SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
3ºONTOBRAS
30 e 31 de agosto de 2010 - Florianópolis/SC

A construção de *wordnets* terminológicas com base em *corpus*

Ariani Di Felippo (UFSCar) ariani@ufscar.br

Resumo: No âmbito do Processamento Automático das Línguas Naturais (PLN), é premente a necessidade de construção de recursos léxico-conceituais especializados ou terminológicos. A WordNet de Princeton (WN.Pr), entendida como um tipo especial de ontologia linguística, tem motivado a construção de inúmeras bases terminológicas no formato "wordnet". Tais bases são comumente construídas com base em recursos estruturados (p.ex.: thesaurus, glossários, etc.). Para certos domínios do conhecimento, tais recursos não estão disponíveis ou mesmo não existem. Dessa forma, no âmbito do projeto TermiNet, propôs-se uma metodologia baseada em corpus para a construção de wordnets terminológicas (ou termintes). Neste artigo, em especial, apresenta-se tal metodologia, a qual se caracteriza por equacionar a tarefa de construção de wordnets terminológicas em três fases: linguística, computacional e implementacional.
Palavras-chave: base de dados lexicais; wordnet; corpus; terminologia.

1. Introdução

Na área do Processamento Automático das Línguas Naturais (PLN), busca-se desenvolver sistemas computacionais "capazes" de processar (interpretar/gerar) as línguas naturais. Dentre eles, citam-se os sistemas de: tradução automática, sumarização automática, etc. Quando baseados em conhecimento linguístico, tais sistemas comumente necessitam de uma "base de dados lexicais" (ou léxico), cuja tarefa é a 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 (HANKS, 2004).

No caso do processamento 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. 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 que representam um conceito. De acordo com a categoria sintática das unidades que compõem os synsets, os mesmos podem estar inter-relacionados pela relação léxico-semântica da antonímia ou pelas relações semântico-conceituais da hiperonímia/hiponímia, holonímia/meronímia, acarretamento e causa. Entendida como uma "ontologia linguística"[1], a WN.Pr tem motivado a construção de bases lexicais no formato

[1] As ontologias linguísticas caracterizam-se por armazenar apenas conceitos lexicalizados (em uma determinada língua), isto é, conceitos expressos por unidades lexicais. Sob esse ponto de vista, uma ontologia é um inventário

"wordnet" para inúmeras línguas. 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. Nesse sentido, é possível encontrar vários trabalhos que objetivam: (i) integrar wordnets genéricas e especializadas (p.ex.: MAGNIN; SPERANZA, 2001; ROVENTINI; MARINELLI, 2004; BENTIVOGLI et al., 2004), (ii) enriquecer wordnets genéricas com o acréscimo de unidades terminológicas (p.ex.: BUITELAAR; SACALEANU, 2002) ou (iii) construir propriamente wordnets terminológicas (e.g.: SAGRI et al., 2004; SMITH; FELLBAUM, 2004).

Em tais projetos, o conhecimento léxico-conceitual especializado é comumente extraído de fontes estruturadas, sejam elas em formato digital ou impressas. Dentre as fontes estruturadas, estão os glossários, dicionários, vocabulários, etc. Entretanto, alguns domínios de especialidade, em especial os emergentes, ainda não foram sistematizados em glossários, dicionários, vocabulários, etc. Nos casos em que não há fontes estruturadas ou mesmo nos casos em que tais fontes não estejam disponíveis, a utilização dos *corpora* textuais caracteriza-se como uma alternativa viável para a construção de wordnets terminológicas.

Diante de tal cenário, optou-se, no âmbito do projeto TermiNet, pela utilização dos *corpora* como fontes de conhecimento especializado. O TermiNet, em especial, é um projeto de dois anos (2009-2011) que, com o auxílio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - 2009/06262-1) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - 471871/2009-5), objetiva: (i) instanciar a metodologia genérica de pesquisa no PLN elaborada por Dias-da-Silva (2006) para o desenvolvimento de wordnets terminológicas ou terminets (do inglês, **termi**nological word**nets**); (ii) 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.

