O DESENVOLVIMENTO DE UMA BASE DE DADOS LÉXICO-CONCEITUAL BILÍNGUE (INGLÊS NORTE-AMERICANO/ PORTUGUÊS BRASILEIRO)

Ariani DI FELIPPO[1]
Bento Carlos DIAS-DA-SILVA[2]

- RESUMO: Para os sistemas computacionais que processam línguas naturais, como os de tradução automática, os recursos léxico-conceituais bilíngues ou multilíngues são de extrema importância. Consequentemente, o desenvolvimento de tais recursos tem ocupado lugar de centralidade no Processamento Automático das Línguas Naturais (PLN). Para o português do Brasil (PB), os recursos desse tipo ainda são escassos. Neste trabalho, apresenta-se REBECA, uma base de dados léxico-conceitual bilíngue desenvolvida para o par de línguas "inglês americano-PB" (Ingl-PB). Na introdução, contextualiza-se o projeto de desenvolvimento dessa base. Na sequência, apresentam-se (i) o equacionamento metodológico do projeto, enfatizando as atividades de pesquisa realizadas em cada uma das etapas previstas pela metodologia, (ii) a construção da base com o auxílio do editor de ontologias Protégé-OWL, (iii) as principais características e potencialidades da base REBECA e, por fim, (iv) as possíveis extensões e algumas considerações finais.

- PALAVRAS-CHAVE: Processamento automático de línguas naturais. Base de dados bilíngue inglês-português. Alinhamento léxico-conceitual. Interlíngua estruturada. MultiNet.

Introdução

Em função das aplicações reais para as quais os sistemas de Processamento Automático de Línguas Naturais são escritos, é premente a compilação de recursos lexicais monolíngues e multilíngues que sejam: (i) manipuláveis pelo sistema do qual fazem parte e (ii) linguisticamente motivados (PALMER, 2001; HANKS, 2004). A construção de bases lexicais, principalmente para o inglês (Ingl), como a WordNet de Princeton (WN.Pr) (FELLBAUM, 1998) e a FrameNet (BAKER; FILLMORE; LOWE, 1998), e para as línguas européias, como a EuroWordNet (VOSSEN, 1998) e a MultiWordNet (PIANTA; BENTIVOGLI; GIRARDI, 2002),

[1] UFSCar – Universidade Federal de São Carlos. Centro de Educação e Ciências Humanas. São Carlos – SP – Brasil. 13560-270 – arianidf@uol.com.br.

[2] UNESP – Universidade Estadual Paulista. Faculdade de Ciências e Letras – Departamento de Letras Modernas, Araraquara – SP – Brasil. 14.800-901 – bento@fclar.unesp.br.

confirma a necessidade de recursos que armazenam informações semântico-conceituais das unidades lexicais.

Nesse cenário, destacam-se os recursos multilíngues em que bases monolíngues de línguas distintas estão alinhadas por meio de uma interlíngua. A EuroWordNet e a MultiWordNet são exemplos paradigmáticos desse tipo de recurso.

O alinhamento nessas bases é feito por uma interlíngua não-estruturada, denominada *Inter-lingual-Index* (ILI), e por relações interlinguais rotuladas. Por exemplo, na Figura 1, ilustra-se que o *synset* {finger}[3] da WN.Pr está indexado ao ILI {finger} pela relação de equivalência sinonímica eq_synonym. Devido a uma diferença léxico-conceitual, o conceito expresso pelo ILI {finger} não é lexicalizado no espanhol; nesse caso, diz-se que há uma lacuna lexical no espanhol. Assim, o *synset*[4] {dedo} da WordNet espanhola liga-se ao mesmo ILI {finger} pela relação eq_has_hyponym. A principal vantagem da interlíngua não-estruturada reside na facilidade de expansão da mesma, pelo acréscimo de conceitos específicos de uma língua (p.ex.: {dedo} do espanhol). A principal desvantagem é o número elevado de *links* entre as unidades lexicais e a interlíngua que as diferenças léxico-conceituais podem causar. Na Figura 1, por exemplo, o *synset* {dedo} liga-se a dois ILIs: {finger} e {toe}.

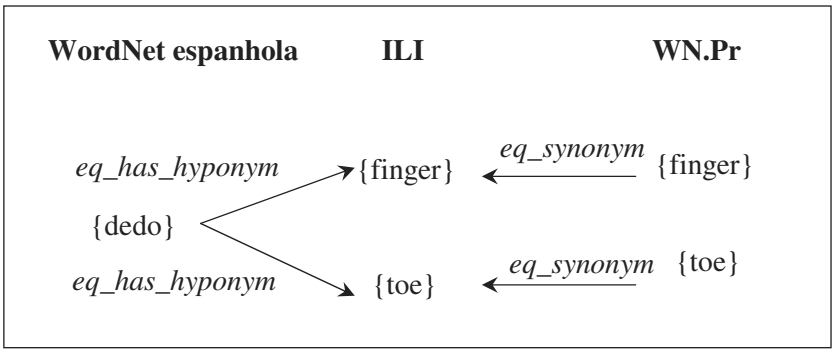

Figura 1 – Indexação léxico-conceitual na EuroWordNet e na MultiWordNet.

Para o português brasileiro (PB), o alinhamento das bases da WordNet.Br (WN.Br) e WN.Pr, que está sendo feito nos moldes da EuroWordNet (DI FELIPPO; DIAS-DA-SILVA, 2007), resultará no único recurso desse tipo que engloba o PB.

[3] Os conceitos, quando codificados em *synsets*, são representados entre parênteses; caso contrário, entre os símbolos <>.

[4] Construto criado para designar a unidade básica de estruturação da rede, isto é, um conjunto de unidades lexicais sinônimas ou quase-sinônimas que permite ao falante inferir o conceito evocado pelas unidades.

Diante desse cenário, apresenta-se aqui a base bilíngue REBECA, desenvolvida para o par de línguas Ingl-PB. Nessa base, um conjunto de conceitos lexicalizados (isto é, expressos por unidades lexicais) no Ingl está alinhado a um conjunto de conceitos lexicalizados no PB por meio de uma interlíngua estruturada.

Para a apresentação do desenvolvimento da base, apresentam-se os seguintes tópicos na sequência: (i) a concepção de PLN segundo a qual a base REBECA fora construída; (ii) a metodologia adotada para a construção dessa base e as atividades realizadas em cada etapa prevista pela metodologia; (iii) as principais características e potencialidades linguístico-computacionais da base REBECA; (iv) as possíveis extensões para a referida base e (v) algumas considerações finais sobre este trabalho.

Os estudos linguístico-computacionais da linguagem

Os sistemas que processam (interpretam/geram) língua natural (registrada em meio escrito), desenvolvidos no PLN, podem ser vistos, segundo Dias-da-Silva (1996, 2006), como um tipo especial de "sistema especialista"[5]. Isso se baseia no fato de que esses sistemas requerem uma parcela específica do conhecimento humano – o conhecimento linguístico – para realizar tarefas específicas como correção ortográfica, tradução automática, etc.

