A interlíngua e o alinhamento léxico-conceitual da base de dados bilíngue REBECA

(The interlingua and the lexical-conceptual alignment in REBECA lexical database)

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

[1]Departamento de Letras - Universidade Federal de São Carlos (UFSCar)
[2]Faculdade de Ciências e Letras - Universidade Estadual Paulista (UNESP/Ar.)

ariani@ufscar.br, bento@fclar.unesp.br

Abstract: In the architecture of a natural language processing system based on linguistic knowledge, two types of component are important: the knowledge databases and the processing modules. One of the knowledge databases is the lexical database, which is responsible for providing the lexical unities and its properties to the processing modules. The systems that process two or more languages require bilingual and/or multilingual lexical databases. These databases can be constructed by aligning distinct monolingual databases. In this paper, we present the interlingua and the strategy of aligning the two monolingual databases in REBECA, which only stores concepts from the "wheeled vehicle" domain.

Keywords: Natural Language Processing; REBECA; lexical database; interlíngua; lexical-conceptual alignment.

Resumo: Quando baseados em conhecimento (linguístico), os sistemas que processam língua natural apresentam dois grupos de componentes: as bases de conhecimento e os módulos de processamento. Uma dessas bases de conhecimento é lexical, responsável por fornecer, aos módulos de processamento, as unidades da língua em questão e as suas respectivas propriedades. Os sistemas que processam duas ou mais línguas requerem bases lexicais bilíngues e/ou multilíngues. Tais bases podem ser construídas em função do alinhamento de diferentes bases monolíngues. Neste trabalho, apresentamos a interlíngua e a estratégia de alinhamento utilizadas na construção da base bilíngue REBECA, que engloba conceitos lexicalizados do domínio dos "veículos com rodas".

Palavras-chave: Processamento Automático das Línguas Naturais; REBECA; bases de conhecimento lexical; interlíngua; alinhamento léxico-conceitual.

Introdução

A arquitetura de um sistema computacional que processa língua natural[1] (p.ex.: sistema de tradução automática) varia de acordo com as especificidades da aplicação para a qual são feitos. No entanto, quando baseados em conhecimento linguístico, dois grupos de componentes são comuns nesses sistemas: as bases de conhecimento e os módulos de processamento que atuam sobre as bases (ALLEN, 1994). Um desses módulos de processamento é o de "análise ou interpretação", responsável pela construção de uma representação do significado das sentenças de um texto de entrada. Para a construção dessa representação, o sistema requer conhecimento semântico das unidades lexicais da língua em questão (SAINT-DIZIER; VIEGAS, 1995).

As unidades lexicais e as suas propriedades morfológicas, sintáticas e semântico-conceituais são fornecidas ao módulo de análise pelo "léxico" ou "base lexical" (HANKS, 2004). A base de conhecimento lexical, juntamente com as bases gramatical e conceitual, forma o conjunto das bases de conhecimento linguístico

[1] No âmbito do Processamento Automático das Línguas Naturais (PLN), o termo "língua natural" engloba as modalidades escrita e oral, desde que ambas estejam registradas em meio escrito.

imprescindível aos sistemas que processam língua natural. Tais léxicos também são denominados "dicionários tratáveis por máquina" (do inglês, *machine tractable dictionaries*) (WILKS et al., 1996).

Para o desenvolvimento de sistemas como os de "tradução automática" e "recuperação de informação multilíngue", que processam duas ou mais línguas, os pesquisadores do Processamento Automático das Línguas Naturais (PLN) necessitam de bases lexicais bilíngues e/ou multilíngues (PALMER, 2001).

