REBECA[1] – uma Base de Dados Léxico-Conceitual Bilíngüe Inglês-Português

Ariani Di Felippo[1,2], Bento Carlos Dias-da-Silva[1,2]

Centro de Estudos Lingüísticos e Computacionais da Linguagem – CELiC[1]
Faculdade de Ciências e Letras – Universidade Estadual Paulista (UNESP)
Caixa Postal 174 – 14.800-901, Araraquara, SP, Brazil
Núcleo Interinstitucional de Lingüística Computacional – NILC[2]
Instituto de Ciências Matemáticas e de Computação – Universidade de SãoPaulo (USP)
Caixa Postal 668 – 13.560-970, São Carlos, SP, Brazil
arianidf@uol.com.br; bento@fclar.unesp.br

Abstract. Lexicons are the heart of many natural language processing systems. Consequently, one of the main issues in last years as regards Natural Language Processing activities is the development of lexical semantic and ontological resources. For Brazilian Portuguese, in particular, there are few resources of this kind available. In this scenario, this paper presents REBECA, a bilingual lexical-conceptual database for BP and (American) English. Accordingly, after contextualizing the project, we describe the two-domain approach methodology that has been applied to the development of REBECA, and its structured artificial language or interlingua. After, we describe the main characteristics and potentialities of REBECA and sketch some future works.

Keywords: natural language processing; lexicon; lexical-conceptual database; structured interlingua; MultiNet.

Doutorado (conclusão: 01 agosto de 2008)

1 Introdução

Atualmente, em função das aplicações reais para as quais os sistemas de Processamento Automático das Línguas Naturais (PLN) são escritos, é premente a compilação de recursos lexicais monolíngües e/ou multilíngües que sejam: (i) manipuláveis pelo sistema do qual fazem parte e (ii) lingüisticamente motivados [1] [2]. A construção de bases lexicais, principalmente para o inglês norte-americano (Ingl), como a WordNet de Princeton (WN.Pr) [3] e a FrameNet [4], e para as línguas européias, como a EuroWordNet [5] e a MultiWordNet [6], confirmam a necessidade de recursos que armazenam informações semântico-conceituais das unidades lexicais.

[1] REBECA, do hebraico *Ribqah*, significa "aquela que une, liga". A escolha do nome Rebeca se deveu, principalmente, a uma característica dessa base ora considerada fundamental. Trata-se da utilização de uma interlíngua conceitual, responsável por "ligar" uma parcela do léxico do inglês norte-americano (Ingl) a uma do português brasileiro (PB).

Nesse cenário, destacam-se os recursos multilíngües em que bases monolíngües de línguas distintas estão alinhadas por meio de uma interlíngua, ou seja, uma coleção única de conceitos. A EuroWordNet e a MultiWordNet são exemplos paradigmáticos desse tipo de recurso.

O alinhamento nessas bases é feito por uma interlíngua não-estruturada[2] e por relações interlinguais rotuladas. Por exemplo, na Figura 1, ilustra-se que o *synset*[3] {finger} da WN.Pr está indexado ao ILI {finger}[4] pela relação de equivalência sinonímica eq_synonym. Devido a uma diferença léxico-conceitual, o conceito expresso pelo ILI {finger} não é lexicalizado no espanhol[5]. Assim, o *synset* {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}.

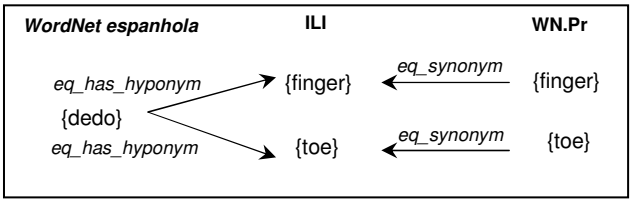

Fig. 1. Indexação léxico-conceitual na EuroWordNet e na MultiWordNet.

O alinhamento das bases da WordNet.Br[6] (WN.Br) e WN.Pr, que está sendo feito nos moldes da EuroWordNet [7], resultará no único recurso desse tipo para o português do Brasil (PB).