Em outras palavras, para as pesquisas que adotam a "concepção linguisticamente motivada de PLN", o computador não poderá satisfatoriamente emular uma língua natural se não conseguir, em alguma medida, compreender o assunto que está em discussão. Logo, é preciso fornecer à máquina descrições e formalizações de dados linguísticos nas dimensões: morfológica, sintática, semântico-conceitual e até mesmo pragmático-discursiva (ROCA, 2000).

Apesar dos sistemas de PLN realizarem satisfatoriamente os passos básicos de processamento da língua, eles não são capazes de "entender" o que os usuários dizem ou fazem (PALMER, 2001). Essa compreensão tem se tornado essencial para alguns sistemas que processam língua, particularmente para aqueles que processam duas ou mais línguas, como os sistemas de "tradução automática". Para tanto, é notória a necessidade de se tratar o conhecimento de nível semântico-conceitual. No caso, para "entender" ou "interpretar" as expressões linguísticas simples ou complexas (sintagmas e sentenças) de um texto, faz-se necessário o desenvolvimento de recursos bilíngues e multilíngues

[5] No âmbito da Inteligência Artificial, um sistema especialista (do inglês, *expert system*) é um sistema computacional inteligente, que toma decisões e resolve problemas referentes a um determinado campo de atuação, como finanças e medicina, utilizando conhecimento e regras analíticas definidas por especialistas no assunto (JACKSON, 1990; HAYES-ROTH, 1990; GIARRATAMO; RILEY, 2004). Um sistema de diagnóstico, por exemplo, necessita saber quais as características das doenças a serem diagnosticadas, pois, sem elas, é impossível elaborar um diagnóstico automaticamente.

que armazenam informação semântico-conceitual sobre as unidades lexicais (SAINT-DIZIER; VIEGAS, 1995; PALMER, 2001; HANKS, 2004). Tais recursos para o PB ainda são escassos.

Diante da necessidade de recursos léxico-conceituais e com base na concepção linguisticamente motivada de PLN, fora construída a base REBECA. O desenvolvimento desse recurso, cujos detalhes são fornecidos na seqüência, é visto, então, como um "exercício de engenharia da linguagem humana".

Metodologia

Para o desenvolvimento da base REBECA, tomou-se por base Dias-da-Silva (1996; 2006), que fornece os passos essenciais para o desenvolvimento de projetos na área do PLN. Para o autor, os sistemas de PLN são vistos como "sistemas especialistas" ou "sistemas baseados em conhecimento" (do inglês, *knowledge-based systems*) (GRISHMAN, 1986). Segundo essa concepção, as pesquisas nesse domínio envolvem uma "engenharia do conhecimento linguístico". Ao conceber um sistema de PLN dessa forma, Dias-da-Silva propõe que as pesquisas sigam as seguintes etapas, as quais se baseiam em Hayes-Roth (1990): "extração do solo" (isto é, explicitação dos conhecimentos e habilidades linguísticas), "lapidação" (isto é, representação formal desses conhecimentos e habilidades) e "incrustação" (isto é, o programa de computador que codifica essa representação).

A realização de uma pesquisa em PLN fatorada nessas fases ou etapas pressupõe que o corpo de conhecimento necessário à construção desse tipo de empreendimento deve ser investigado em três domínios, a saber: domínio linguístico, domínio linguístico-computacional e domínio computacional (DIAS-DA-SILVA, 1996).

A seguir, apresentam-se as atividades de pesquisa e desenvolvimento da base REBECA organizadas em função dos domínios a que pertencem.

Domínio linguístico

As atividades relativas ao domínio linguístico, em particular, ficaram especialmente concentradas nas atividades de: (i) delimitação do tipo conceitual; (ii) delimitação do domínio conceitual; (iii) compilação dos conceitos que compõem a interlíngua; (iv) identificação dos conceitos lexicalizados e a subsequente montagem da base monolíngue do Ingl; (v) investigação e identificação dos conceitos lexicalizados e a montagem da base monolíngue do PB. A seguir, cada uma dessas etapas é descrita.

Delimitação do tipo conceitual

Nessa etapa, era preciso delimitar quais tipos de conceito seriam armazenados na base (p.ex.: aqueles expressos por nomes, verbos, adjetivos, etc.). Decidiu-se por armazenar apenas os conceitos do tipo "objeto concreto discreto". Segundo Lyons (1977), os conceitos desse tipo são entidades de primeira ordem e, por isso, intuitivamente categorizam referentes perceptíveis pelos sentidos, localizadas no tempo e no espaço, que são contáveis e indivisíveis. Quanto à expressão linguística, tais conceitos realizam-se por expressões nominais, sejam elas simples, compostas ou complexas. A escolha pelos objetos concretos discretos justifica-se pela sua natureza hierárquica, que torna passível uma sistematização formal desses objetos.

Delimitação do domínio conceitual

Partindo-se do princípio de que os conceitos não estão isolados na mente, mas sim organizados (CRUSE, 2004), delimitou-se o domínio conceitual "veículo com roda" (no inglês, *wheeled vehicle*). A escolha desse domínio não se justifica por questões teóricas, mas sim práticas; no caso: delimitação bem-definida e extensão reduzida.

Compilação dos conceitos constitutivos da interlíngua

O conjunto dos conceitos constitutivos da interlíngua foi manualmente extraído da WN.Pr (2.1). Precisamente, foram selecionados todos os *synsets* organizados sob o *synset* {wheeled vehicle}. A escolha da WN.Pr como fonte dos conceitos teve três motivações principais. A primeira diz respeito ao fato de que a WN.Pr, organizada em campos conceituais, engloba o campo "veículos com rodas". A segunda foi o fato de que a WN.Pr é uma rede semântica e, por isso, seus conceitos/*synsets* podem ser reestruturados em termos do modelo de representação MultiNet, segundo o qual a interlíngua da base REBECA foi formalmente representada. No total, foram obtidos 217 conceitos. Para cada conceito da interlíngua, foi elaborada uma glosa (ou seja, uma definição informal) em PB com base principalmente nos dicionários monolíngues do Ingl (LANDAU, 2001; SUMMERS, 2005).