Nesse cenário, destacamos a base multilíngue EuroWordNet (VOSSEN, 1998), que agrupa bases no formato *wordnet*[2] desenvolvidas para várias línguas europeias. Nessa base, o alinhamento é feito por uma interlíngua (isto é, conjunto único de conceitos) não-estruturada, denominada *Inter-lingual-Index* (ILI), e por relações interlinguais rotuladas. A interlíngua é composta pelo conjunto dos *synsets* da WordNet de Princeton[3] (versão 1.5) (WN.Pr) (FELLBAUM, 1998) e suas respectivas glosas, ou seja, definições informais dos conceitos subjacentes as *synsets.* Na Figura 1, ilustramos que o *synset* {finger}[4] da WN.Pr está indexado ao ILI {finger}[5] 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, há uma lacuna lexical (do inglês, *lexical gap*) no espanhol. Assim, o *synset*[6] {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 bases monolíngues 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}.

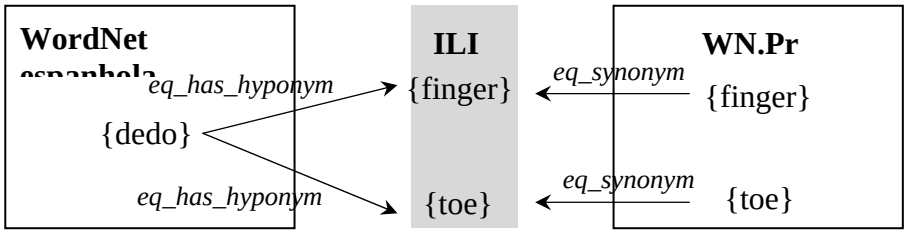

Figura 1. Indexação léxico-conceitual na EuroWordNet

Para o português do Brasil (PB), os recursos lexicais computacionais ainda são bastante escassos. Nesse cenário, destacamos a WordNet.Br, que, em seu estágio atual de desenvolvimento, está sendo alinhada à WN.Pr nos moldes da EuroWordNet (DIAS-DA-SILVA et al., 2008). Quando alinhadas, as *wordnets* brasileira e norte-americana

[2] Em uma base no formato *wordnet*, estão armazenados apenas conceitos lexicalizados, ou seja, expressos por unidades lexicais. De uma forma geral, esses conceitos são codificados por conjuntos de formas sinônimas, os chamados *synsets* (do inglês, *synonym sets*), os quais se relacionam a outros conjuntos por meio de várias relações (antonímia, hiponímia, meronímia, acarretamento e causa) (FELLBAUM, 1998).
[3] A WordNet de Princeton armazena unidades lexicais do inglês norte-americano (Ingl).
[4] Os conceitos, quando codificados em *synsets*, são representados entre parênteses; caso contrário, entre os símbolos <>.
[5] Por questão de simplificação, o ILI está representado apenas pelo *synset.*
[6] 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.

constituirão uma importante base bilíngue para o processamento automático do par de línguas PB-Ingl.

Neste trabalho, em especial, apresentamos a interlíngua e a estratégia de alinhamento léxico-conceitual utilizadas na construção de outra base bilíngue (PB-Ingl), a REBECA (DI FELIPPO; DIAS-DA-SILVA, 2008). A interlíngua da REBECA caracteriza-se por ser estruturada e formal e o alinhamento, por sua vez, por não utilizar relações rotuladas.

Para tanto, dividimos este artigo em 5 Seções. Na Seção 2, (i) apresentamos a composição, ou seja, o conjunto de conceitos que formam a interlíngua da base REBECA e (ii) descrevemos a macro e a microestrutura da interlíngua em função do formalismo utilizado para representar os seus conceitos constitutivos. Na Seção 3, descrevemos brevemente as bases monolíngues que compõem a REBECA. Na seção 4, descrevemos a implementação e a estratégia de alinhamento das bases monolíngues. Por fim, na seção 5, algumas considerações finais são apresentadas.