Diante desse cenário, apresenta-se aqui a base bilíngüe REBECA, desenvolvida para o par de línguas Ingl-PB. Nessa base, um conjunto de conceitos lexicalizados (isto é, expressos por unidades lexicais[7]) no Ingl está alinhado a um conjunto de conceitos lexicalizados no PB por meio de uma interlíngua estruturada. Mais especificamente, na Seção 2, apresenta-se a metodologia adotada para a construção dessa base e as atividades realizadas em cada etapa prevista pela metodologia. Na Seção 3, apresentam-se as principais características e potencialidades lingüísticas e computacionais da base REBECA. Na Seção 4, apresentam-se

---

[2] A interlíngua (ou *Inter-lingual-Iindex,* ILI) da EuroWordNet e da MultiWordNet é composta por conceitos que não são inter-relacionados. Especificamente, esses conceitos são os *synsets* da WN.Pr 1.5. Um índice da interlíngua ou um ILI é composto por um *synset* e pelo número de registro e glosa desse *synset*.

[3] Construto criado para designar a unidade básica de estruturação da WN.Pr. Um *synset* é um conjunto de unidades lexicais sinônimas ou quase-sinônimas que permite ao falante inferir o conceito (ou descrição mental de um tipo de entidade) evocado por elas.

[4] Por questão de brevidade, um ILI é ilustrado aqui apenas pelo *synset*.

[5] Nesse caso, diz-se que há uma "lacuna lexical" [13].

[6] A WN.Br armazena atualmente apenas os *synsets* e a relação de antonímia. A especificação das demais relações está prevista nas etapas futuras de desenvolvimento.

[7] Entende-se por unidades lexicais as expressões que se espera encontrar como entradas ou subentradas em dicionários monolíngües.

etapas futuras de desenvolvimento da referida base. Por fim, na Seção 5, tecem-se algumas considerações finais sobre este trabalho.

2 Metodologia

Para o desenvolvimento da base REBECA, tomou-se por base [8], que fornece os passos essenciais para o desenvolvimento de projetos na área do PLN. Para [8], os sistemas de PLN são vistos como "sistemas especialistas" ou "sistemas baseados em conhecimento" (do inglês, *knowledge-based systems*) [9]. Segundo essa concepção, as pesquisas nesse domínio envolvem uma "engenharia do conhecimento[8] lingüístico". Ao conceber um sistema de PLN dessa forma, [8] propõe que as pesquisas sigam as seguintes etapas, as quais se baseiam em [11]:
(i) "extração do solo", isto é, explicitação dos conhecimentos e habilidades lingüísticas;
(ii) "lapidação", isto é, representação formal desses conhecimentos e habilidades;
(iii) "incrustação", isto é, o programa de computador que codifica essa representação.

Tais etapas foram denominadas, respectivamente, domínio lingüístico, domínio lingüístico-computacional e domínio computacional [8]. Neste trabalho, as atividades de pesquisa ficaram restritas apenas aos domínios lingüístico e lingüístico-computacional, posto que as atividades do domínio computacional não fazem parte do escopo deste trabalho.

A seguir, apresentam-se as atividades realizadas nos domínios lingüístico e lingüístico-computacional tendo em vista a construção da base bilíngüe REBECA.

2.1 Domínio lingüístico

As atividades relativas ao domínio lingüístico ficaram especialmente concentradas em:

(i) A delimitação do tipo conceitual
Nessa etapa, delimitou-se o tipo de conceito que seria armazenado na base (p.ex.: aqueles expressos por nomes, verbos, adjetivos, etc.). Decidiu-se por armazenar apenas conceitos do tipo "objeto concretos discretos". Segundo [12], tais conceitos são entidades de 1ª ordem e, por isso, intuitivamente categorizam referentes perceptíveis pelos sentidos, localizadas no tempo e no espaço, que são contáveis e indivisíveis. Quanto à expressão lingüística, tais conceitos realizam-se por expressões nominais, sejam elas simples, compostas ou complexas. A escolha dessa classe de conceitos justifica-se pelo fato de que eles, devido a sua natureza hierárquica, são passíveis de uma sistematização formal.

(ii) A delimitação do domínio conceitual
Partindo-se do princípio de que os conceitos não estão isolados na mente, mas sim organizados [13], delimitou-se o domínio conceitual dos "veículos com rodas". A escolha desse domínio não se justifica por questões teóricas, mas sim práticas; no caso: delimitação bem-definida e extensão reduzida.

