Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Extração de conhecimento terminológico no projeto TermiNet

Ana Catarina Gianoti
Ariani Di Felippo

NILC-TR-11-01

Agosto, 2011

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório, estão descritas as atividades realizadas no projeto iniciação científica intitulado "*Extração automática de termos segundo o paradigma linguístico*", o qual integrou o projeto "*Instanciação e aplicação de uma metodologia para o desenvolvimento de wordnets terminológicas em português do Brasil*" (doravante, TermiNet) (FAPESP Proc. 2009/06262-1 e CNPq Proc. 471871/2009-5). Em especial, descreve-se a extração automática de terminologia (EAT) baseada no paradigma linguístico no âmbito do projeto TermiNet, o qual objetiva construir uma base de dados lexicais terminológicos do domínio da Educação a Distância (EaD) em português do Brasil (PB) no formato da WordNet de Princeton (FELLBAUM, 1998). A EAT consiste em reconhecer/ extrair possíveis unidades terminológicas, a partir de um *corpus* ("extensos conjuntos de textos em formato digital"), de forma automática e é uma das tarefas previstas na metodologia proposta no TermiNet (DI-FELIPPO, ALMEIDA, 2010) para a construção de bases lexicais terminológicas no formato *wordnet*. Da EAT, são obtidos os termos que, uma vez validados, são organizados em função da sinonímia, pois uma *wordnet* armazena essencialmente (i) conceitos lexicalizados, codificados em "conjuntos de sinônimos" (em inglês, *synonym sets* ou *synsets*), e (ii) diferentes relações semântico-conceituais entre *synsets*. Como resultado concreto, o projeto foi responsável pela identificação de 59 unigramas, 102 bigramas, 63 trigramas e 5 quadrigramas, os quais serão diretamente inseridos na base de dados terminológica do domínio da EaD denominada WordNet.EaD.

Este trabalho contou com o apoio financeiro
da FAPESP e do CNPq/UFSCar.

Sumário

0. Introdução

Quando baseados no paradigma simbólico, os pesquisadores do Processamento Automático das Línguas Naturais (PLN) buscam desenvolver sistemas que processam língua natural (p.ex.: tradutor automático) cuja arquitetura seja semelhante à arquitetura proposta para o sistema linguístico (Allen, 1995). Como decorrência, um sistema de PLN possui dois grupos de componentes imprescindíveis para o processamento automático de uma língua: as bases de dados e os módulos de processamento que atuam sobre essas bases (DIAS-DA-SILVA, 1996). À base lexical, em especial, cabe a tarefa de armazenar as unidades lexicais da(s) língua(s) e seus traços morfológicos, sintáticos, semânticos e pragmático-discursivos, os quais subsidiam a interpretação e/ou geração da(s) língua(s) pelo sistema (HANKS, 2004).

Desde o seu lançamento na década de 1990, a WordNet de Princeton (WN.Pr) (Fellbaum, 1998) tem sido amplamente utilizada para o processamento do inglês norte-americano devido a sua acessibilidade, reutilização, adequação linguística e computacional e abrangência (MORATO *et al.,* 2004). Por conseguinte, a WN.Pr tem motivado a construção de bases lexicais de língua geral no formato *wordnet* para inúmeras línguas. A *wordnet* para o português do Brasil (PB), por exemplo, a WordNet.Br (DIAS-DA-SILVA *et al.,* 2008), está em pleno desenvolvimento.

Uma base no formato *wordnet* caracteriza-se primordialmente por armazenar as unidades lexicais em arquivos distintos, os quais correspondem às quatro categorias sintáticas: nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em conjuntos de sinônimos (do inglês, *synonym sets* ou *synsets*) (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 e pelas relações semântico-conceituais da hiperonímia/hiponímia, holonímia/meronímia, acarretamento e causa.

Objetivando-se processar textos especializados, vários projetos recentes têm focalizado: (i) integrar *wordnets* de língua geral e terminológicas[1] (p.ex.: MAGNIN, SPERANZA 2001; ROVENTINI, MARINELLI 2004; BENTIVOGLI *et al.*, 2004); (ii) enriquecer *wordnets* de lingual geral com o acréscimo de unidades terminológicas ou termos[2] (p.ex.: BUITELAAR, SACALEANU 2002); ou (iii) construir *wordnets* terminológicas[3], como a JurWordnet (SAGRI *et al.,* 2004), a Medical Wordnet (SMITH, FELLBAUM 2004) e a BioWordnet (POPRAT *et al.,* 2008). Esses trabalhos são comumente pautados em metodologias que se caracterizam pela extração manual do conhecimento a partir de fontes estruturadas (p.ex.: dicionários, taxonomias, etc.). Diante principalmente da escassez de recursos estruturados disponíveis para vários domínios em PB, propôs-se, no projeto TermiNet[4] (FAPESP Proc. 2009/06262-1 e CNPq Proc. 471871/2009-5), uma metodologia suficientemente clara e genérica para a construção de *wordnets* terminológicas ou *terminets* (do inglês, *terminological wordnets*) referentes a quaisquer domínios em PB. Essa metodologia caracteriza-se por ser semiautomática e

[1] Relativas a domínio ou áreas específicas do conhecimento.

[2] Unidades lexicais da língua geral que se caracterizam por expressarem conhecimento especializado, produzido no âmbito das ciências e das técnicas (CABRÉ, 1999).

[3] Bases de dados lexicais no formato *wordnet* que armazenam o repertório ou conjunto de termos de uma área específica.

[4] www.nilc.icmc.usp.br/~arianidf/terminet

baseada em *corpus*[5] (isto é, fonte não-estruturada) (DI-FELIPPO, ALMEIDA, 2010). Tendo em vista a escassez de recursos computacionais terminológicos para o PB, a validação de tal metodologia está sendo feita pela construção de uma base do domínio da Educação a Distância (EaD) em PB, denominada WordNet.EaD.