A composição e a representação da interlíngua da base REBECA

A composição da interlíngua

O conjunto dos conceitos constitutivos da interlíngua foi manualmente extraído da WN.Pr (versão 2.1) (FELLBAUM, 1998), que é um recurso linguisticamente confiável e livremente disponível. Precisamente, foram selecionados todos os conceitos mais específicos que o conceito <wheeled vehicle>,[7] ou seja, todos os *synsets* relacionados ao *synset* {wheeled vehicle} por meio da relação de hiponímia, como ilustrado na Figura 2.

wheeled vehicle
=> baby buggy, baby carriage, carriage, perambulator, pram, stroller, go-cart, pushchair, pusher
=> bassinet
=> bicycle, bike, wheel, cycle
=> bicycle-built-for-two, tandem bicycle, tandem
=> mountain bike, all-terrain bike, off-roader
=> ordinary, ordinary bicycle
….

Figura 2. O *synset* {wheeled vehicle} e seus hipônimos na WN.Pr 2.1

A escolha do domínio "veículos com rodas" pautou-se em dois aspectos: a delimitação bem-definida e a extensão reduzida do domínio. O domínio dos "veículos com rodas" engloba um tipo específico de conceito, os "objetos (conceituais) concretos discretos". Tais conceitos intuitivamente categorizam referentes perceptíveis pelos sentidos, localizados no tempo e no espaço, que são contáveis e indivisíveis (LYONS, 1977). A escolha desse tipo de conceito pautou-se no fato de que eles, devido a sua natureza hierárquica, são passíveis de uma sistematização formal. No total, foram obtidos 217 conceitos (ou *synsets*). Na REBECA, cada um desses conceitos foi associado a uma glosa em PB. Para elaboração das glosas, partimos das próprias glosas

[7] As unidades lexicais do inglês nada mais são do que um recurso mnemônico para a descrição dos conceitos.

(em inglês) armazenadas na WN.Pr. As glosas da WN.Pr consideradas bem-formadas foram diretamente traduzidas para o PB. Caso contrário, novas glosas foram elaboradas com base nas definições de dicionários monolíngues do Ingl. (LANDAU, 2000; SUMMERS, 2005).

Vale ressaltar que, ao contrário da EuroWordNet, os conceitos da interlíngua da base REBECA não estão codificados em *synsets*. Para a implementação da base, utilizamos uma única unidade lexical do inglês para identificar um conceito. No caso, utilizamos a primeira unidade que compõe um *synset*. Por exemplo, o conceito <bicycle>, codificado na EuroWordNet pelo *synset* {bicycle, bike, wheel, cycle}, está armazenado na REBECA por meio do rótulo <bicycle>.

A representação da interlíngua

Para a representação dos conceitos da interlíngua, adotamos o modelo de representação do conhecimento denominado MultiNet (do inglês, *Multilayered Extended Semantic Networks*) (HELBIG, 2006), que tem sido usado como interlíngua semântica de interfaces de busca em língua natural. O MultiNet é um modelo de representação do conhecimento que se baseia na metalinguagem formal das redes semânticas. 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; (ii) adequação cognitiva, isto é, todo conceito tem uma representação única por meio da qual toda a informação a ele associada torna-se acessível. Seguindo os pressupostos do MultiNet, todo conceito da interlíngua foi representado em função dos construtos da Figura 3, os quais são responsáveis pela macro e microestruturação dos mesmos.