Neste trabalho, em especial, apresenta-se a instanciação da metodologia genérica de pesquisa no PLN elaborada por Dias-da-Silva (2006) para o desenvolvimento de wordnets terminológicas ou terminets.[2] Para tanto, apresenta-se, na Seção 2, a metodologia genérica de pesquisa no PLN de Dias-da-Silva (2006). Na Seção 3, apresenta-se a instanciação (ou especificação) da referida metodologia para o desenvolvimento de terminets. Na Seção 4, algumas considerações finais sobre este trabalho são apresentadas.

---

dos sentidos de uma dada língua, ou seja, é um inventário somente daqueles conceitos compartilhados por uma comunidade linguística (VOSSEN, 1998).

[2] Vistas aqui como "ontologias linguísticas especializadas", as terminets armazenam somente daqueles conceitos (lexicalizados) compartilhados pela comunidade linguística que caracteriza o domínio de especialidade que se quer sistematizar.

2. 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 seguintes etapas: "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 (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 computacionalmente tratáveis. E, por fim, no domínio implementacional, as atividades ficam concentradas nas questões relativas à implementação do sistema de PLN.

3. 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:

a) **domínio linguístico:** delimitação do domínio de conhecimento especializado; delimitação das fontes e da estratégia de aquisição do conhecimento necessário à criação de uma wordnet, e delimitação e compilação do conhecimento léxico-conceitual;

b) **domínio representacional:** representação do conhecimento delimitado no domínio linguístico no formato wordnet;

c) **domínio implementacional:** transformação da representação do conhecimento linguístico em uma base lexical propriamente dita, ou seja, em um objeto computacional.

A Figura 1 ilustra os domínios e as tarefas previstas na metodologia.

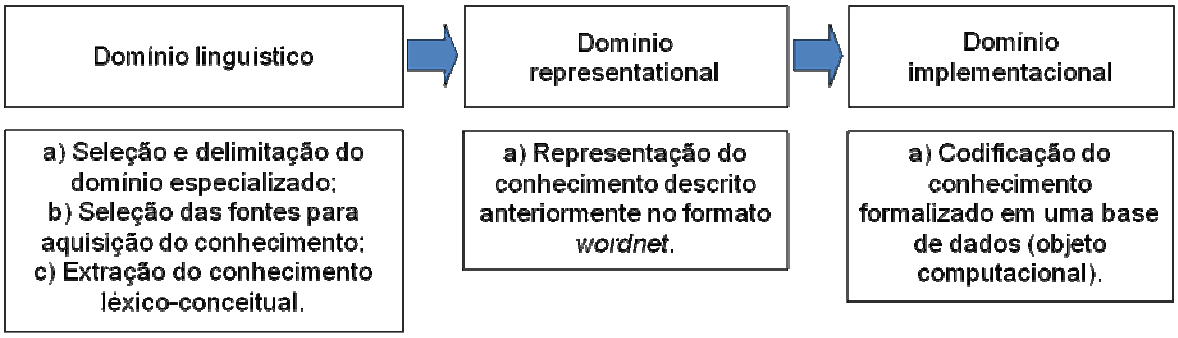

Figura 1: Os domínios e as tarefas de construção de *wordnets* terminológicas.

4.1. As tarefas do domínio linguístico

4.1.1. A seleção e 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. 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*; (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; (b) 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 as 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) 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 no projeto TermiNet, os *corpora* constituem a principal fonte da qual o conhecimento léxico-conceitual deve ser extraído.

Com base nos pressupostos da Linguística de Corpus, a construção do *corpus* deve seguir 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. No projeto TermiNet, propõe-se que a projeção do *corpus* seja pautada em três grupos de critérios (DI FELIPPO; SOUZA, 2009): (i) as características definitórias de *corpus*, (ii) o tipo de recurso lexical para o qual o *corpus* servirá de base e (iii) as decisões de projeto. Com base nesses três grupos de critérios, tem-se a tipologia apresentado no Quadro 1.

Quadro 1: Tipologia do *corpus* para a construção de uma terminet em PB.