Neste relatório, descreve-se o método linguístico e a ferramenta computacional (extrator automático) utilizados para a construção da WordNet.EaD em uma tarefa específica dentre as várias previstas na metodologia do TermiNet, a saber: a "extração automática dos candidatos a termos"[6] (EAT) a partir de *corpus*. Além disso, apresentam-se as estratégias de validação dos candidatos e os resultados obtidos pelo método/ferramenta em questão.

Para tanto, equaciona-se este texto em 6 Seções. Na Seção 1, apresenta-se o formato *wordnet* para bases de dados lexicais. Na Seção 2, apresenta-se a metodologia trifásica proposta no TermiNet com o objetivo de contextualizar a EAT no projeto. Na Seção 3, discorre-se sobre a delimitação de "unidade lexical terminológica" que guiou a tarefa de EAT. Na Seção 4, apresenta-se o *corpus* do domínio da EaD (o *Corpus.*EaD) a partir do qual os candidatos a termos foram extraídos. Na Seção 5, descrevem-se o método linguístico e a ferramenta de extração automática, além das estratégias e dos resultados da validação dos candidatos. E, por fim, na Seção 6, algumas considerações finais sobre o trabalho ora descrito são apresentadas.

1. 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, 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[7] 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* {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

[5] Por corpus, entende-se: "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*" (SINCLAIR, 2005).

[6] Apesar de ser qualificada como "automática", a EAT caracteriza-se pela extração do conhecimento por meio de ferramentas computacionais seguida pelas tarefas de limpeza e validação humanas ou manuais.

[7] 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.

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 1, baseado em Fellbaum (1998), resume o conjunto principal de relações em função das categorias sintáticas.

Quadro 1: As relações semânticas da WN.Pr em função das categorias sintáticas.

Relações	**Categorias sintáticas**	**Exemplos**
Antonímia (oposição conceitual)	Adj, Adv N, V	*mulher* é antônimo de *homem*[8] *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		

A seguir, na Seção 2, apresenta-se a instanciação da metodologia genérica de pesquisa no PLN proposta por Dias-da-Silva (2006) para o desenvolvimento de *wordnets* terminológicas.

2. O projeto TermiNet e sua metodologia trifásica

A especificação de uma metodologia clara e genérica para a construção de *terminets* teve como ponto de partida a metodologia genérica de pesquisa no PLN proposta por Dias-da-Silva (2006). Essa metodologia destaca-se por equacionar todo empreendimento no PLN em três domínios e, sobretudo, por evidenciar a importância do conhecimento linguístico nesse tipo de pesquisa. 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", que é equacionada em função das etapas de 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

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

computador que codifica essa representação) (DIAS-DA-SILVA, 1998; DI-FELIPPO, DIAS-DA-SILVA, 2009).

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

Com base no formato *wordnet* e na, a instanciação da metodologia ficou assim delimitada (DI-FELIPPO, ALMEIDA, 2010):

- *Domínio linguístico*: (i) seleção e delimitação do domínio especializado; (ii) delimitação do conhecimento léxico-conceitual, ou seja, das categorias sintáticas, do tipo de unidade lexical, das relações lexicais (sinonímia e antonímia) e das relações semântico-conceituais (hiponímia, meronímia, acarretamento e causa); (iii) seleção ou construção dos recursos-fonte e seleção da estratégia de extração do conhecimento, (iv) seleção de métodos e ferramentas de extração e a efetiva extração do conhecimento léxico-conceitual.
- *Domínio representacional*: representação do conhecimento delimitado no domínio linguístico no modelo *wordnet*, ou seja, com base nas noções de *forma lexical* (do inglês, *word form*), *synset*, *matriz lexical* e *ponteiros relacionais* (do inglês, *relational pointers*);
- *Domínio implementacional*: transformação do conhecimento representado no formato *wordnet* para uma base lexical relacional; especificamente, essa etapa engloba as tarefas de (i) seleção de um editor/sistema de gerenciamento de bases relacionais (SGBDR) (do inglês, *relational database management system*) e (ii) inserção dos dados no editor e construção da base.

Quanto às atividades do domínio linguístico, salienta-se que, para a validação da metodologia, selecionou-se o domínio da EaD com base nos critérios propostos por Almeida e Correia (2008): (i) interesse dos especialistas do domínio por um produto terminológico (no caso, uma *wordnet*); (ii) relevância educacional, social, político, econômico, científico e/ou tecnológico do domínio para o país, e (iii) facilidade de obtenção de fontes de conhecimento. O domínio da EaD caracteriza-se basicamente por: (i) a separação física (espaço-temporal) entre aluno e professor e (ii) a intensificação do uso de tecnologias de informação e comunicação como mediadoras da relação ensino-aprendizagem. Em suma, a EaD foi delimitada como uma modalidade educacional que faz uso das tecnologias denominadas "telemáticas" (ou seja, baseadas nas telecomunicações e na informática) (MILL, 2010).

As demais tarefas do domínio linguístico são descritas nas próximas Seções, pois estas respondem respectivamente às questões: (i) *O que extrair para a construção de uma terminet?*; (ii) *De onde extrair?* e, sobretudo, (iii) *Como extrair e validar?*.

3. A delimitação do conhecimento lexical terminológico

Para a construção de uma base *wordnet,* seja ela de língua geral ou especializada, é preciso especificar o conhecimento lexical e o semântico-conceitual a serem sistematizados. Quanto ao conhecimento lexical, é preciso especificar os elementos constitutivos dos *synsets,* ou seja, os tipos de unidades lexicais que formarão os conjuntos de sinônimos. Quanto ao conhecimento semântico-conceitual, é necessário especificar a noção de sinonímia para a construção dos *synsets* e as relações semântico-conceituais que interligarão os *synsets.*

Especificamente quanto ao conhecimento lexical, as unidades constitutivas dos *synsets* em uma *terminet* são as "unidades terminológicas" ou "termos" do domínio especializado. Um "termo" pode ser definido como signo linguístico, composto de forma e conteúdo, que representa certo conceito no interior de um domínio específico (Cabré, 1999). Os tipos de unidades a serem armazenadas em uma *terminet* foram delimitados em função de dois critérios: formal e funcional. Do ponto de vista formal ou de sua estrutura morfossintática, uma *terminet* poderá armazenar termos simples[9], constituídos de apenas um radical, com ou sem afixos (p.ex.: "ambiente"), e complexos, constituídos de dois ou mais radicais, aos quais podem-se acrescentar outros elementos[10] (p.ex.: "ambiente digital", "ambiente de aprendizagem", "ambiente virtual de aprendizagem").