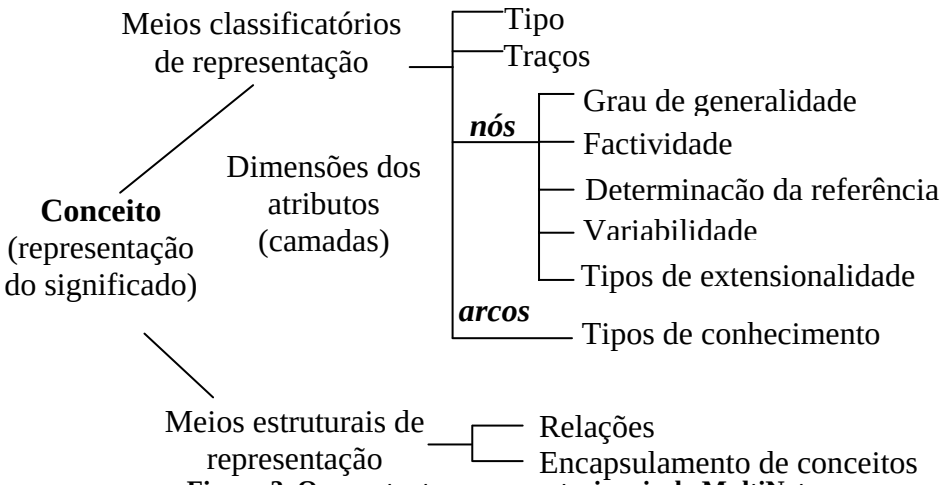

Figura 3. Os construtos representacionais do MultiNet

A macroestrutura

Tendo em vista a adoção do MultiNet, a interlíngua da base REBECA é, do ponto de vista de sua macroestrutura, uma rede semântica, composta por nós (conceitos) e arcos (relações).

Especificamente, os *meios estruturais* de representação 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 de

hiperonímia/ hiponímia é a mais importante para organizar tais conceitos. Assim, do ponto de vista da macroestrutura, a interlíngua está organizada exclusivamente em função dessa relação que, no MultiNet, é descrita pelo rótulo SUB (subsunção). Além de SUB, os conceitos da interlíngua estão especificados pelas relações PARS (parte-todo ou meronímia) e PURP (propósito), também consideradas fundamentais para a caracterização do tipo de conceito sob análise.[8] 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; são uma espécie de propriedade dos nós.

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 <car> estiver associado a <air bag> por meio da relação PARS, os conceitos hipônimos de <car> herdam essa relação. Isso acontece porque a relação PARS é tida como conhecimento prototípico, o qual é herdado por *default* pelos conceitos mais específicos.

A microestrutura

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 roda" 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 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 na REBECA são: (a) grau de generalidade (GENER); (b) factividade (FACT); (c) determinação da referência (REFER); (d) variabilidade (VARIA); e (e) extensionalidade (ETYPE) (cf. Figura 3).

Os atributos GENER, REFER, VARIA, FACT e ETYPE, segundo o modelo, têm vários valores. Como os conceitos que pertencem à interlíngua são tidos como genéricos (p.ex.: <car>), 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 a

[8] O MultiNet possui um elenco de mais de 100 relações. No caso, as relações SUB, PARS e PURP foram consideradas as mais relevantes para a descrição formal dos conceitos do tipo em questão.

descrição no nível pré-extensional de um conceito genérico x é um elemento prototípico do conjunto <todos os X>.

O atributo do arco, em especial, é denominado tipo de conhecimento (K-TYPE). O arco relativo à relação SUB é rotulado por K (do inglês, *categorial knowledge*), indicando que o conhecimento é categorial ou imanente 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. A parte categorial do conceito genérico é necessariamente herdada por todos os conceitos subordinados. Já sua parte prototípica é herdada como conhecimento típico. Assim, uma informação prototípica é herdada até que não haja informação mais específica disponível. Se houver, essa informação particularizante prototípica é sobrescrita.

Na Figura 4, o conceito <cart> (no PB, *carroça*), elemento constitutivo da interlíngua, é representado pelo MultiNet.