---

8 Vale ressaltar que "conhecimento" (do inglês, *knowledge*) é um "termo guarda-chuva", empregado para denotar qualquer tipo de informação manipulada por um sistema computacional [10].

(iii) A 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 conceitos, codificados em *synsets*, mais específicos que o conceito subjacente ao *synset* {wheeled vehicle}, ou seja, todos os hipônimos de {wheeled vehicle}. A escolha da WN.Pr como fonte dos conceitos teve duas 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 monolíngües do Ingl [14] [15].

(iv) A identificação dos conceitos lexicalizados e a montagem da base monolíngüe do Ingl
Com base nos dicionários monolíngües do Ingl [14] [15], 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 lingüísticas que compõem os seus respectivos *synsets* não são entradas ou subentradas em tais dicionários. Assim, a base monolíngüe 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[9].

(v) A investigação dos conceitos lexicalizados e a montagem da base monolíngüe do PB
Partindo-se dos conceitos da interlíngua, foi possível identificar em uma primeira fase, por meio de consultas manuais a dicionários bilíngües Ingl-PB [16] [17], os conceitos que eram expressos por unidades lexicais no PB. Em uma segunda fase, dicionários monolíngües [18] [19] e de sinônimos [20] [21] foram manualmente consultados para a identificação de unidades sinônimas às compiladas nos dicionários bilíngües e subseqüente montagem dos *synsets* do PB. 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[10] e textos disponíveis na *Web*. Os textos em PB disponíveis na *Web* foram consultados através do motor de busca Google[11], 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. Além das unidades lexicais, foram identificados os chamados "sintagmas livres recorrentes" (SLRs) (do inglês, *recurrent free phrases*), ou seja, expressões que não são dicionarizadas, mas que comumente expressam determinado conceito. 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 importantes para o tratamento computacional das "lacunas lexicais", pois provêem expressões correspondentes para

9 http://www.webcorp.org.uk/index.html
10 O PLN-BR FULL contém cerca de 29 milhões de palavras e está disponível para consultas através do Philologic, ferramenta Web para análise de *corpora* desenvolvida na Universidade de Chicago.
11 http://www.google.com.br/

conceitos que não são lexicalizados. Os SLRs formam um conjunto próprio: um *phraset*. Para cada SLR, uma frase-exemplo também fora compilada dos referidos *corpora*. Dos 205 conceitos lexicalizados no Ingl, foram identificadas 84 lexicalizações no PB, sendo que, para 12 delas, foi possível identificar também um SLR. Das 121 lacunas, em apenas 40 casos foi possível identificar um SLR. Vale ressaltar que, para os 12 conceitos da interlíngua que não são lexicalizados no Ingl, a ausência de lexicalizações no PB não foi considerada lacuna lexical. Ao final, tem-se a base monolíngüe do PB composta por 84 conceitos organizados em *synsets*.

2.2 Domínio lingüístico-computacional

As atividades relativas ao domínio lingüístico-computacional ficaram especialmente concentradas em:

(i) A representação formal da interlíngua

Nessa etapa, os conceitos da interlíngua foram representados em função do modelo de representação do conhecimento (RC) denominado MultiNet [22] (do inglês, *Multilayered Extended Semantic Networks*).

Ao conceber o PLN como uma espécie de "engenharia do conhecimento lingüístico", as atividades do domínio lingüístico-computacional podem ser beneficiadas pelas estratégias da Engenharia do Conhecimento. Seguindo essa concepção, adotou-se o MultiNet [22], um modelo de RC que se baseia na metalinguagem formal das redes semânticas e cujos construtos básicos estão ilustrados na Figura 2.