A identificação dos conceitos lexicalizados e a montagem da base monolíngue do Ingl

Com base nos referidos dicionários monolíngues do Ingl, foi possível identificar que, dos 217 conceitos da interlíngua, 12 não são efetivamente lexicalizados no Ingl (p.ex.: *self-propelled vehicle*; no PB, *veículo autopropulsado*),

ou seja, as expressões linguísticas que compõem os seus respectivos *synsets* não são entradas ou subentradas em tais dicionários. Ressalta-se que a ausência de uma expressão no PB para os 12 conceitos não-lexicalizados (p.ex.: *self-propelled vehicle*; no PB, *veículo autopropulsado*) não caracteriza lacuna lexical. Assim, a base monolíngue do Ingl é composta pelos 205 conceitos da interlíngua que são lexicalizados no Ingl. Tais conceitos são os próprios *synsets* da WN.Pr. Ressalta-se que, para cada unidade lexical constitutiva de um *synset* do Ingl, uma frase-exemplo (isto é, sentença que fornece o contexto de uso mínimo) fora manualmente extraída ou da WN.Pr ou da *Web*. Para a extração da *Web*, utilizou-se o portal WebCorp[6], que pode ser definido, em linhas gerais, como um conjunto de ferramentas que permite o acesso à *Web* como um *corpus* (ou seja, como uma coleção de textos a partir dos quais fatos sobre a língua podem ser observados e extraídos).

A investigação dos conceitos lexicalizados e a montagem da base monolíngue do PB

Nessa fase, foi preciso investigar e identificar os conceitos pertencentes ao domínio em questão que são lexicalizados no PB. Isso se deve ao fato de que não há uma sistematização desses dados para o PB. Tal identificação teve o Ingl como língua-fonte. Com base na delimitação informal dos conceitos realizada por meio da elaboração de glosas, as unidades do PB foram manualmente identificadas e extraídas, em uma primeira fase, de dicionários bilíngues Ingl-PB (HOUAISS; CARDIM, 1982; WEISZFLOG, 2000). Em uma segunda fase, dicionários monolíngues (WEISZFLOG, 1998, FERREIRA, 2004; HOUAISS; VILLAR; FRANCO, 2001) e de sinônimos (BARBOSA, 2000; FERNANDES, 1997) foram manualmente consultados para a identificação de unidades sinônimas e subsequente montagem dos *synsets*. Em uma terceira etapa, verificou-se manualmente a ocorrência de uso das unidades extraídas dos recursos lexicográficos em *corpora*. Essa verificação foi feita porque, por vezes, as unidades extraídas de tais recursos estão em desuso. Para tanto, foram utilizados os corpora: PLN-BR FULL[7] e textos disponíveis na *Web*. Os textos em PB disponíveis na *Web* foram consultados através do motor de busca Google[8], lançando-se mão do recurso de restrição das buscas às páginas do Brasil. Dos mesmos *corpora*, foram extraídas as frases-exemplo para as unidades lexicais.

---

[6] http://www.webcorp.org.uk/index.html

[7] O PLN-BR FULL é um *corpus* do gênero informativo (e subgênero jornalístico) composto por textosdo jornal a Folha de São Paulo, mais especificamente, por textos publicados em apenas um mês de cada ano, no intervalo de 1994 a 2005. No total, o PLN-BR FULL contém aproximadamente 29 milhões de palavras e está disponível para consultas na *webpage* do Philologic (isto é, ferramenta *Web* para buscas, recuperação e análise de *corpora*).

[8] http://www.google.com.br/

Além das unidades lexicais, foram identificados os chamados "sintagmas livres recorrentes" (SLRs) (do inglês, *recurrent free phrases*) do PB (BENTIVOGLI; PIANTA, 2004). Por exemplo, o conceito "caminhão grande destinado ao transporte de cargas pesadas; usualmente sem laterais", expresso no Ingl por *lorry*, é expresso no PB pelo SLR *caminhão de carga*. De modo geral, os SLRs são combinações livres frequentes e são importantes para o tratamento computacional das "lacunas lexicais", uma vez que proveem expressões correspondentes para conceitos que não são lexicalizados (BENTIVOGLI; PIANTA, 2004). Os SLRs formam um conjunto próprio, um *phraset*, sendo que, para cada SLR, uma frase-exemplo também fora selecionada dos referidos *corpora*. Dos 205 conceitos lexicalizados no Ingl que pertencem ao domínio "veículo com roda", apenas 84 estão lexicalizados no PB, o que equivale aproximadamente a 40,9% do total de conceitos analisados. Dessa forma, nota-se que, no domínio conceitual "veículo com roda", menos da metade dos conceitos analisados são lexicalizados no PB. Para os demais 121 conceitos (ou 59,1%), o PB apresenta lacunas lexicais, ou seja, o PB não possui unidades lexicais para expressar tais conceitos. Tais dados estão sistematizados na Tabela 1.

Dentre os 84 conceitos lexicalizados no PB e codificados em termos de *synsets*, 11 deles possuem um *phraset* sinônimo como informação adicional, o que equivale a 13% do total de conceitos que o PB lexicaliza. Os demais 73 (ou 87%) não possuem *phraset* sinônimo. Dentre as 121 lacunas, observa-se que, em 40 casos, foi possível identificar um *phraset* que expressa no PB o conceito que é expresso por unidades lexicais no Ingl. Em outras palavras, pode-se dizer que, para 33% das lacunas, foi possível montar um conjunto de SLRs. Para as demais 81 lacunas, não foi possível identificar SLRs correspondentes, o que equivale a 67% do total de lacunas lexicais identificadas no PB. Na Tabela 2, estão descritos alguns exemplos de lacunas no PB. Para dois deles, não foi possível identificar um *phraset* correspondente.

Tabela 1 – As estatísticas das lexicalizações identificadas no PB.

Descrição	***Quant.***	***Porcentagem***
Conceitos lexicalizados no PB (synsets)	84	40,9% (de 205)
com phraset sinônimo	11	13% (de 84)
Gaps	121	59,1% (de 205)
com phraset sinônimo	40	33% (de 121)

Tabela 2 – Alguns casos de lacuna no PB.

Conceito	*Glosa*	*Phraset*
<lorry>	"carroça grande e baixa sem laterais"	-
<funny wagon>	"ambulância usada para transportar pacientes de e para hospitais psiquiátricos"	-
<cattle car>	"vagão de carga fechado usado para transportar gado"	{vagão gaiola; vagão de gado}
<sound truck>	"caminhão equipado com alto-falantes, usado para fazer propaganda"	{caminhão de som}

Domínio linguístico-computacional ou representational

Como mencionado, os *synsets* do Ingl e do PB identificados no domínio linguístico foram alinhados em função dos conceitos que expressam. Esse alinhamento foi feito por meio de uma interlíngua estruturada, ou seja, pela formalização dos 217 *synsets* extraídos da WN.Pr. Tal alinhamento, juntamente como a inserção das glosas e frases-exemplo, deu origem à base REBECA.

Assim, nesse domínio, as atividades de pesquisa concentram-se principalmente na escolha do formalismo de representação do conhecimento semântico e, consequentemente, na especificação da arquitetura da base de dados.

A arquitetura da base REBECA e sua interlíngua

Para a representação formal dos conceitos da interlíngua, escolheu-se o modelo de representação do conhecimento (RC) denominado MultiNet (HELBIG, 2006) (do inglês, *Multilayered Extended Semantic Networks*).