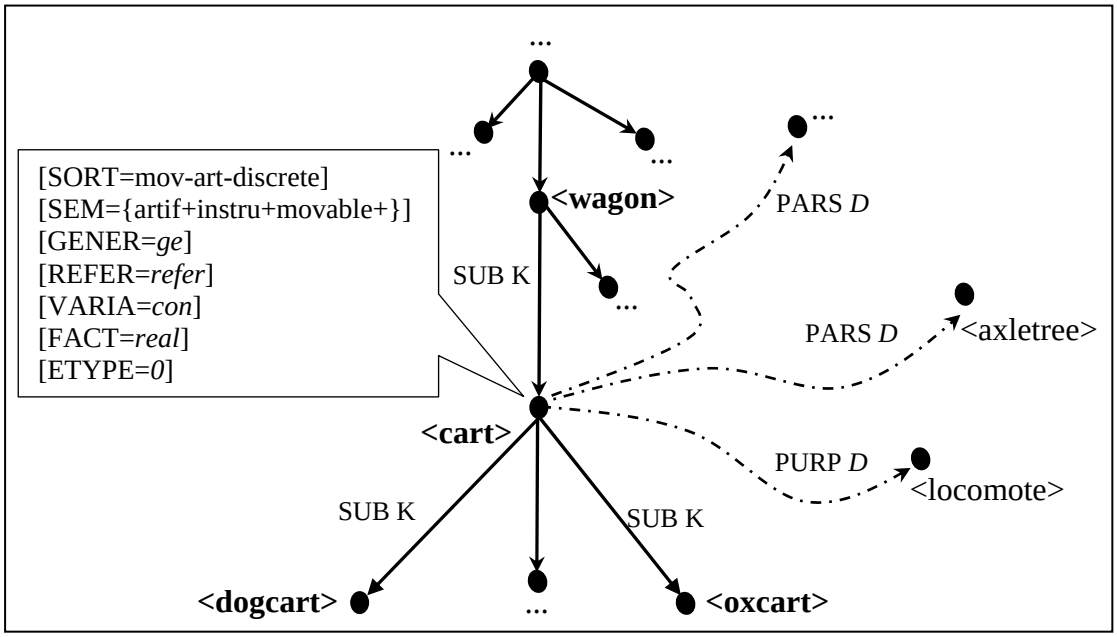

Figura 4. Representação de um conceito segundo o MultiNet

As bases lexicais monolíngues da REBECA

A base monolíngue do Ingl

A base lexical monolíngue do Ingl é composta por 205 conceitos lexicalizados do domínio dos "veículos com rodas". Esses conceitos, na verdade, são os próprios *synsets* da WN.Pr que serviram de base para a constituição da interlíngua. A diferença entre o número de conceitos da interlíngua (217) e o número de conceitos da base monolíngue (205) se deve ao fato de que 12 dos conceitos extraídos da WN.Pr (p.ex.: {self-propelled vehicle}) não são efetivamente lexicalizados no Ingl, ou seja, a(s) expressão(s) linguística(s) que compõe os seus respectivos *synsets* não são entradas ou subentradas nos dicionários monolíngues do Ingl utilizados na construção da REBECA

(LANDAU, 2000; SUMMERS, 2005). Esses conceitos foram mantidos na interlíngua para a estruturação mais adequada da mesma.

Além dos conceitos lexicalizados, codificados em *synsets*, a base monolíngue do Ingl também armazena uma frase-exemplo (isto é, sentença que fornece o contexto de uso mínimo) para cada unidade lexical constitutiva de um *synset*. Tais frases foram manualmente extraídas ou da própria WN.Pr ou da *Web*. Para a extração da *Web*, utilizamos o portal *WebCorp*.[9]

A base monolíngue do PB

A construção da REBECA englobou a construção prévia da base monolíngue do PB. Partindo-se dos conceitos da interlíngua, foi possível identificar em uma primeira fase, por meio de consultas manuais a dicionários bilíngues Ingl-PB (HOUAISS; CARDIM, 1982; WEISZFLOG, 2000), os conceitos que eram expressos por unidades lexicais no PB. Em uma segunda fase, dicionários monolíngues (FERREIRA, 2004; HOUAISS; VILLAR, 2001) e de sinônimos (BARBOSA, 2000; FERNANDES, 2001) foram manualmente consultados para a identificação de unidades sinônimas às compiladas nos dicionários bilíngues e subsequente montagem dos *synsets* do PB.[10] 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[11] e textos disponíveis na *Web*. Os textos em PB disponíveis na *Web* foram consultados através do motor de busca Google.[12] Dos mesmos *corpora*, foram extraídas as frases-exemplo para cada unidade lexical.