Os termos compostos também são unidades lexicais formadas por dois ou mais radicais. No entanto, eles diferem dos complexos pelo alto grau de lexicalização e pelo conjunto de morfemas lexicais e/ou gramaticais que os constitui. No caso, os termos compostos por aglutinação (p.ex.: "teleducação" [tel(e)+educação]) e por justaposição (com ou sem hífen) (p.ex.: "livro-texto" e "weblog" [web+log]) são considerados termos simples[11].

Os termos complexos podem ter tamanhos distintos; os sintagmas terminológicos chegam a compor-se de até 5 unidades lexemáticas (p.ex.: *amilóide cutânea genuína localizada maculosa*) (BARROS, 2004, p. 101). Tendo em vista que os termos complexos de tamanho 5 ou pentagramas são menos frequentes, optou-se por restringir a constituição dos *synsets* de uma *terminet* às unidades dos tipos unigrama, brigrama, trigrama e quadrigrama, como as exemplificadas em (1).

(1) a. ambiente (unigrama ou 1-grama)
b. ambiente virtual (bigrama ou 2-grama)
c. ambiente de aprendizagem (trigrama ou 3-grama)
d. ambiente virtual de aprendizagem (quadrigrama ou 4-grama)

Quanto ao critério funcional, optou-se por restringir a construção de uma *wordnet* terminológica às unidades lexicais da categoria dos nomes (substantivos), pois estas ocupam um lugar de destaque nas terminologias, ou seja, no conjunto de termos de uma área especializada (BARROS, 2004). Devido à forte interdependência entre a formação dos conceitos e a formação de suas expressões linguísticas, as classes de palavras são associadas às classes de conceitos. Dessa associação, fala-se em "conceitos nominais",

[9] Do ponto de vista computacional, os termos simples são definidos como n-gramas de tamanho 1 (unigramas), ou seja, sequências únicas de caracteres separados por espaços em branco.
[10] Do ponto de vista computacional, os termos complexos são n-gramas de tamanho maior que 1, isto é, sequências de 2 unigramas (bigramas), 3 unigramas (trigramas), etc.
[11] Do ponto de vista computacional, os termos compostos são considerados unigramas.

"conceitos verbais", etc. (SAGER, 1993). Uma *terminet* codificará, assim, conceitos nominais (isto é, expressos por termos da classe dos nomes). Dessa forma, fez-se um recorte no conhecimento lexical comumente armazenado em uma base do tipo *wordnet.* Como consequência desse recorte, a WordNet.EaD deve armazenar apenas unigramas, brigramas, trigramas e quadrigramas da categoria dos nomes.

Após a delimitação do conhecimento lexical a ser armazenado em uma *terminet,* o próximo passo, segundo a metodologia proposta no TermiNet, consiste na seleção e/ou construção dos recursos-fonte a partir dos quais esse conhecimento deve ser extraído e na seleção da estratégia de extração.

4. A delimitação dos recursos-fonte e das estratégias de extração

O conhecimento lexical delimitado para a construção de uma *terminet* pode ser extraído de fontes estruturadas e/ou não-estruturadas (em formato eletrônico ou não). No projeto TermiNet, as fontes não-estruturadas, ou seja, os *corpora,* são considerados a principal fonte de conhecimento especializado. Quanto a estratégia de extração do conhecimento lexical a partir de *corpus,* optou-se no TermiNet pela semiautomática, ou seja, extração por meio de uma ferramenta computacional com posteriores pós-processamento e avaliação manuais. Essa escolha baseou-se no fato de a estratégia semiautomática promover mais rapidez ao desenvolvimento do recurso léxico-computacional e garantir maior facilidade de recuperação de informações relativas às unidades extraídas (RIGAU, 1999). Para a construção de uma *terminet,* deve-se investigar a disponibilidade de *corpora* do domínio especializado sob sistematização com o objetivo de se selecionar (ou adaptar) o mais adequado à tarefa. Caso contrário, é preciso construir um *corpus* do domínio. Para os casos em que a construção se faz necessária, o TermiNet especificou uma tipologia que reúne as características essenciais de um *corpus* para a construção de uma *terminet* (DI-FELIPPO, SOUZA, 2010) (Quadro 2).

Quadro 2: Tipologia do *corpus* para a construção de *terminets* em PB.

Critérios	**Características**
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 e Informativo
Quantidade de línguas	Monolíngue
Anotação	Anotado (nível morfossintático)
Comunidade produtora	Falantes nativos
Mutabilidade	Aberto
Variação histórica	Sincrônico (contemporâneo)
Disponibilidade	Disponível via *Web*[12]

Para a validação da metodologia de construção de *terminets,* construiu-se, a partir da tipologia do Quadro 1, um *corpus* do domínio da EaD, denominado *Corpus.*EaD. Esse

[12] O *Corpus.*EaD estará disponível em breve no seguinte endereço eletrônico: http://www.nilc.icmc.usp.br/~arianidf/terminet/corpus.html

corpus engloba 347 textos em PB do domínio da EaD que foram manualmente compilados de fontes de qualidade, ou seja, de conteúdo confiável. Tais fontes foram restritas a sites de instituições governamentais reconhecidamente vinculadas a projetos de EaD. A distribuição quantitativa dos textos em função dos tipos e gêneros textuais é apresentada no Quadro 3.

Quanto ao tamanho, o *Corpus*.EaD apresenta o total de 1.350.683 ocorrências.