Tipologia	
Tamanho	Médio-grande (no mínimo, 1 milhão de palavras)
Balanceado	Por gênero
Modalidade	Escrito
Tipo de texto	Escrito (língua escrita registrada em meio escrito)
Mídia	Jornais, livros, manuais, periódicos e outras
Cobertura da língua	Especializado
Gênero	Técnico-científico, científico de divulgação, instrucional, informativo e técnico-administrativo
Quantidade de línguas	Monolíngue
Anotação	Anotado (nível morfossintático)
Comunidade produtora	Falantes nativos
Mutabilidade	Aberto
Variação histórica	Sincrônico (contemporâneo)
Disponibilidade	Disponível via *Web*

Para a compilação dos textos, indica-se a coleta de material disponível na *web* devido ao custoso trabalho de digitalização de material impresso. Além disso, essa indicação justifica-se pelo fato de a *web* ser uma mina de dados linguísticos de riqueza e acessibilidade sem precedentes (KILGARRIF; GEFENSTETTE, 2003). Para tal coleta, utiliza-se comumente a abordagem manual, que consiste na seleção manual de páginas e documentos na *web*.

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 (http://www.nilc.icmc.usp.br/lacioweb/).

Finalmente, o *corpus* deve passar por um processo de anotação morfossintática ou sintática 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*). 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). A anotação sintática consiste no reconhecimento da estrutura sintática das sentenças do *corpus*. Para a anotação sintática de textos em PB, tem-se disponível o *parser* PALAVRAS.

4.1.3. A delimitação e compilação do conhecimento léxico-conceitual

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 unidades 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

Na literatura, existem três paradigmas de extração (CABRÉ et al., 2001; JACQUEMIN, BOURIGAULT, 2003; PAZIENZA et al., 2005; BERNHARD, 2006): (i) paradigma linguístico; (ii) paradigma estatístico; (iii) paradigma híbrido.

No paradigma linguístico, busca-se identificar os candidatos a termos com base em certos conhecimentos linguísticos, os quais são dependentes de língua e, por vezes, de domínio também. Para que os extratores desenvolvidos segundo o paradigma linguístico funcionem, o *corpus* sob análise deve ser pré-processado. No caso, esse pré-processamento comumente consiste na etiquetação morfossintaticamente (do inglês, *tagging*) das ocorrências do *corpus* e na identificação dos sintagmas do mesmo. Esses dois pré-processos são comumente realizados pelas seguintes

ferramentas computacionais, respectivamente: etiquetador morfossintático ou *tagger* e analisador (ou etiquetador) sintático ou *parser*.

Para a extração de termos simples (ou seja, aqueles formados por apenas um elemento) no paradigma linguístico, os extratores comumente utilizam três tipos de conhecimento linguístico: morfemas greco-latinos, categoria sintática ou nuclearidade sintagmática.

Para a extração de termos complexos, ou seja, formados por mais de um elemento, os extratores utilizam frequentemente três tipos de informação linguística: padrão morfossintático, expressão indicativa ou nuclearidade sintagmática. Os padrões morfossintáticos são sequências de etiquetas morfossintáticas que podem ser descritas por meio de expressões regulares do tipo: (Nome + (Adjetivo | Nome)) (p.ex.: *missão aérea*). As expressões indicativas (ou padrões léxico-sintáticos) podem ser vistas como indicadores estruturais que introduzem definições e os termos definidos, por exemplo: "é definido como", "é chamado de", "é um tipo de", etc.

No paradigma estatístico, os candidatos a termo são extraídos com base na aplicação de medidas estatísticas como *frequência, informação mútua, log-likelihood ratio* e *coeficiente Dice*. Para a extração de unidades simples (unigramas), utiliza-se comumente a frequência simples, que pode ser entendida como a quantidade de vezes que um *token* (isto é, sequência de caracteres separados por espaços em branco) ocorre em um único texto do *corpus* ou no *corpus* inteiro. Vale ressaltar que a frequência simples também pode ser usada para a extração de termos complexos (ou seja, bigramas, trigramas, etc.). As demais estatísticas (*informação mútua, log-likelihood ratio* e *coeficiente Dice*), no entanto, são utilizadas para a extração dos candidatos complexos, pois buscam identificar a estabilidade de expressões sintagmáticas[3]. Tais medidas podem ser aplicadas por meio da utilização do pacote NSP (do inglês, *N-gram Statistics Package*) (http://www.d.umn.edu/~tpederse/nsp.html), que é constituído por um conjunto de programas que auxilia na análise de *n*-gramas (isto é, sequência de *n* elementos em texto) em arquivos no formato *.txt*.

No paradigma híbrido, a extração automática de candidatos é comumente feita com base em métodos desenvolvidos segundo ambos os paradigmas: linguístico e estatístico.