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 proveem expressões correspondentes para 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.

A implementação e a estratégia de alinhamento léxico-conceitual

Como a interlíngua da REBECA caracteriza-se como uma espécie de "ontologia linguística", utilizamos, para a construção da base, um dos editores de ontologia mais

[9] http://www.webcorp.org.uk/index.html
[10] Nas bases monolíngues da REBECA, os conceitos estão codificados em *synsets.*
[11] 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.
[12] http://www.google.com.br/

difundidos na literatura, o Protégé (3.3.1).[13] Especificamente, utilizamos a versão desenvolvida com base na linguagem OWL,[14] o Protégé-OWL.

A escolha desse editor baseou-se nas seguintes características: (i) interoperabilidade, que busca consentir a compatibilidade com outros sistemas de representação do conhecimento; o Protégé-OWL gera a base de dados no formato OWL, que permite sua manipulação computacional e integração a sistemas de PLN; (ii) usabilidade, que busca garantir a facilidade de uso da ferramenta; o Protégé apresenta uma interface gráfica que facilita a inserção e manipulação dos dados; e (iii) aplicabilidade, que busca garantir o emprego diversificado das bases por meio da exportação das mesmas em diversos formatos; o editor Protégé-OWL, além de manipular a linguagem OWL, também suporta outros formatos, como HTML (do inglês, *HyperText Markup Language*) e RDF/XML (do inglês, *Resource Description Framework e Extensible Markup Language*).

Os dados da REBECA foram inseridos da seguinte forma no editor:

(a) os conceitos da interlíngua foram inseridos como "classes" do Protégé-OWL;

(b) os demais conceitos, que se vinculam aos da interlíngua pelas relações de PARS e PURP, o tipo conceitual, os traços semânticos 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 o tipo, os traços e os atributos multidimensionais como DatatypeProperty (isto é, construto para representar demais informações sobre as classes);

(c) os *synsets* que compõem a base monolíngue do Ingl e os *synsets* e *phrasets* que compõem a base do PB foram inseridos como "instâncias" ou "indivíduos" das classes;

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

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

A Figura 5 ilustra essa implementação.

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

[14] A OWL é uma linguagem desenvolvida pelo *World Wide Web Consortium* (W3C) (http://www.w3.org/) para promover a Web Semântica, uma proposta de estruturação dos documentos da *Web*.