O paradigma de representação do conhecimento MultiNet

Ao conceber o PLN como uma espécie de "engenharia do conhecimento linguístico", as atividades nesse domínio podem ser beneficiadas pelas estratégias da Engenharia do Conhecimento. Seguindo essa concepção, adotou-se o modelo de RC MultiNet (HELBIG, 2006), que se baseia na metalinguagem formal das redes semânticas e cujos construtos básicos estão ilustrados na Figura 2.

O MultiNet tem sido empregado principalmente como interlíngua semântica para recuperação de informação na *Web* por meio de interfaces em língua natural (LEVELING, 2004).

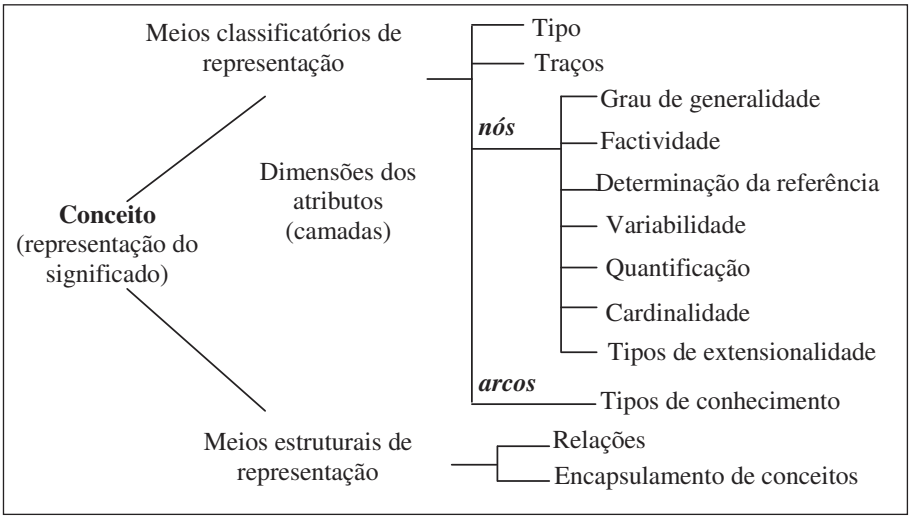

Figura 2 – Os construtos de representação do MultiNet.

A escolha do MultiNet pautou-se principalmente nos critérios de: (i) homogeneidade, isto é, seus meios de representação são capazes de expressar conceitos subjacentes a unidades lexicais, sintagmas e sentenças; e (ii) adequação cognitiva, isto é, todo conceito tem uma representação única por meio da qual toda a informação a ele associada torna-se acessível. Segundo o MultiNet, cada conceito da interlíngua fora representado em função dos construtos da Figura 2, os quais são responsável pela macro e microestruturação da interlíngua.

O MultiNet e a macroestrutura da interlíngua

Tendo em vista a adoção do MultiNet, a interlíngua da base REBECA é, na verdade, uma rede semântica, composta por nós (conceitos) e arcos (relações). Os meios estruturais do MultiNet, ou seja, as relações e o encapsulamento de conceitos, são responsáveis pela macroestrutura da rede. No caso do tipo de conceito escolhido para ser armazenado, a relação SUB (subsunção), responsável pela organização hierárquica, é a mais importante para organizar tais conceitos. Assim, do ponto de vista da macroestrutura, a interlíngua está organizada exclusivamente em função dessa relação.

Além de SUB, os conceitos da interlíngua estão especificados pelas relações PARS (parte-todo) e PURP (propósito), também consideradas fundamentais para

a caracterização do tipo de conceito sob análise. As relações SUB, PARS e PURP de cada conceito da interlíngua também foram extraídas da WN.Pr. Os conceitos relacionados por PARS e PURP, no entanto, não fazem propriamente parte da interlíngua; eles são especificações dos conceitos que constituem a interlíngua.

O encapsulamento de conceitos, por sua vez, garante que o conhecimento estabelecido por um tipo de relação seja adequadamente herdado pelos nós/conceitos mais específicos. Por exemplo, se o conceito codificado pelo *synset* {car, auto, automobile, machine, motorcar} estiver associado a {air bag} através de PARS, os conceitos hipônimos de {car, auto, automobile, machine, motorcar} herdam essa relação. Isso acontece porque a relação PARS é tida como conhecimento prototípico, o qual é herdado por *default* pelos conceitos mais específicos.

O MultiNet e a microestrutura da interlíngua

Os meios classificatórios são responsáveis pela microestrutura da rede, ou seja, pela representação interna de cada nó/conceito. Tais meios dividem-se em: "tipo conceitual", "traços semânticos" e "atributos multidimensionais". O tipo conceitual indica a classe mais geral a que o conceito pertence. No caso, os conceitos do domínio "veículo com roda" são do tipo [mov-art-discrete], ou seja, conceitos cujos referentes são objetos do tipo móvel, manufaturado e não contínuo. Assim, todo conceito da interlíngua está associado ao tipo conceitual cujo valor é [mov-art-discrete]. Além dos tipos, o MultiNet conta também com traços (do inglês, *features*), que desempenham papel fundamental na classificação dos objetos e na análise sintático-semântica. Os traços facilitam a formulação de restrições de seleção e da subcategorização dos itens lexicais. No caso, os conceitos do tipo [mov-art-discrete] estão associados aos traços [ARTIF+], [INSTRU+] e [MOVABLE+]. Consequentemente, todo conceito da interlíngua também está associado a esses traços semânticos.

A característica essencial do MultiNet é o conjunto de atributos multidimensionais especificado para os nós e arcos, os quais buscam capturar aspectos extensionais e intensionais do significado das línguas naturais (HELBIG, 2006). Os atributos dos nós são: (a) grau de generalidade (GENER); (b) factividade (FACT); (c) determinação da referência (REFER); (d) variabilidade (VARIA); (e) quantificação (QUANT); (f) cardinalidade (CARD); e (g) extensionalidade (ETYPE). O atributo do arco, em especial, é denominado tipo de conhecimento (K-TYPE). Tais atributos têm vários valores. Como os conceitos que pertencem à interlíngua são tidos como genéricos (p.ex.: <carro>), eles são especificados pelos seguintes pares de atributo-valor: [GENER=ge], [REFER=refer], [VARIA=con] e [FACT=real]. O valor *ge* de GENER indica a natureza genérica do conceito. O valor *refer* de REFER indica que esse tipo de conceito não determina a referência; ele é relacionado a

um elemento prototípico não-especificado. O valor *con* de VARIA indica que esse tipo de conceito não varia no nível pré-extensional[9]. Já o valor *real* de FACT indica que os conceitos em questão fazem referência a objetos reais. Por fim, o tipo de extensionalidade dos conceitos genéricos é geralmente [ETYPE=0], posto que a descrição no nível pré-extensional de um conceito genérico x é um elemento prototípico do conjunto <todos os X>. Quanto ao atributo do arco, ressalta-se que o arco relativo à relação SUB é rotulado por K (do alemão, *Kategorisch*), indicando que o conhecimento é categorial e, por isso, herdado sem nenhuma exceção por todos os subconceitos. Os arcos relativos às relações PARS e PURP são rotulados por D (do inglês, *default knowledge*), indicando que o conhecimento é prototípico e, por isso, herdado como conhecimento padrão.