Para a extração de candidatos a termo a partir de textos em PB, ressalta-se o sistema híbrido denominado OntoLP (RIBEIRO Jr, 2008). A ferramenta OntoLP é, na verdade, um *plug-in*[4] para o editor de ontologias Protégé (http://protege.stanford.edu/), bastante utilizado na comunidade científica. O OntoLP, em especial, auxilia o usuário do Protégé nas tarefas iniciais de construção de uma ontologia, que são: (i) extração de termos candidatos a conceitos de um domínio e (ii)

[3] Segundo Pazienza et al. (2005), tal estabilidade de "expressões sintagmáticas" (ou seja, unidades linguísticas compostas por elementos fortemente associados) é denominada *unihood*.

identificação da relação hierárquica (hiperonímia/ hiponímia) entre os termos. Além do OntoLP, destaca-se atualmente o extrator linguístico E_xATO_{LP} (Extrator Automático de Termos para Ontologias em Língua Portuguesa) (LOPES et al., 2009). Em particular, essa ferramenta identifica os candidatos a termos com base em uma única informação linguística: nuclearidade em SNs.

d) A delimitação e identificação das relações internas às terminets

Essa etapa consiste na identificação no *corpus* da relação de sinonímia e 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 hiperonímia/hiponímia.

Para a identificação e extração da relação de hiperonímia/hiponímia, em particular, vários trabalhos (p.ex.: MORIN; JACQUEMIN (2004) e MITITELU (2006)) têm aplicado a abordagem linguística, que se baseia no reconhecimento dos padrões léxico-sintáticos identificados por Hearst (1992). Especificamente, 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 extração da relação de hiperonímia/hiponímia de textos em PB, tem-se o OntoLP. No caso, essa ferramenta extrai tais relações com base nos padrões de Hearst (1992) e Morin e Jacquemin (2004) que foram adaptados ao PB.

4.2. As tarefas do domínio representacional

O formato wordnet fundamenta-se em três construtos formais (FELLBAUM, 1998): (i) o método diferencial, segundo o qual os conceitos são ativados na mente por meio de formas lexicais sinônimas, eliminando a necessidade de determinar o valor semântico das unidades; (ii) os synsets, que são conjuntos de formas lexicais determinados pela relação de pertença e munidos de dois tipos de ponteiros, os que especificam relações entre formas (antonímia) e os que especificam relações entre conceitos (synsets); (iii) a noção de matriz lexical, que especifica uma relação biunívoca entre conceitos e synsets. A montagem das bases wordnets 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 remeta a pesquisa às atividades do domínio implementacional.

[4] Pequeno programa de computador que serve normalmente para adicionar funções a outros programas maiores, provendo alguma funcionalidade especial ou muito específica.

4.3. As tarefas do domínio implementacional

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 (http://nlp.fi.muni.cz/projekty/visdic/). 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. Caso necessário, uma ferramenta desse tipo poderá ser desenvolvida no âmbito do projeto.

4.3.2. A inserção das informações no editor

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

5. Considerações finais

Acredita-se que o projeto TermiNet fornece uma metodologia suficientemente clara e genérica para a construção de bases terminológicas no formato wordnet em PB. Essa metodologia, em especial, caracteriza-se por ser baseada em *corpus* e não em fontes estruturadas e por conciliar as faces linguística e a computacional das atividades no PLN. Tal metodologia está sendo validada por meio da construção de um *corpus* em PB do domínio da Educação a Distância, o qual será a fonte para a construção da terminet do referido domínio, a WordNet.EaD. Com isso, pretende-se incentivar o desenvolvimento de terminets em PB, pois tais objetos computacionais podem beneficiar não só 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.

Agradecimento

À FAPESP e CNPq, pelo apoio financeiro.

Referências

AGBAGO, A., BARRIÈRE, C. Corpus construction for Terminology. In: **CORPUS LINGUISTICS CONFERENCE**, 2005. Proceedings... Birmingham, 2005. p. 14-17.

ALMEIDA, G.M.B. A Teoria Comunicativa da Terminologia e a sua prática. **Alfa**, v. 50, p. 81-97, 2006.

____.; CORREIA, M. Terminologia e corpus: relações, métodos e recursos. In: Stella E. O. Tagnin and Oto Araújo Vale (orgs.). **Avanços da Lingüística de Corpus no Brasil**. 1 ed. Humanitas/FFLCH/USP; São Paulo, volume 1, 63-93.