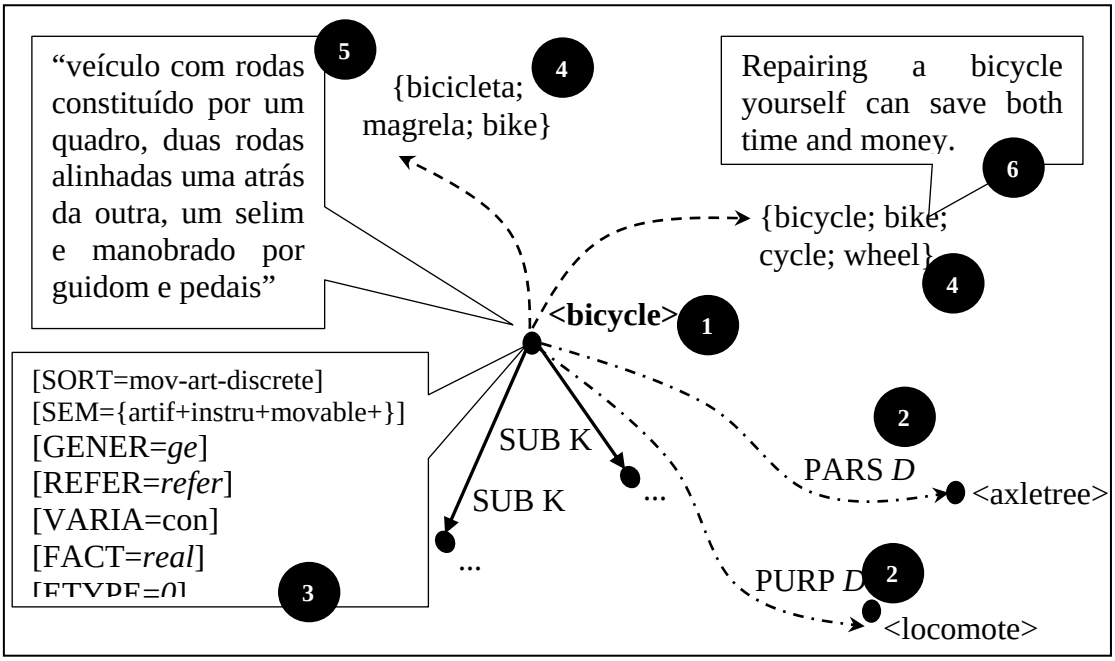

Figura 5. Ilustração da implementação da REBECA no Protégé-OWL

Por terem sido implementadas como instâncias das classes, as quais estão organizadas hierarquicamente, as unidades constitutivas dos *synsets* e dos *phrasets* ligam-se ou alinham-se a apenas um índice da interlíngua, como ilustrado na Figura 6. Nessa figura, por exemplo, os *synset* {bicicleta; magrela; bike} do PB e {bicycle; bike; cycle; wheel} do Ingl lexicalizam o mesmo conceito, representado pelo índice da interlíngua <bicycle>. Em outras palavras, tais *synsets* são "instâncias linguísticas" do conceito <bicycle>.

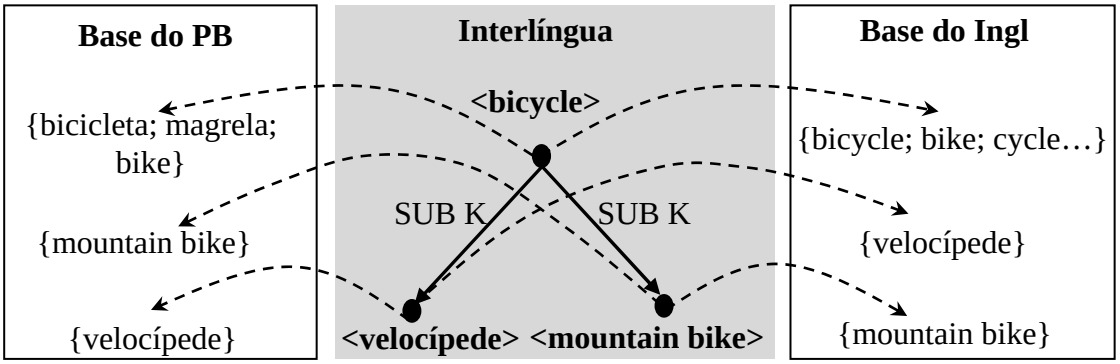

Figura 6. O alinhamento na base REBECA

Nos casos em que há lacunas lexicais por causa de divergências léxico-conceituais, não é preciso criar outros *links* entre as bases monolíngues e a interlíngua

para se encontrar uma possível tradução. É possível percorrer a interlíngua e encontrar em um nível superior uma ou mais unidades lexicais que lexicalizam conceitos menos específicos, como ilustrado na Figura 7. Nessa figura, vemos que o conceito <bicycle-built-for-two>, expresso em Ingl por {bicycle-built-for-two; tandem bicycle; tandem}, não é lexicalizado no PB, havendo, portanto, uma lacuna lexical, indicada pelo rótulo {GAP}. Nesse caso, a partir de qualquer unidade do Ingl pertencente ao *synset* {bicycle-built-for-two; tandem bicycle; tandem}, é possível percorrer a interlíngua e encontrar, no nível superior, as unidades {bicicleta; magrela; bike}, que, apesar de lexicalizações de um conceito mais genérico, podem servir de possíveis traduções no PB.