Quadro 3: *Corpus*.EaD: número de textos por gênero.

Gêneros Textuais	**Tipos Textuais**	**Quantidade**	**Total**
Científico de divulgação	Artigos de divulgação	138	307
Instrucional	Livro-texto	9	
	Apostila	2	
Informativo	Notícias/reportagens	158	
Técnico-científico	Tese	3	40
	Dissertação	14	
	Projetos de pesquisa	1	
	Artigos científicos	22	

5. A extração automática dos candidatos a termos: base teórica

Quanto à extração das potenciais unidades lexicais formadoras de *synsets*, ressalta-se que, na literatura, há três paradigmas de EAT (CABRÉ *et al.*, 2001; JACQUEMIN, BOURIGAULT, 2003; PAZIENZA *et al.*, 2005; BERNHARD, 2006): linguístico, estatístico e híbrido.

No paradigma linguístico, busca-se identificar os candidatos a termo com base em certos conhecimentos linguísticos. Para que os extratores desenvolvidos segundo o paradigma linguístico funcionem, o *corpus* sob análise deve ser pré-processado. Dependendo do método linguístico implementado, esse pré-processamento pode incluir os processos de: (i) tokenização (no inglês, *tokenization*), (ii) etiquetação morfossintática (no inglês, *tagging*), (iii) lematização (no inglês, *lemmatization*), (iv) análise sintática (no inglês, *parsing*) e/ou (v) exclusão de *stopwords*. Para a extração dos candidatos simples (unigramas), os extratores comumente utilizam dois tipos de conhecimento linguístico, os quais dão origem a dois métodos de EAT, a saber: categoria sintática ou nuclearidade sintagmática. Quando baseados no método que identifica os candidatos a termos pela categoria sintática (verbo, adjetivo, advérbio e verbo), os extratores possibilitam aos usuários a especificação da categoria que desejam extrair do *corpus*; uma vez especificada, somente as unidades da categoria em questão são extraídas. Quando baseados no reconhecimento de estruturas sintagmáticas, as ferramentas apenas extraem candidatos considerados núcleo de sintagmas (em especial, de sintagmas nominais (SNs)). Para a extração de candidatos complexos (n-gramas > 1), os extratores podem utilizar três tipos de informação linguística, ou seja, três métodos distintos, a saber: padrão morfossintático, expressão indicativa ou nuclearidade sintagmática. Os padrões morfossintáticos são sequências de etiquetas morfossintáticas, exemplificadas aqui pela expressão regular [Nome+Adjetivo] (p.ex.: "educação infantil"). As expressões indicativas (ou padrões léxico-sintáticos) introduzem definições e os termos definidos, por exemplo: "é definido como", "é um tipo de" (p.ex.: "A Educação a Distância (EaD) *é um tipo de* modalidade de ensino [...]."), etc. Dentre os trabalhos que se encaixam nessa linha, citam-se, por exemplo, os de Aussenac-Gilles e Séguéla (2000), Suárez e Cabré

(2002), Cederberg e Widdows (2003); Morin e Jacquemin (2004), Morin e Jacquemin (2004), Mititelu (2006), etc.

No paradigma estatístico, os candidatos são extraídos com base na aplicação de medidas como frequência, informação mútua, *log-likelihood ratio* e coeficiente Dice[13]. Para a extração de unidades simples, utiliza-se comumente a frequência simples, que pode ser entendida como a quantidade de vezes que um *token* (isto é, sequência de caracteres separada por espaços em branco) ocorre no *corpus*. A frequência simples também pode ser usada para a extração de termos complexos. As demais estatísticas (informação mútua, *log-likelihood ratio* e coeficiente Dice) são utilizadas apenas para a extração dos candidatos complexos, pois buscam identificar a estabilidade de expressões sintagmáticas, ou seja, a correlação entre n-gramas.

Para paradigma híbrido, os sistemas de EAT combinam métodos baseados em ambos os paradigmas: linguístico e estatístico. Os métodos híbridos são desenvolvidos com base no pressuposto de que a confiabilidade das medidas estatísticas é maior quando estas são aplicadas a candidatos a termo linguisticamente "justificados".

No caso específico da extração de candidatos a partir de textos em PB, há trabalhos desenvolvidos segundo os três paradigmas. Não há, no entanto, uma avaliação comparativa única e abrangente entre os métodos e extratores que indique quais são os de melhor desempenho para cada tipo de n-grama. As comparações existentes são parciais, diferindo, por exemplo, quanto a: o *corpus*, a *stoplist*[14] e o tipo de n-grama extraído. Consequentememte, optou-se por utilizar os principais métodos/ferramentas de cada um dos paradigmas (estatístico, linguístico e híbrido) disponíveis para o PB, os quais estão descritos no Quadro 4.

Quadro 4: Métodos/ferramentas de extração utilizados no TermiNet.

Paradigma	N-Grama	Conhecimento	Ferramenta
Linguístico	1, 2, 3, 4	Núcleo de sintagma nominal	ExATOlp (Lopes *et al.*, 2009)
Híbrido	1	Classe gramatical/tf-idf	OntoLP (Ribeiro Jr., 2008)
	2	Padrões morfossintáticos/tf-idf	
	3	Padrões morfossintáticos/tf-idf	
	4	Padrões morfossintáticos/tf-idf	
Estatístico	1, 2, 3, 4	Frequência simples	Pacote NSP (Banerjee, Pedersen, 2003)

Neste relatório, em especial, descrevem-se as atividades de extração dos candidatos apenas segundo o paradigma linguístico.

5.1. A extração linguística

Para a extração dos candidatos a termo com base no paradigma linguístico, utilizou-se o extrator ExATOlp (Lopes *et al.*, 2009, 2010). Nessa ferramenta, foi implementado um método linguístico de extração de candidatos que se baseia na informação de

[13] Mais informações sobre essas medidas estatísticas são encontradas em Ribeiro Jr. (2008).