BENTIVOGLI, L.; BOCCO, A.; PIANTA, E. ArchiWordnet: integrating Wordnet with domain-specific knowledge. In: **INTERNATIONAL GLOBAL WORDNET CONFERENCE**, 2, 2004. Proceedings... Masaryk University, Brno, 2004. p. 39-47. Disponível em: <http://www.fi.muni.cz/gwc2004/proc/101.pdf>. Acesso em: 10 maio 2008.

BERNHARD, D. Multilingual term extraction from domain-specific corpora using morphological Structure. In: **CONFERENCE OF THE EUROPEAN CHAPTER OF THE ACL**, 11, 2006. Proceedings… Trento, Italy, 2006, p. 171-174.

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

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

BUITELAAR, P.; SACALEANU, B. Extending synsets with medical terms. In: **INTERNATIONAL CONFERENCE ON GLOBAL WORDNET**, 1, 2002. Proceedings… Mysore, India, 2002.

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

___; ESTOPÀ, R.; PALATRESI, J. V. 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., 2001, p. 53-87.

____. Application-driven terminology engineering. **Terminology**, v.11(2), p. 1-19, 2005.

DIAS-DA-SILVA, B.C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

DI FELIPPO, A.; SOUZA, J. W. C. Projetando o corpus para a construção de uma wordnet terminológica. In **Livro de Resumos do VIII Encontro de Linguística de Corpus**. pp. 70-72, Rio de Janeiro, Rio de Janeiro: UERJ, Brasil, 13 e 14 de Novembro.

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

HANKS, P. Lexicography. In: Mitkov, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, p. 48-69.

HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: **INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS**, 14, 1992. Proceedings... Nantes, 1992. p. 539-545.

JACQUEMIN, C.; BOURIGAULT, D. Term extraction and automatic indexing. In: Mitkov, R. (Ed.). **Handbook of Computational Linguistics**. Oxford University Press, 2003, p.599-615.

KILGARIFF, A.; GREFENSTETTE, G. Introduction to the special issue on the Web as Corpus. **Computational Linguistics**, v. 29, 2003.

LOPES, L. et al. ExATOlp - an automatic tool for term extraction from Portuguese language corpora. In: **LTC,** 2009. Proceedings…, Poznam, Poland, 2009.

MAGNINI, B.; SPERANZA, M. Integrating generic and specialized wordnets. In: **CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING**, 2, 2001. Proceedings…Tzigov Chark, Bulgaria. 2001.

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

MITITELU, V.B. Automatic extraction of patterns displaying hyponym-hypernym co-cccurrence from corpora. In: **CENTRAL EUROPEAN STUDENT CONFERENCE IN LINGUISTICS**, 1, 2006. Proceedings... Budapest, Hungary, 2006.

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

RATNAPARKHI, A. A maximum entropy part-of-speech tagger. In: **EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING CONFERENCE**, 1, 1996. Proceedings… Philadelphia, 1996. p.133-142.

RIBEIRO JR., L.C. **OntoLP: construção semi-automática de ontologias a partir de textos da língua portuguesa.** São Leopoldo, 2008, 131p. Dissertação (Mestrado em Computação Aplicada) – Univ. do Vale do Rio dos Sinos, 2008.

ROVENTINI, A.; MARINELLI, R. Extending the Italian Wordnet with the specialized language of the maritime domain. In: **INTERNATIONAL GLOBAL WORDNET CONFERENCE**, 2, 2004. Proceedings... Masaryk University, Brno, 2004. p. 193-198.

SAGRI et al. Jur-Wordnet. In: **INTERNATIONAL GLOBAL WORDNET CONFERENCE**, 2, 2004. Proceedings... Masaryk University, Brno, 2004. p. 305-310.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, 1994. Proceedings… Manchester, UK, 1994. p. 44-49.

SMITH, B.; FELLBAUM, C. Medical Wordnet: a new methodology for the construction and validation of information resources for consumer health. In: **INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS**,20,2004. Proceedings ..Geneva, 2004.

VOSSEN, P. EuroWordNet: Linguistic ontologies in a multilingual database. **Communication and Cognition for Artificial Intelligence** (Special Issue), v. 15, n. (1-2), p. 37-80, 1998.