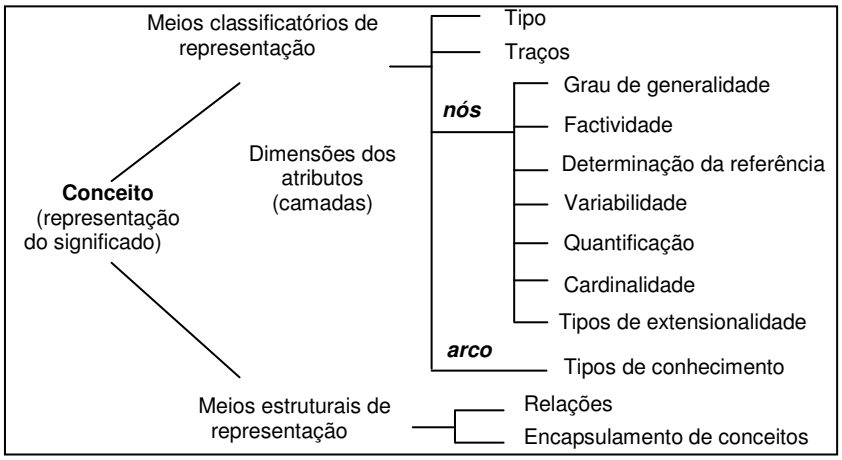

Fig 2. Os construtos de representação do MultiNet.

O MultiNet tem sido empregado principalmente como interlíngua semântica para recuperação de informação na *Web* por meio de interfaces em língua natural [23]. 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.

Devido à adoção do Multinet, a interlíngua é, do ponto de vista de sua macroestrutura, 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 conceitual escolhido (os "objetos concretos discretos"), a relação de subsunção, representada pelo rótulo SUB, é a mais importante para organizar tais conceitos e, por isso, a responsável pela organização global da interlíngua. Além de SUB, os conceitos da interlíngua ligam-se a outros conceitos por meio das relações PARS (parte-todo) e PURP (propósito), também consideradas fundamentais para a caracterização do tipo de conceito sob análise. Tais conceitos, como <axletree>[12] da Figura 3, não fazem parte da interlíngua e sim especificam os conceitos da mesma. As relações SUB, PARS e PURP estabelecidas por cada conceito da interlíngua também foram identificadas com base na WN.Pr 2.1. O encapsulamento de conceitos, por sua vez, prevê 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} está associado a {air bag} por meio 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.

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 "veículos com rodas" são do tipo [mov-art-discrete]. Assim, todo conceito da interlíngua está associado ao tipo conceitual cujo valor é [mov-art-discrete]. Além dos tipos, o MultiNet prevê que os conceitos sejam associados a traços semânticos (do inglês, *features*), que facilitam a formulação das 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+]. Conseqüentemente, 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 [22]. Os atributos dos nós são: (a) grau de generalidade (GENER); (b) determinação da referência (REFER); (c) variação (da referência) (VARIA); (d) factividade (FACT) e (e) 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*], [FACT=*real*] e [ETYPE=0]. 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. 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

[12] Os sinais de < > são empregados para expressar "conceitos".

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 inglês, *categorial knowledge*), 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> (no PB, *carroça*), elemento constitutivo da interlíngua, é representado pelo MultiNet. Vale ressaltar que os arcos pontilhados e os nós sem preenchimento indicam os conceitos que não fazem parte da interlíngua.

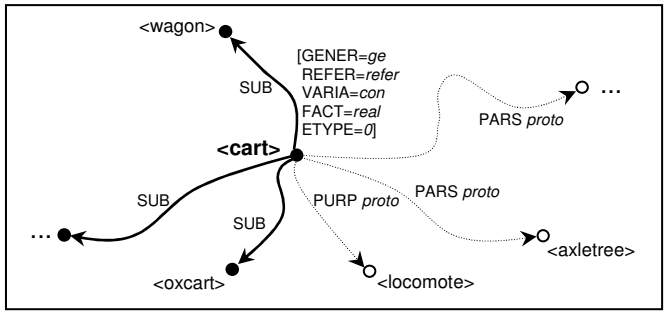

Fig. 3. Ilustração da representação formal segundo o MultiNet.

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" [24] [25].

(ii) A construção da base REBECA: sua implementação no editor Protégé-OWL

Para a construção da base REBECA, utilizou-se um dos editores de ontologia mais difundidos na literatura, o Protégé (3.3.1)[13]. Especificamente, utilizou-se a versão desenvolvida com base na linguagem OWL[14]. 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 lingüístico e representadas no domínio lingüístico-computacional pudessem ser adequadamente inseridas. Tais adaptações foram: (i) os conceitos da interlíngua 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 como `ObjectProperty` (isto é, construto para representar propriedades intrínsecas às classes) e os atributos multidimensionais como `DatatypeProperty` (isto é, construto para representar demais informações sobre as classes); (iii) os *synsets* que compõem a base monolíngüe do Ingl e os *synsets* e *phrasets* que

[13] http://protege.stanford.edu/
[14] http://www.w3.org

compõem a base do PB foram inseridos como "instâncias" ou "indivíduos" das classes; (iv) as glosas foram inseridas como "comentários" das classes (conceitos); e (v) as frases-exemplo foram inseridas como "comentários" das instâncias (unidades lexicais ou SLRs).