[14] Lista composta por *stopwords,* ou seja, palavras irrelevantes que aparecem com muita frequência nos textos.

nuclearidade em SNs. A seguir, descrevem-se as especificações de extração e o pós-processamento das listas geradas pelo ExATOlp.

5.1.1. As etapas metodológicas de extração

A anotação linguística do *Corpus*.EaD foi feita por meio do *parser*[15] PALAVRAS (Bick, 2000). As dados de entrada do PALAVRAS são os textos do *corpus* no formato ASCII (txt). Após a análise sintática automática, o *parser* gera, como saída do processamento, um arquivo no formato TigerXML para cada texto do *corpus.* Na Figura 1, ilustra-se a análise sintática feita pelo PALAVRAS para a sentença "Acessibilidade é tema de novo curso a distância" de um dos textos do *Corpus*.EaD.

O arquivo XML gerado pelo PALAVRAS contém todas as sentenças de um texto analisadas sintaticamente, sendo que, para cada sentença, o *parser* constroi uma árvore sintática em que os nós terminais são as palavras das sentenças e os não-terminais representam as categorias sintagmáticas. Os arquivos no formato TigerXML são a entrada da ferramenta ExATOlp.

.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <xml>
  - <corpus>
    - <body>
      - <s id="s1" ref="1" source="Running text" forest="1" text="Acessibilidade é tema de novo curso a distância.">
        - <graph root="s1_500">
          - <terminals>
              <t id="s1_1" word="Acessibilidade" lemma="acessibilidade" pos="n" morph="F S" sem="percep-f" extra="--" />
              <t id="s1_2" word="é" lemma="ser" pos="v-fin" morph="PR 3S IND VFIN" sem="--" extra="vK fmc mv" />
              <t id="s1_3" word="tema" lemma="tema" pos="n" morph="M S" sem="ac" extra="--" />
              <t id="s1_4" word="de_novo" lemma="de_novo" pos="adv" morph="--" sem="--" extra="--" />
              <t id="s1_5" word="curso" lemma="curso" pos="n" morph="M S" sem="dir per occ" extra="--" />
              <t id="s1_6" word="a" lemma="o" pos="art" morph="F S" sem="--" extra="--" />
              <t id="s1_7" word="distância" lemma="distância" pos="n" morph="F S" sem="am" extra="--" />
              <t id="s1_8" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--" />
            </terminals>
          - <nonterminals>
            - <nt id="s1_500" cat="s">
                <edge label="STA" idref="s1_501" />
              </nt>
            - <nt id="s1_501" cat="fcl">
                <edge label="S" idref="s1_1" />
                <edge label="P" idref="s1_2" />
                <edge label="Cs" idref="s1_3" />
                <edge label="fA" idref="s1_4" />
                <edge label="fCs" idref="s1_5" />
                <edge label="fCs" idref="s1_502" />
                <edge label="DN" idref="s1_6" />
                <edge label="H" idref="s1_7" />
                <edge label="PU" idref="s1_8" />
              </nt>
              <nt id="s1_502" cat="np" />
            </nonterminals>
          </graph>
        </s>
```

Figura 1: Análise sintática do PALAVRAS no formato TigerXML.

[15] Ferramenta computacional que reconhece a estrutura sintática de uma sentença, atribuindo funções sintáticas aos constituintes reconhecidos (CARROL, 2004).

a) Os parâmetros de extração

A extração dos candidatos pelo ExATOlp foi feita a partir do *Corpus.*EaD anotado pelo *parser* PALAVRAS. A ferramenta utiliza as informações relativas aos nós não-terminais para a extração dos candidatos.

O ExATOlp dispõe de um conjunto de regras para refinar o processo. Essas regras foram elaboradas com base na estrutura típica dos SNs com o propósito de eliminar ou refinar SNs que ou são frutos de análises equivocadas do *parser* ou não têm potencial terminológico. No caso do projeto TermiNet, foram utilizadas as regras-padrão[16], a saber:

- Remover pronomes, conjunções, preposições e artigos que ocorrem no início ou final de SNs extraídos do *corpus*; no caso, os SNs são armazenados na ferramenta sem esses elementos; p.ex.: o SN "estas condições" é armazenado como "condições", "todas as crianças" como "crianças" e "dosagem diária para" como "dosagem diária";
- Recusar SNs extraídos do *corpus* que contenham numerais e/ou caracteres especiais; p.ex.: "20 anos", "seis meses", "dupla mãe/neonato" e "remédio+profilaxia"; no caso, são aceitos apenas sintagmas que possuem letras (acentuadas ou não) e hífen;
- Extrair SNs do *corpus* que ocorram como sujeitos ou complementos;
- Extrair SNs cujos núcleos sejam substantivos comuns, substantivos próprios, adjetivos e verbos no particípio passado;
- Extrair SNs implícitos por remoção de adjetivos.

Além dessas regras, é preciso especificar o tamanho dos SNs a extrair. Especificamente, o ExATOlp agrupa todos os sintagmas extraídos em dez listas, cada uma armazenada em um arquivo específico. As listas de 1 a 9 contém, respectivamente, os sintagmas compostos por 1-gramas, 2-gramas, 3-gramas, 4-gramas, 5-gramas, 6-gramas, 7-gramas, 8-gramas e 9-gramas. A lista de número 10 armazena os SNs formados por 10 ou mais unidades (Lopes *et al.*, 2009). No caso do projeto TermiNet, somente os arquivos relativos aos 1-gramas, 2-gramas, 3-gramas e 4-gramas foram considerados. No interior desses arquivos, os candidatos são ranqueados em função de sua frequência simples no *corpus*, isto é, em função do número de vezes em que o SN ocorreu no *corpus* inteiro.

b) A delimitação do ponto de corte

A especificação de um ponto de corte é essencial, pois, em geral, a extração automática gera uma lista de candidatos extensa, que inclui unidades relevantes e irrelevantes. Neste sentido, o ponto de corte reduz o tamanho das listas, excluindo o mínimo possível de candidatos relevantes.