Na Figura 3, o conceito <cart>[10] (no PB, *carroça*), elemento constitutivo da interlíngua, é representado segundo os pressupostos do MultiNet.

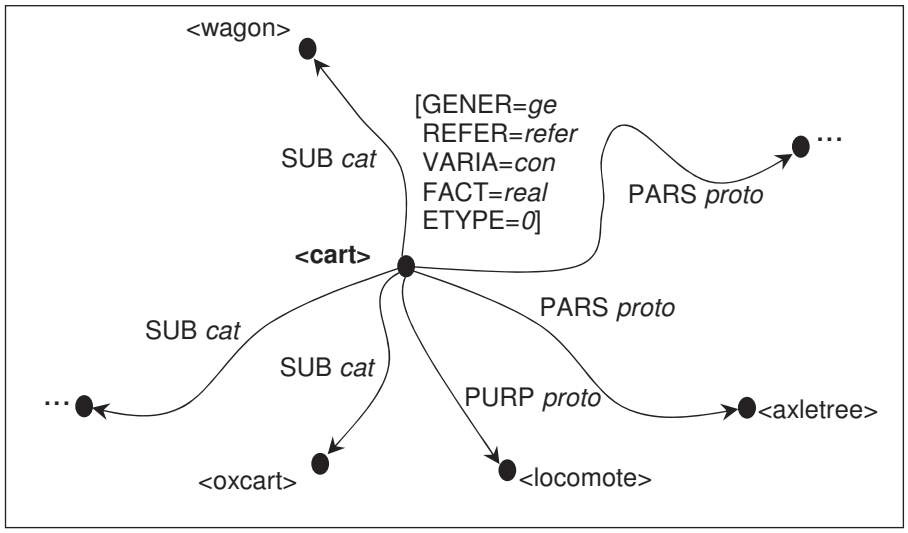

Figura 3 – Representação de um conceito lexicalizado segundo o MultiNet.

Vale ressaltar aqui que, uma vez representados por um modelo de RC (o MultiNet), a interlíngua caracteriza-se como uma "ontologia", ou seja, "uma especificação formal de uma conceitualização compartilhada" (GRUBER, 1995; BORST, 1997). Nessa definição, "formal" significa que os conceitos estão descritos explicitamente; "conceitualização" significa que uma ontologia fornece uma visão

[9] O MultiNet distingue dois planos de representação básicos: o "plano intensional" e o "plano pré-extensional". O primeiro modela as relações entre os conceitos e o segundo modela os conceitos e suas extensões. No plano pré-extensional, são modeladas, por exemplo, a cardinalidade das extensões e as relações entre essas extensões, modeladas pela teoria dos conjuntos.

[10] Vale ressaltar que os rótulos em inglês dos nós (conceitos) da interlíngua são apenas recursos mnemônicos; o rótulo para o conceito <cart>, por exemplo, poderia ser um mero código como C1 (Conceito 1).

simplificada do conhecimento; "compartilhada" significa que a visão simplificada é consensual. Assim, pode-se dizer que a base REBECA utiliza uma ontologia como interlíngua.

As atividades de domínio implementacional

No domínio da implementação, as atividades foram subdivididas em três tarefas bastante distintas.

A primeira, eminentemente computacional, consistiu na escolha de uma ferramenta computacional para a montagem da base de dados. Essa ferramenta desempenhou a função específica de editor, que possibilitou ao linguista inserir e editar as informações da interlíngua e das duas bases monolíngues. A segunda tarefa, essencialmente linguística, concentrou-se na inserção do conhecimento léxico-conceitual no editor, para a qual, aliás, foi preciso realizar certas adaptações dos construtos do editor aos tipos de informação que compõem a base REBECA. A terceira, por fim, também computacional, consistiu na investigação da possibilidade de se gerar uma representação gráfica da base REBECA, que facilitaria, por exemplo, a identificação das diferenças léxico-conceituais entre as línguas em questão.

A seguir, cada uma dessas tarefas é relatada.

A construção da base REBECA no editor Protégé-OWL

Para a construção da base REBECA, utilizou-se um dos editores de ontologia mais difundidos na literatura, o Protégé[11]. Especificamente, utilizou-se a versão desenvolvida com base na linguagem OWL[12]. Esse editor fora escolhido principalmente por sua: (i) interoperabilidade, que busca consentir a compatibilidade com outros sistemas de representação do conhecimento, (ii) usabilidade, que busca garantir a facilidade de uso da ferramenta, e (iii) aplicabilidade, que busca garantir o emprego diversificado das bases por meio da exportação das mesmas em diversos formatos ou linguagens.

Para a utilização do Protégé-OWL, algumas adaptações foram feitas para que as informações especificadas no domínio linguístico pudessem ser adequadamente inseridas. Tais adaptações foram:

---

[11] http://protege.stanford.edu

[12] A OWL é a mais recente linguagem desenvolvida pelo *World Wide Web Consortium* (W3C) (http://www.w3.org/) para promover a *Web* Semântica, que consiste em uma proposta de estruturação dos documentos da *Web*. Nesse cenário, a OWL foi projetada como anotação-padrão para o conteúdo semântico a ser disponibilizado na *Web*.

(i) os conceitos da interlíngua/ontologia foram inseridos como "classes";

(ii) os demais conceitos, que se vinculam aos da interlíngua pelas relações de PARS e PURP, e os atributos multidimensionais foram inseridos como "propriedades" das classes; mais especificamente, as relações PARS e PURP foram inseridas enquanto ObjectProperty e os atributos multidimensionais enquanto DatatypeProperty;

(iii) as expressões linguísticas, ou seja, unidades lexicais que compõem os *synsets* do Ingl e do PB e os SLRs que compõem os *phrasets* do PB, foram inseridas como "instâncias" ou "indivíduos" das classes;

(iv) as glosas foram inseridas como "comentários" das classes (conceitos);

(v) as frases-exemplo foram inseridas como "comentários" das instâncias (unidades lexicais ou SLRs).

A visualização gráfica da interlíngua e das expressões linguísticas de seus conceitos constitutivos

Na Figura 4, apresenta-se a interface de visualização gráfica do editor Protégé-OWL. Nessa figura, exibem-se um dos 217 conceitos da interlíngua da base REBECA e as expressões linguísticas desse conceito no Ingl e no PB.