3 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 monolíngüe 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 lingüísticas alternativas: os SLRs e a(s) unidades lexicais (ou SLRs) que expressam um conceito hiperônimo.

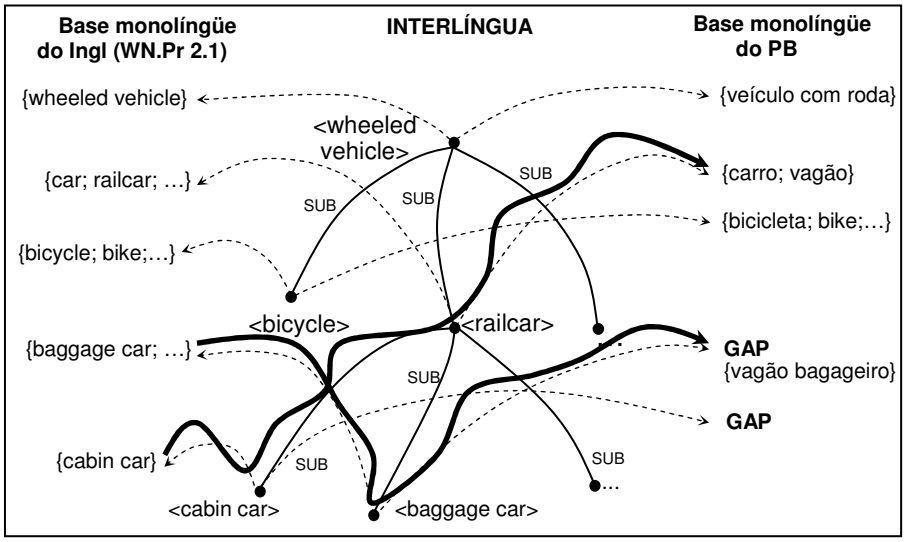

Fig. 4. Os alinhamentos léxico-conceituais na base de dados REBECA.

Na Figura 4, por exemplo, observa-se que os conceitos <cabin car> e <baggage car> não são lexicalizados no PB, configurando lacunas lexicais nessa língua ("GAP"). Nessa Figura, as

setas mais espessas indicam os caminhos para a identificação das expressões lingüísticas 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 monolíngües do Ingl e do PB, respectivamente, que estão indexadas ao mesmo conceito da interlíngua. 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, no nível superior, que o conceito <railcar> é lexicalizado no PB, expresso especificamente por *carro* e *vagão*.

Dessa forma, sob o ponto de vista lingüístico, 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 (cf. Figura 4). Conseqüentemente, sob o ponto de vista tecnológico, evidencia-se seu potencial de uso em várias aplicações do PLN, por exemplo, na recuperação de informação multilíngüe, pela expansão de unidades lexicais de uma língua a unidades lexicais relacionadas em outra língua via a interlíngua estruturada.

4 Trabalhos futuros

Como desenvolvimento futuro deste trabalho, propõe-se: (i) o refinamento do domínio conceitual dos "veículos 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 de *corpora* e poderá ser feita com o auxílio do *plug-in* do Protégé-OWL denominado OntoLT [26]. Tal tarefa poderá contar também com recursos computacionais e lexicográficos do Ingl. A tarefa (ii) é semelhante à (i) e consistirá na extração de *corpora* de conceitos (e unidades lexicais) que são lexicalizados especialmente no PB; tal extração poderá ser feita com o auxílio do *plug-in* OntoLP [27], que é a adaptação do OntoLT para o tratamento de textos em PB. 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 dos "veículos 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.