O ExATOlp, em especial, disponibiliza as seguintes opções de ponto de corte: (i) ponto de corte absoluto segundo a frequência relativa, em que um limiar mínimo (ou seja, um número real entre 0 e 1) deve ser informado; (ii) ponto de corte absoluto segundo a frequência absoluta, em que um limiar mínimo (ou seja, um número inteiro superior a 1) deve ser informado; (iii) ponto de corte absoluto único, em que um número específico de termos (ou seja, um inteiro) deve ser informado; (iv) ponto de corte relativo, em que um percentual do número de termos (ou seja, um valor entre 0% e 100%) deve ser informado (LOPES *et al.*, 2009).

[16] As regras de token e não-token e a limpeza das listas geradas pelo NSP assemelham-se às regras estabelecidas no ExATOlp.

No projeto TermiNet, o ponto de corte escolhido foi o (ii) absoluto segundo a frequência absoluta. O cálculo usado para estabelecer esse ponto de corte é: (tamanho do *corpus*/100.000) + 1. Segundo Lopes *et al.* (2009), esse cálculo é baseado na premissa de que, em determinado *corpus*, as unidades extraídas de menor frequência não possuem valor terminológico. Geralmente, nos textos de domínios especializados, os termos ou unidades terminológicas ocorrem com frequências altas (Estopá Bagot, 1999).

No caso do *Corpus.*EaD, cujo número total de ocorrências é de 1.350.683, o ponto de corte foi estabelecido em 14, pois:

Ponto de corte → (1.350.683/100.000) + 1

Consequentemente, apenas os SNs de frequência absoluta igual ou superior a 14 ocorrências no *corpus* foram considerados, ou seja, SNs com frequência menor que 14 não foram incluídos na lista final de candidatos extraídos. Essa seleção de candidatos insere, no processo de extração, que segue claramente uma abordagem linguística, um componente estatístico.

c) O pós-processamento das listas de candidatos

Após a extração, realizou-se o pós-processamento das listas de candidatos, a fim de preparar e/ou refinar os candidatos para a validação. No caso, o pós-processamento englobou especificamente a tarefa de limpeza manual das listas. No caso, essa limpeza englobou a retirada principalmente de alguns nomes próprios e de SNs sem valor terminológico (p. ex.: "século XIX").

No Quadro 5, apresenta-se o total de candidatos contidos nas listas do ExATOlp antes e depois do processo de limpeza.

Quadro 5: Número de candidatos extraídos pelo ExATOlp.

N-grama	Pré-limpeza (corte 14)	Pós-limpeza
Unigrama	1.016	977
Bigrama	305	164
Trigrama	208	134
Quadrigrama	30	10
TOTAL	1.559	1.285

5.1.2. A validação dos candidatos

A extração dos candidatos a termo engloba a tarefa de validação desses candidatos. Para tanto, há, de certa forma, quatro estratégias de validação, as quais podem ser empregas isoladamente ou em conjunto.

A primeira consiste na avaliação da lista de candidatos por especialistas do domínio (PAZIENZA *et al.*, 2005; ALMEIDA *et al.*, 2007).

A segunda estratégia consista na comparação automática da lista de candidatos extraídos a uma lista de termos de referência do domínio em questão, comumente denominada "lista de referência dourada" (do inglês, *gold reference list*). Essa lista contém um conjunto considerável de termos do domínio atestado por especialistas e/ou linguistas. Uma lista de termos de referência pode ser elaborada especificamente para a tarefa de validação; nesse caso, ela é comumente elaborada de forma manual por linguistas e/ou especialistas do domínio com base no mesmo *corpus* do qual a lista de candidatos é extraída. Pode-se utilizar também como lista de referência o conjunto de

termos contidos em produtos terminográficos, como glossário e dicionários, desde que estes tenham sido atestados e que sejam reconhecidamente relevantes (PAZIENZA *et al.*, 2005).

A terceira estratégia consiste na comparação da lista de candidatos, associados às suas respectivas frequências no *corpus* do domínio, a uma lista de palavras de um *corpus* de referência, também associadas às suas frequências. O *corpus* de referência é aquele comumente usado para fins de contraste com o *corpus* de estudo (no caso, o *corpus* do domínio de especialidade). Em geral, espera-se que o *corpus* de referência seja aproximadamente 5 vezes maior que o de estudo, não contenha os textos do *corpus* de estudo e, de preferência, seja composto de textos que vão colocar em evidência as características peculiares do *corpus* de estudo, no caso, os termos. Dessa forma, espera-se que um candidato a termo não esteja presente na lista do *corpus* de referência ou, se estiver, espera-se que sua frequência seja menor que a apresentada no *corpus* de estudo, evidenciando, assim, seu estatuto de termo.

A quarta estratégia, por fim, consiste na aplicação de alguns linguísticos critérios pelos terminológicos. O principal critério é o da relevância semântica, ou seja, da importância (ou não) do candidato para o domínio especializado, independente da frequência com a qual o termo ocorrer no *corpus*. Além da relevância semântica, outros critérios podem ser utilizados, principalmente para verificar o grau de lexicalização dos sintagmas terminológicos e determinar os limites das unidades terminológicas sintagmáticas, por exemplo: (i) não-separabilidade dos componentes (ou seja, os elementos que compõem um termo sintagmático são marcados pela interdependência, constituindo, portanto, uma unidade conceitual única); (ii) existência de uma definição, isto é, o termo se encontra definido em dicionários especializados da área em questão; (iii) compatibilidade sistêmica, isto é, comparação do candidato a um conjunto de unidades de um sistema terminológico; (iv) substituição sinonímica, isto é, a substituição de um candidato complexo por um termo simples pode evidenciar (ou não) se os mesmos designam o mesmo conceito, (v) uso e frequência de uso, etc. (BARROS, 2007).