Essa exibição é possível devido ao *plug-in*[13] de visualização TGVizTab, que permite aos usuários visualizar a ontologia de conceitos por meio de representações gráficas dinâmicas e interativas, contribuindo, por conseguinte, para a compreensão da estrutura ontológica, análise das relações, etc. O TGVizTab (do inglês, *TouchGraph Visualisation Tab*) (ALANI, 2003), que equivale a uma aba na interface principal do Protégé-OWL (círculo vermelho da Figura 4), baseia-se na tecnologia denominada TouchGraph, que oferece vários recursos de visualização de uma rede conceitual, como alto grau de interação, rápida renderização[14], visão panorâmica e *zoom*, entre outros[15].

---

[13] Pequenos programas de computador que servem normalmente para adicionar funções a outros programas maiores, provendo alguma funcionalidade especial ou muito específica (MICROSOFT PRESS, 1998, p.583). Mais informações sobre os vários *plug-ins* que podem ser associados ao Protégé podem ser encontradas no endereço: http://protege.stanford.edu/download/plugins.html

[14] O termo *renderização* pode ser entendido como a produção de uma imagem gráfica a partir de um arquivo de dados em um dispositivo de saída, como um monitor ou impressora (MICROSOFT PRESS, 1998, p.633).

[15] Tais recursos, aliás, têm sido considerados fundamentais para a visualização de redes conceituais extensas. Os recursos do TGVizTab aplicam-se sobre uma visualização que se baseia na técnica denominada *spring-layout*, no qual os nós (classes ou conceitos) se repelem e os arcos ou arestas (relações) atraem os nós (ALANI, 2003). Dessa forma, os nós semanticamente similares ficam dispostos próximos uns aos outros. A tecnologia TouchGraph tem sido empregada em várias aplicações, como o GoogleBrowser, responsável por exibir páginas *Web* relacionadas, e o AmazonBrowser, responsável por exibir em grafo itens de compra similares, entre outros.

Na Figura 4, observa-se, especificamente, o conceito <wheeled vehicle> como nó central da rede, juntamente com os conceitos a ele imediatamente relacionados, e as expressões que atualizam esse conceito no PB e no Ingl.

Vale ressaltar que, no campo do editor denominado ClassBrowser, exibe-se a hierarquia conceitual em formato arbóreo. Além disso, para uma visualização mais direta das diferenças léxico-conceituais entre o PB e o Ingl, os nós que representam graficamente os conceitos lexicalizados no PB foram destacados pela cor amarela. Os nós em azul representam os conceitos não-lexicalizados nem mesmo no Ingl; para esses conceitos, a ausência de unidades lexicais no PB não fora contabilizada como lacuna lexical. Os demais nós, por exclusão, indicam os conceitos não lexicalizados no PB.

Quando um conceito é selecionado no grafo, a lista das expressões linguísticas associadas a ele é mostrada no campo denominado InstanceBrowser (retângulo vermelho inferior da Figura 4). No caso da Figura 4, observa-se que o conceito <wheeled vehicle> realiza-se no Ingl por meio da unidade lexical *wheeled vehicle*, a qual constitui o *synset* unitário {wheeled vehicle}. No PB, tal conceito não é lexicalizado, sendo expresso pelo SLR *veículo com roda*, o que resulta em uma lacuna lexical. O SLR *veículo com roda*, de forma análoga à unidade *wheeled vehicle*, constitui o *phraset* unitário {veículo com roda}. Em outras palavras, pode-se dizer que o *synset* {wheeled vehicle} e o *phraset* {veículo com roda} estão indexados ao mesmo conceito da interlíngua.

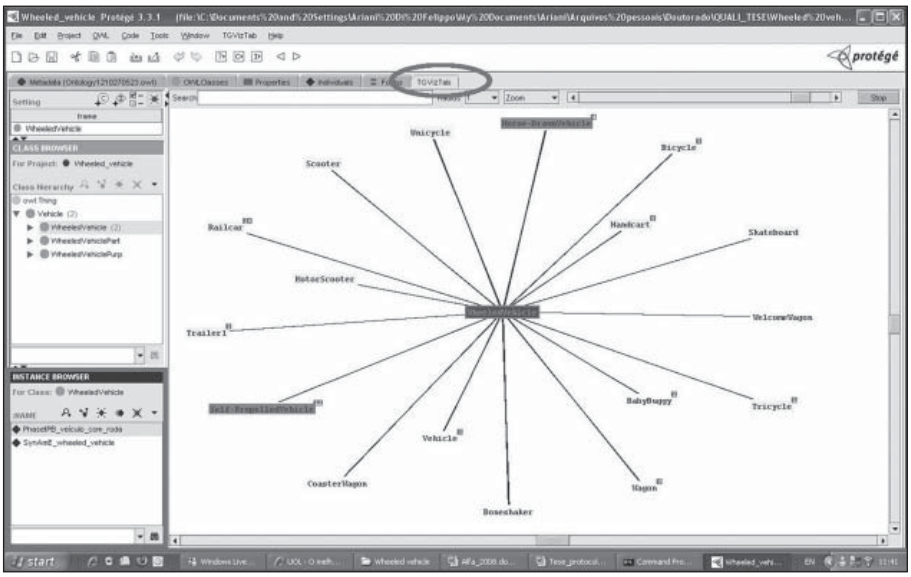

Figura 4 – A *interface* do *plug-in* TGVizTab, exibindo o conceito <wheeled vehicle> no centro do grafo.

As principais características e potencialidades da base REBECA

De um modo geral, a base REBECA caracteriza-se, nos moldes da EuroWordNet e MultiWordNet, por: (i) armazenar conceitos lexicalizados e, por isso, capturar as lexicalizações e as relações entre as unidades lexicais do PB; (ii) fornecer definições informais para cada conceito da interlíngua e (iii) fornecer uma frase-exemplo para cada unidade lexical de ambas as línguas e para os SLRs do PB. A base REBECA diferencia-se dessas outras bases por (i) utilizar uma interlíngua hierarquicamente estruturada e formal e (ii) englobar apenas conceitos do tipo "objeto concreto discreto" e pertencentes ao domínio dos "veículos com rodas".

Quanto ao alinhamento, em especial, ressalta-se que a inserção no Protégé-OWL (i) dos conceitos da interlíngua como "classes" hierarquicamente organizadas e (ii) das unidades lexicais (ou *synsets*) do Ingl e do PB e dos SLRs do PB (ou *phrasets*) como "instâncias" das "classes" permitiu que os elementos constitutivos de cada base monolíngue fossem indexados a um único conceito da interlíngua, evitando-se o número excessivo de *links*, característico do uso de uma interlíngua desestruturada. No entanto, a expansão da interlíngua torna-se um pouco mais complicada, pois requer uma reestruturação da mesma. Ressalta-se ainda que, nos casos em que há lacunas no PB, a base REBECA é capaz de fornecer dois tipos de expressões linguísticas alternativas: os SLRs e a(s) unidades lexicais (ou SLRs) que expressam um conceito hiperônimo.

