vol. 41, pp. 197–229.

MITITELU, V. B. (2006) Automatic extraction of patterns displaying hyponym-hypernym co-cccurrence from corpora. *Proceedings of the 1st Central European Student Conference in Linguistics*. Hungria: Budapest, pp. 1-8.

MITKOV, R. (ed.) (2004) *The Oxford handbook of computational linguistics*. New York: Oxford University Express.

MORATO, J. et al. (2004) Wordnet applications. *Proceedings of the 2nd International Global Wordnet Conference.* Brno, Masaryk University, pp. 270-278.

MORIN, E.; JACQUEMIN, C. (2004) Automatic acquisition and expansion of hypernym links. *Computer and the Humanities*, vol. 38 (4), pp. 343-362l.

NASCIMENTO, M. F. B. (2003) O papel dos corpora especializados na criação de bases terminológicas. CASTRO, I.; DUARTE, I. (orgs.). *Razões e emoções, miscelânea de estudos em homenagem a Maria Helena Mateus*. Lisboa: Imprensa Nacional-Casa da Moeda, vol. II, pp. 167-179.

NENADIC, G. *et al.* (2004) Mining term similarities from corpora. *Terminology*, vol. 10(1), pp. 55-81.

OFFICE DE LA LANGUE FRANÇAISE (1985) *Vocabulaire systématique de la terminologie*, Québec.

PALMER, M. (2001) Multilingual resources, multilingual information management: current levels and future abilities. *Linguistica Computazionale*, vol. XIV-XV, pp. 1-33.

PAZIENZA, M. T. *et al.* (2005) Terminology extraction: an analysis of linguistic and statistical approaches. *Studies in Fuzziness and Soft Computing*, vol. 185, pp. 255-280.

POPRAT, M.; BEISSWANGER, E.; HAHN, U. (2008) Building a BioWordnet using Wordnet data structures and Wordnet's software infrastructure – a failure story. *Proceedings of the ACL Workshop on*

Software Engineering, Testing, and Quality Assurance for Natural Language Processing. EUA: Ohio, pp. 31-39.

RATNAPARKHI, A. (1996) A maximum entropy part-of-speech tagger. *Proceedings of the 1st Empirical Methods in Natural Language Processing Conference.* EUA-Philadelphia, pp. 133-142.

RENOUF, A. (ed.) (1998) *Explorations in Corpus Linguistics*. Amsterdam: Rodopi.

ROVENTINI, A.; MARINELLI, R. (2004) Extending the Italian Wordnet with the specialized language of the maritime domain. *Proceedings of the 2nd International Global Wordnet Conference.* Masaryk University, Brno, pp. 193-198.

SAGRI, T. M.; TISCORNIA, D.; BERTAGNA, F. (2004) Jur-Wordnet. *Proceedings of the 2nd International Global Wordnet Conference.* Masaryk University, Brno, pp. 305-310.

BERBER SARDINHA, T. (2000) Linguística de Corpus: histórico e problemática. *Delta*, vol. 16 (2), pp. 323-367.

____ (2004) *Lingüística de Corpus*. Barueri-SP: Manole.

SCHMID, H. (1994) Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK, pp. 44-49.

SCOTT, M. (2008) *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.

SINCLAIR, J. (2005) Corpus and text: basic principles. In: WYNNE, M. (ed.). *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. pp.1-16. Disponível em: <http://ahds.ac.uk/linguistic-corpora/>. Acesso em 30 de outubro de 2006.

SMITH, B.; FELLBAUM, C. (2004) Medical Wordnet: a new methodology for the construction and validation of information resources for consumer health. *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, pp. 371-382.

SUÁREZ, M.; CABRÉ, M. T. (2002) La variación denominativa en los textos de especialidad: indicios linguísticos para su recuperación automática. *Proceedings of the 8th Simposio Iberoamericano de Terminología*. Cartagena de Indias, pp. 1-12.

USCHOLD, M., GRUNINGER, M. (1996) Ontologies: principles, methods and applications. *Knowledge Engineering Review*, vol. 11(2), pp. 93-155.

VOSSEN, P. (ed.) (2002) *EuroWordnet general document (Version 3 – Final)*. Disponível em: <http://www.vossen.info/docs/2002/EWNGeneral.pdf>. Acesso em 16 de julho de 2010.

ZAVAGLIA, C. *et al.* (2007) Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. *Proceedings of the 5th Workshop in Information and Human Language Technology.* RJ, pp. 1575-84.