No TermiNet, o julgamento do especialista é a estratégia principal de validação dos candidatos a termos extraídos automaticamente a partir do *corpus.* No entanto, no caso específico da lista unificada de unigramas, uma estratégia prévia à validação pelo especialista foi adotada com o objetivo de reduzir a quantidade de unigramas. A lista unigramas foi comparada à lista de unigramas de um *corpus* de referência. O *corpus* de referência utilizado foi a porção corrigida do Corpus NILC, composta por textos jornalísticos publicados no jornal Folha de São Paulo no ano de 2005, em um total de aproximadamente 1.7 milhões de ocorrências. Da comparação entre a lista de unigramas extraídos do *Corpus.*EaD e da lista de unigramas do *corpus* de referência, foram identificados os candidatos a termos de frequência 0 no *corpus* de referência. Dos 977, 131 candidatos tinham frequência 0 no *corpus* de referência. Assim, no Quadro 6, estão descritas as listas finais de candidatos extraído pelo ExATOlp que foram submetidos ao processo de validação pelo especialista.

Quadro 6: Listas submetidas à validação.

N-gramas	Quantidade
1-grama	130
2-grama	164
3-grama	134
4-grama	10

Aqui, no entanto, cabem algumas ressaltas quanto a essa estratégia adotada para a redução da lista de unigramas. A primeira delas diz respeito ao fato de que o *corpus* de referência é composto por textos publicados em 2005 e o *Corpus.*EaD, por sua vez, é composto por textos mais atuais. Consequentemente, na comparação das listas, certos candidatos tiveram frequência zero ou muito baixa no *corpus* de referência. Esse fato foi considerado um indicativo do potencial estatuto de termo desses candidatos, posto que vários termos da EaD surgiram nos últimos anos, não estando, por conseguinte, presentes em textos datados de 2005. A segunda observação diz respeito ao fato de que a utilização dessa estratégia reduz, por um lado, o número de candidatos, validando, de certa forma, os que passam por esse filtro, mas, por outro lado, pode excluir candidatos a termos que, por serem unidades também de língua geral, tenham alta frequência no *corpus* de referência. Com base na validação, foi possível calcular as medidas tradicionais de avaliação do desempenho das ferramentas de EAT: precisão, cobertura e f-*measure* (CABRÉ *et al.,* 2001).

A precisão indica a capacidade do método/ferramenta de identificar efetivamente os termos do domínio, ou seja, os termos que foram validados pelo especialista. Essa medida é calculada pela equação em (2), em que o número de candidatos validados, extraídos por um método (e ferramenta) específico, é dividido pelo número total de candidatos extraídos por esse método/ferramenta.

$$(2) \quad \text{Precisão} = \frac{\text{Termos validados}}{\text{Termos extraídos}}$$

A cobertura indica a quantidade de termos validados extraídos pelo método/ferramenta. Essa medida é calculada pela equação em (3), em que o número de candidatos validados, extraídos por um método (e ferramenta) específico, é dividido pelo número total de termos validados. No caso, para a cômputo da cobertura, considerou-se a lista total de termos validados, composta por termos provenientes dos três métodos/ferramentas utilizados no TermiNet (cf. Quadro 3), como uma lista de referência do domínio. Especificamente, essa lista de referência é composta por 118 unigramas, 139 bigramas, 117 trigramas e 6 quadrigramas.

$$(3) \quad \text{Cobertura} = \frac{\text{Termos validados}}{\text{Total de termos validados}}$$

A medida f-*measure*, por fim, é considerada uma média harmônica entre a precisão e cobertura, sendo calculada pela equação em (4).

$$(4) \quad \text{F-measure} = \frac{2 * (\text{Precisão} * \text{Cobertura})}{\text{Precisão} + \text{Cobertura}}$$

A seguir, no Quadro 7, apresentam-se as medidas de precisão, cobertura e f-*measure* para o método linguístico/ferramenta em função do tipo de n-grama extraído.

Quadro 7: Avaliação da EAT segundo o paradigma linguístico.

Extrator	N-grama	ExATOlp
Precisão	1-grama	45% (59/130)
	2-grama	62% (102/164)
	3-grama	47% (63/134)
	4-grama	50% (5/10)
Cobertura	1-grama	50% (59/118)
	2-grama	73% (102/139)
	3-grama	53% (63/117)
	4-grama	83% (5/6)
F-*measure*	1-grama	47%
	2-grama	67%
	3-grama	49%
	4-grama	62%

6. Considerações finais

Tendo em vista o objetivo principal do projeto, que foi a extração dos candidatos a termos do *Corpus.*EaD com base no paradigma linguístico, ressalta-se este fora plenamente alcançado.

Do ponto de vista teórico-metodológico, salienta-se que o principal resultado do trabalho de IC ora descrito reside na experiência adquirira pelo grupo de pesquisa em que este está inserido no que diz respeito aos processos de EAT com base no paradigma linguístico e de validação dos candidatos. Além disso, o trabalho gerou resultados concretos para a construção da WordNet.EaD, ou seja, as listas de unigramas, brigramas, trigramas e quadrigamas, provenientes do ExATOlp, validadas pelo especialista de domínio. No caso dos unigramas, tal validação também contou com a comparação a uma lista de palavras de um *corpus* de referência.

Agradecimentos

Agradecemos à Coordenadoria de Iniciação Científica e Tecnológica da Pró-Reitoria de Pesquisa da UFSCar pela bolsa concedida no âmbito do Programa Institucional de Bolsas de Iniciação Científica – PIBITI/CNPq/UFSCar. Agradecemos também à Lucelene Lopes, desenvolvedora da ferramenta ExATOlp, pela ajuda com a manipulação do *software.*

7. Referências bibliográficas

ALLEN, J. **Natural language understanding**. Redwood City, CA: Benjamin/Cummings, 1995.