Na Figura 5, por exemplo, observa-se que os conceitos <cabin car> e <baggage car> não são lexicalizados no PB, configurando lacunas lexicais nessa língua ("GAPs"). Nessa Figura, as setas mais espessas, uma pontilhada e outra contínua, indicam os caminhos para a identificação das expressões linguísticas alternativas para essas lacunas. No caso de <baggage car>, é possível, a partir das expressões do Ingl (p.ex.: *baggage car*), chegar ao SLR *vagão bagageiro* do PB por meio da interlíngua, posto que *baggage car* e *vagão bagageiro* são as instâncias das bases monolíngues do Ingl e do PB, respectivamente, que estão indexadas ao mesmo conceito da interlíngua (<baggage car>). No caso de <cabin car>, não há um SLR correspondente no PB. No entanto, devido à estruturação da interlíngua, é possível, a partir das expressões do Ingl (p.ex.: *cabin car*), percorrer a hierarquia conceitual e identificar que, no nível superior, o conceito <railcar> é lexicalizado no PB, expresso especificamente por *carro* e *vagão*.

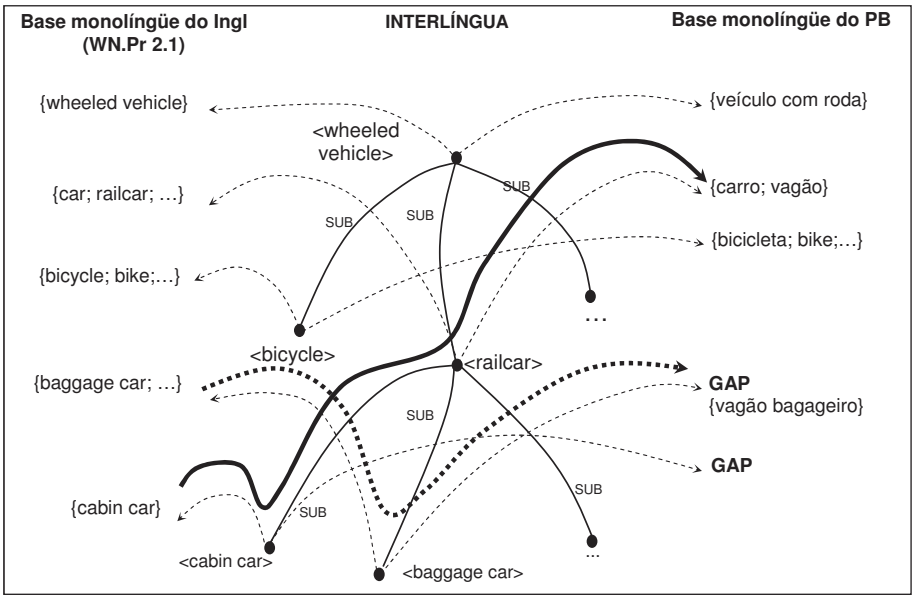

Figura 5 – Os alinhamentos léxico-conceituais na base de dados REBECA.

Dessa forma, sob o ponto de vista linguístico, vê-se que a base REBECA propicia a observação das diferenças nos padrões de lexicalização entre as línguas e no relacionamento léxico-conceitual interno às línguas, pois tais diferenças e relacionamentos ficam evidentes no alinhamento à interlíngua (Figura 5). Consequentemente, sob o ponto de vista tecnológico, evidencia-se seu potencial de uso em várias aplicações do PLN, por exemplo, na recuperação de informação multilíngue, pela expansão de unidades lexicais de uma língua a unidades lexicais relacionadas em outra língua via a interlíngua estruturada.

Extensões

Para a ampliação da base REBECA, propõe-se: (i) o refinamento do domínio conceitual "veículo com roda", (ii) a inclusão dos conceitos "específicos" do PB, e (iii) a inclusão de outros domínios conceituais.

A tarefa (i) pressupõe a identificação de conceitos que ainda não estão armazenados na WN.Pr. Essa identificação poderá consistir na extração de conceitos a partir de *corpora* e poderá ser feita com o auxílio do *plug-in* do Protégé-OWL denominado OntoLT (BUITELLAR, 2004). Tal tarefa poderá contar também com recursos computacionais e lexicográficos do Ingl.

A tarefa (ii) é semelhante à (i) e consistirá na extração de conceitos (e unidades lexicais) lexicalizados especialmente no PB a partir de *corpora*; tal extração

poderá ser feita com o auxílio do *plug-in* OntoLP (RIBEIRO JUNIOR, 2008), que é a adaptação do OntoLT para o tratamento de textos em PB. O OntoLP caracteriza-se pelas tarefas semi-automáticas de extração de unidades lexicais a partir de *corpus* anotado linguisticamente (morfossintaticamente) e identificação de possíveis relações semântico-conceituais entre elas. Para a extração das unidades lexicais e relações semântico-conceituais, o OntoLP utiliza os métodos linguístico (baseado em reconhecimento de padrões léxico-sintáticos) e estatístico (baseado em métricas estatísticas). Uma vez inseridos na interlíngua, o alinhamento do Ingl a esses conceitos específicos no PB poderá resultar na identificação da sua lexicalização ou de lacunas no Ingl.

Quanto à atividade (iii), ressalta-se que a metodologia aplicada na investigação do domínio "veículo com roda" poderá ser empregada na investigação de outros domínios conceituais (p.ex.: o dos recipientes, dos alimentos, etc.). Essa metodologia, que se baseia especialmente em informações extraídas de recursos lexicográficos, poderá ser estendida pela utilização de informações provenientes de corpora, por meio da utilização do OntoLP.

Considerações finais

A construção da base REBECA reflete os primeiros resultados da investigação sobre os padrões de lexicalização (isto é, associação entre um conceito e uma unidade lexical) do Ingl e do PB no âmbito do desenvolvimento de uma base léxico-conceitual bilíngue. É reconhecido que a identificação de tais padrões e o subsequente alinhamento dos conceitos lexicalizados contribui para o tratamento computacional dos problemas causados pelas diferenças léxico-conceituais. Com a extensão da base REBECA, buscar-se-á contribuir diretamente para o tratamento computacional do par de línguas Ingl-PB em aplicações como tradução automática e/ou recuperação de informação multilíngue. Além disso, a pesquisa que resultou na base REBECA busca promover a visão linguisticamente motivada das atividades do PLN e, consequentemente, fortificar o trabalho colaborativo entre cientistas e engenheiros da linguagem.

Agradecimentos

Ao CNPq, pelo financiamento da pesquisa da qual este trabalho é parte.

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. The development of a bilingual (North-American English and Brazilian Portuguese) lexical-conceptual database. *Alfa*, São Paulo, v.53, n.1, p.77-97, 2009.