5 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. É reconhecido que a identificação de tais padrões e o subseqüente 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 língua Ingl-PB em aplicações como tradução automática e/ou recuperação de informação multilíngüe. Além disso, a pesquisa que resultou na base REBECA busca promover a visão lingüisticamente

motivada das atividades do PLN e, conseqüentemente, fortificar o trabalho colaborativo entre cientistas (lingüistas) e engenheiros da linguagem (cientistas da computação).

Referências

1. Palmer, M., Multilingual resources, multilingual information management: current levels and future abilities. Linguistica Computazionale, 14-15, 1--33 (2001)
2. Hanks, P. Lexicography. In: Mitkov, R. (ed.). The Oxford handbook of computational linguistics. Oxford University Press: Oxford. 48--69 (2004)
3. Fellbaum, C., WordNet: an electronic lexical database. MIT Press: Cambridge, MA. (1998)
4. Baker, C.F., Fillmore, C.J., Lowe, J.B. (998). The Berkeley FrameNet project. In: 17th International Conference on Computational Linguistics (COLING/ACL), pp. 86—90. Montreal (1998)
5. Vossen, P., Introduction to EuroWordNet. Computers and the Humanities, 32, 73-89 (1998)
6. Pianta, E., Bentivogli, L., Girardi, C. MultiWordNet: developing an aligned multilingual database. In: 1st International Conference on Global WordNet, pp. 22--25, Mysore, Índia (2002)
7. Di Felippo, A., Dias-da-Silva, B. C., Towards an automatic strategy for acquiring the WordNet.Br hierarchical relations. In: 5th Workshop in Information and Human Langague Technology, RJ (2007)
8. Dias-da-Silva, B.C., O estudo lingüístico-computacional da linguagem. Letras de Hoje, 41, 103--38 (2006)
9. Grishman, R., Computational linguistics. Cambridge University Press: Cambridge (1986)
10. Nirenburg, S. et al., Machine translation. Morgan Kaufmann: San Mateo (1992)
11. Hayes-Roth, F., Expert systems. In: Shapiro, E. (ed.) Encyclopedia of artificial intelligence. New York, Wiley, 287--298 (1990)
12. Lyons, J., Semantics. Cambridge University Press: Cambridge, 2, (1977)
13. Cruse, A., Meaning in language: an introduction to semantics and pragmatics. Oxford University Press: Oxford (2004)
14. Landau, S. I., Cambridge dictionary of American English. Cambridge University Press: Ca. (2001)
15. Summers, D. (ed.). Longman dictionary of contemporary English online. Longman Group Ltda (2005). http//www.ldoceonline.com
16. Houaiss, A., Cardim, I. (orgs.) Dicionário eletrônico Webster's inglês-português/ português-inglês. Ed. Record: Rio de Janeiro (1982) 1 CD-ROM
17. Weiszflog, W., Michaelis: moderno dicionário inglês (inglês-português/ português-inglês). Editora Melhoramentos, 2000. http://michaelis.uol.com.br/moderno/ingles/index.php
18. Ferreira, A.B.H. Novo dicionário eletrônico Aurélio da língua portuguesa. Positivo: Curitiba (2004)
19. Houaiss, A., Villar, M. de S., Dicionário eletrônico Houaiss da língua portuguesa. (versão 1.0). Editora Objetiva: Rio de Janeiro (2001) 1 CD-ROM.
20. Barbosa, O., Grande dicionário de sinônimos e antônimos. Ediouro: Rio de Janeiro (2000)
21. Fernandes, F., Dicionário de sinônimos e antônimos da língua portuguesa. Globo: São Paulo (1997)
22. Helbig,H., Knowledge representation and semantics for natural language. Springer-Verlag (2006)
23. 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 (RIAO), pp. 431--446. Avignon, France (2004)
24. Gruber, T., Toward principles for the design of ontologies used for knowledge sharing. International Journal Human-Computer Studies, 43 (5-6), 907-928 (1995)
25. Borst, W.N., Construction of engineering ontologies. Holanda, 1997. Tese (Doutorado). http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf
26. Buitellar, A. et al., A Protégé plug-in for ontology extraction from text based on linguistic analysis. In: 1st European Semantic Web Symposium (ESWS), pp. 31--44.1. Heraklion, Greece (2004)
27. Ribeiro Jr., Vieira, R., Geração de ontologias para a web semântica a partir de textos da língua portuguesa. In: 3rd Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA), Ribeirão Preto (2006)