ALMEIDA, G.M.B.; CORREIA, M. Terminologia e corpus: relações, métodos e recursos. In: Stella E. O. Tagnin; Oto Araújo Vale. (Org.). **Avanços da Lingüística de Corpus no Brasil**. 1 ed. São Paulo: Humanitas/FFLCH/USP, 2008, v. 1, p. 63-93.

BARROS, L.A. **Conhecimento de Terminologia Geral para a prática tradutória**. São José do Rio Preto, SP: Editora NovaGraf, 2007.

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

AUSSENAC-GILLES, N.; SÉGUÉLA, P. Les relations sémantiques: du linguistique au formel. **Cahiers de Grammaire**, n.25, p. 175–198, 2000.

BANERJEE, S.; PEDERSEN, T. The design, implementation, and use of the N-gram Statistics Package. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 4, 2003, **Proceedings**… Mexico City, Mexico, 2003, p. 370-381.
BARROS, L. A. **Curso básico de Terminologia**. São Paulo: EDUSP, 2004.
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.
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.
CARROL, J. Parsing. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 12, p. 233-248.
CEDERBERG, S.; WIDDOWS, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 11, 2003. **Proceedings…** Edmonton, Canada, 2003, p. 111-118.
DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais**. Araraquara, 1996. 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.
___. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.
___.; DI-FELIPPO, A.; NUNES, M.G.V. The automatic mapping of Princeton Wordnet lexical-conceptual relations onto the Brazilian Portuguese Wordnet database. In: International Conference on Language Resources and Evaluation, 6, 2008. **Proceedings**... Marrakech, Morocco, 2008.
DI-FELIPPO, A.; ALMEIDA, G. M. B. Uma metodologia para o desenvolvimento de *wordnets* terminológicas em português do Brasil. **TradTerm**, n.16, p. 365-395, 2010. ISSN 0104-639X.
___.; DIAS-DA-SILVA, B. C. O Processamento Automático de Línguas Naturais enquanto Engenharia do Conhecimento Linguístico. **Calidoscópio**, São Leopoldo (RS), v.7, n.03, p. 183-191, set/dez. 2009. ISSN 1679-8740
DI-FELIPPO, A.; SOUZA, J. W. C. O projeto do *corpus* para a construção de uma *wordnet* terminológica. In: SHEPHERD, T., *et al.* (Orgs). ENCONTRO DE LINGUÍSTICA DE CORPUS, 8, 2009. **Anais**... Rio de Janeiro (RJ), 2010. *In press.*
ESTOPÁ BAGOT, R. **Extracció de terminologia: elements per a construcció d'un SEACUSE** (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada). Tese de (Doutorado) – Universidade Pompeu Fabra, Barcelona, 1999.
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.
HAYES-ROTH, F. Expert systems. In: Shapiro, E. (Ed.). **Encyclopedia of artificial intelligence**. New York, Wiley, 1990, p. 287-298.

JACQUEMIN, C.; BOURIGAULT, D. Term extraction and automatic indexing. In: MITKOV, R. (Ed.). **Handbook of Computational Linguistics**. Oxford University Press, 2003, p.599-615.
LOPES, L. *et al.* ExATOlp: An automatic tool for term extraction from Portuguese language corpora. In: LANGUAGE & TECHNOLOGY CONFERENCE LTC'09, 4, 2009, Polish. **Proceedings…** Polish, 2009, pp 427–431.
LOPES, L. *et al.*, EXATOLP - an automatic tool for term extraction from Portuguese language corpora. In: INTERNATIONAL LANGUAGE AND TECHNOLOGY CONFERENCE (LTC), 4, 2009. **Proceedings …** Poznam, Poland, 2009. p. 1-5.
LOPES, L.; OLIVEIRA, L. H. M.; VIEIRA, R. Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches. In: PROPOR, 9, 2010, Porto Alegre. **Proceedings**… Porto Alegre, 2010, p. 1-6
MAGNINI, B.; SPERANZA, M. Integrating generic and specialized wordnets. In: CONFERENCE ON RECENT ADVANCES IN NATURAL LANGUAGE, 2, 2001, Tzigov Chark. **Processing…** Tzigov Chark, Bulgaria. 2001.
MILL, D. et al (Orgs.). **Polidocência na Educação a Distância**: múltiplos enfoques. São Carlos: Edufscar, 2010. *No prelo*.
MITITELU, V.B. Automatic extraction of patterns displaying hyponym-hypernym co-cccurrence from corpora. In: CENTRAL EUROPEAN STUDENT CONFERENCE IN LINGUISTICS, 1, 2006, Budapest. **Proceedings...** Budapest, Hungary, 2006.
MORATO, J. *et al.* Wordnet applications. In: INTERNATIONAL GLOBAL WORDNET CONFERENCE, 2, 2004, Masaryk University. **Proceedings...** Masaryk University, Brno, 2004. p. 270-278.
MORIN, E.; JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. **Computer and the Humanities**, v. 38 (4), p. 343-362l, 2004.
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.
POPRAT, M. *et al.* Building a BioWordnet using Wordnet data structures and Wordnet's software infrastructure - a failure story. In: ACL WORKSHOP ON SOFTWARE ENGINEERING, TESTING, AND QUALITY ASSURANCE FOR NATURAL LANGUAGE PROCESSING, 2008, Ohio. **Proceedings…** Ohio, 2008. p.31-39.
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.
RIGAU, G. **Automatic acquisition of lexical knowledge from MRDs**. Tesis doctoral, Departament de Llenguatges i Sistemes Informàtics, UPC, Barcelona, 1998.
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.
SAGER, J.C. **Curso práctico sobre el procesamiento de la Terminologia**. Madri: Fundación Germán Sánchez Ruipérez, 1993.
SAGRI *et al.* Jur-Wordnet. In: INTERNATIONAL GLOBAL WORDNET CONFERENCE, 2, 2004. **Proceedings...** Masaryk University, Brno, 2004. p. 305-310.
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.
SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice.** Oxford: Oxbow Books, 2005. P.1-16. Disponível em: <www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>. Acesso em: 02 ago. 2010.
SUÁREZ, M.; CABRÉ, M.T. La variación denominativa en los textos de especialidad: indicios linguísticos para su recuperación automática. In: Simposio Iberoamericano de Terminología, 8, 2002. **Actas…** Cartagena de Indias, 2002. p.1-12.