▪ *ABSTRACT: The natural languages processing in some applications (e.g. machine*

translation) requires bilingual or multilingual lexical-conceptual resources. Accordingly, one of the main issues of Natural Language Processing research has been the development of such resources. In particular, there are few resources of this kind available for Brazilian Portuguese (BP). In this scenario, this paper presents REBECA, a bilingual lexical-conceptual database for BP and (North-American) English. Accordingly, after contextualizing the project, it is presented (i) the natural language processing framework in which the database is couched, (ii) the methodology that has been applied to the development of REBECA, (iii) the construction of REBECA itself with the help of the Protégé-OWL ontology editor, and (iv) the main features and potentialities of REBECA. Finally, we sketch some future work and present the final considerations.

- *KEYWORDS: Natural language processing. Bilingual database. Lexical-conceptual alignment. Structured Interlingua. MultiNet.*

REFERÊNCIAS

ALANI, H. *TGVizTab*: an ontology visualisation extension for Protégé. In: WORKSHOP ON VISUALIZATION INFORMATION IN KNOWLEDGE ENGINEERING, 2., 2003, Sanibel Island. *Proceedings...* Sanibel Island: VIKE, 2003. p.01-06.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley FrameNet project. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 17., 1998, Montreal. *Proceedings...* Montreal: COLING/ACL, 1998. p.86-90.

BARBOSA, O. *Grande dicionário de sinônimos e antônimos*. Rio de Janeiro: Ediouro, 2000.

BENTIVOGLI, L.; PIANTA, E. Extending wordnet with syntagmatic information. In: GLOBAL WORDNET CONFERENCE, 2., 2004, Brno. *Proceedings...* Brno: Association Masaryk University, [2004]. p.47-53.

BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. 1997. 227f. Thesis (Doctor) – Universiteit Twente, Enschede, 1997. Disponível em: <http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf>. Acesso em: 05 abr. 2006.

BUITELLAR, A. et al. A Protégé plug-in for ontology extraction from text based on linguistic analysis. In: EUROPEAN SEMANTIC WEB SYMPOSIUM, 1., 2004, Heraklion. *Proceedings...* Heraklion: ESWS, [2004]. p.31-44.

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

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. Towards an Automatic Strategy for Acquiring the WordNet.Br Hierarchical Relations. In: WORKSHOP IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 5., 2007, Rio de Janeiro. *Proceedings...*

Rio de Janeiro: TIL, [2007].

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

_____. *A face tecnológica dos estudos da linguagem*: o processamento automático das línguas naturais. 1996. 272f. Tese (Doutorado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.

FELLBAUM, C (Ed.). *WordNet*: an electronic lexical database. Cambridge: MIT Press, 1998.

FERNANDES, F. *Dicionário de sinônimos e antônimos da língua portuguesa*. São Paulo: Globo, 1997.

FERREIRA, A. B. H. *Novo dicionário eletrônico Aurélio da língua portuguesa*. Curitiba: Positivo, 2004. 1 CD-ROM.

GIARRATANO, J. C.; RILEY, G. D. *Expert systems*: principles and programming. Boston: Course Techonology, 2004.

GRISHMAN, R. *Computational linguistics*: an introduction. Cambridge: Cambridge University Press, 1986.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, New York, v.43, n.5-6, p.907-928, 1995.

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. S. C. (Ed.). *Encyclopedia of artificial intelligence*. New York: J. Wiley& Sons, 1990. p. 287-298.

HELBIG, H. *Knowledge representation and the semantics of natural language*. Berlin: Springer-Verlag, 2006.

HOUAISS, A.; VILLAR, M. de S.; FRANCO, F. M. de M. *Dicionário eletrônico Houaiss da língua portuguesa*. Rio de Janeiro: Objetiva, 2001. 1 CD-ROM.

HOUAISS, A.; CARDIM, I. (Org.). *Dicionário eletrônico Webster's inglês-português*: português-inglês. Rio de Janeiro: Record, 1982. 1 CD-ROM

JACKSON, P. *Introduction to expert systems*. Wokingham: Addison-Wesley, 1990.

LANDAU, S. I. *Cambridge dictionary of American English*. Cambridge: Cambridge University Press; 2001.

LEVELING, J. Feedback mechanisms for a natural language interface: an application of the critic paradigm. In: RECHERCHE D'INFORMATION ASSISTEE PAR ORDINATEUR – COMPUTER ASSISTED INFORMATION RETRIEVAL, 7., 2004, Avignon. *Proceedings…* Avignon: Le Centre de Hautes Etudes Internationales d'informatique Documentaire, [2004]. p.431-446.

LYONS, J. *Semantics*. Cambridge: Cambridge University Press, 1977. v.2.

MICROSOFT PRESS. *Microsoft press dicionário de informática*. Rio de Janeiro: Campus, 1998.

PALMER, M. Multilingual resources, multilingual information management: current levels and future abilities. *Linguistica Computazionale*, Piza, v.14-15, p.1-33, 2001.

PIANTA, E.; BENTIVOGLI, L.; GIRARDI, C. MultiWordNet: developing an aligned multilingual database. In: INTERNATIONAL CONFERENCE ON GLOBAL WORDNET, 1., 2002, Mysore. *Proceedings...* Mysore: The Central Institute of Indian Languages, [2002]. p.22-25.

RIBEIRO JUNIOR, L. C. *OntoLP*: construção semi-automática de ontologias a partir de textos da Língua Portuguesa. 2008. 131f. Dissertação (Mestrado em Computação Aplicada) – Universidade do Vale do Rio dos Sinos, São Leopoldo, 2008.

ROCA, S. C. Individuación e información parte-todo. Representación para el procesamiento computacional del lenguaje. *Estudios de Lingüística Española*, Madrid, v.8, 2000. Disponível em: <http://elies.rediris.es/elies8/>. Acesso em: 10 jun. 2005.

SAINT-DIZIER, P.; VIEGAS, E. *Computational lexical semantics*. Cambridge: Cambridge University Press, 1995.

SUMMERS, D. (Ed.). *Longman dictionary of contemporary English*. Essex: Longman, 2005. Disponível em: <http://www.ldoceonline.com/>. Acesso em: 28 maio 2008.

VOSSEN, P. Introduction to EuroWordNet. *Computers and the Humanities,* Flushing, v.32, n.2-3, p.73-89, 1998.

WEISZFLOG, W. *Michaelis*: moderno dicionário da Língua Portuguesa. São Paulo: Melhoramentos, 1998. Disponível em: <http://michaelis.uol.com.br/moderno/portugues/index.php>. Acesso em: 20 maio 2008.

WEISZFLOG, W. *Michaelis*: moderno dicionário inglês (inglês-português/português-inglês). São Paulo: Melhoramentos, 2000. Disponível em <http://michaelis.uol.com.br/moderno/ingles/index.php>. Acesso em: 10 abr. 2008.

Recebido em setembro de 2008.

Aprovado em novembro de 2008.