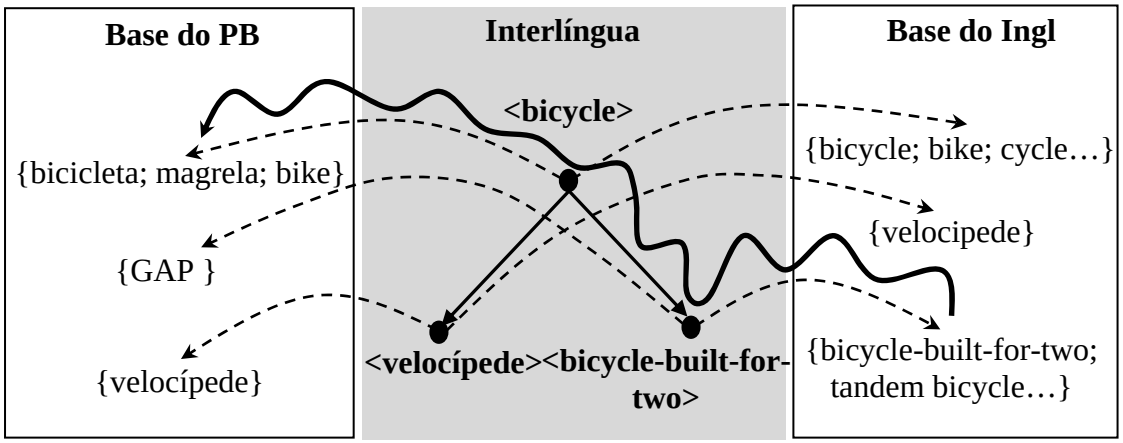

Figura 7. O alinhamento e as lacunas lexicais na base REBECA

Considerações finais

Neste trabalho, apresentamos a interlíngua e a estratégia de alinhamento adotadas na construção da base de dados bilíngue REBECA. Nessa base, a interlíngua é representada segundo os construtos do MultiNet, isto é, um modelo de representação do conhecimento que tem como base as redes semânticas. A adoção do MultiNet é responsável pelas três principais características da REBECA. A primeira é a organização hierárquica (relação SUB) da interlíngua, que a faz altamente estruturada. A segunda é o tipo de alinhamento. Devido à organização hierárquica da interlíngua, os *synsets* das bases monolíngues (e os *phrasets*, no caso da base do PB) alinham-se a apenas um índice da interlíngua. Dessa forma, evita-se o número excessivo de *links*, como acontece na EuroWordNet. A terceira característica é o grau de formalização dos conceitos da interlíngua. O MultiNet fornece meios representacionais que permitem uma descrição bastante explícita dos conceitos subjacentes às unidades lexicais. Para o PLN, tal descrição é essencial, posto que, quanto mais explícito for o conhecimento, mais uma base lexical se torna manipulável pelo sistema do qual faz parte. Assim, a base REBECA apresenta potencial linguístico-tecnológico para o tratamento computacional do par de línguas PB-Ingl em aplicações como a tradução automática.

Por fim, salienta-se que, apesar da utilização satisfatória do editor Protégé-OWL, esta foi feita de forma totalmente adaptada, posto que o Protégé não foi desenvolvido para suportar os meios representacionais do MultiNet. Além disso, a adequação da interlíngua e da estratégia de alinhamento fora testada apenas com os conceitos do domínio dos "veículos com rodas". Consequentemente, para testar a potencialidade linguístico-computacional da REBECA, outros domínios conceituais precisam ser inseridos na base. Outra limitação do trabalho consiste na construção

manual da base monolíngue do PB, posto que não temos recursos lexicais "legíveis" ou "tratáveis por máquinas" que possam servir de fontes para a delimitação dos conceitos e identificação das expressões linguísticas.

Agradecimento

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

REFERÊNCIAS BIBLIOGRÁFICAS

ALLEN, J. *Natural language understanding*. 2. ed.Massachusetts, CA: Addison Wesley, 1994. 654 p.

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

DIAS-DA-SILVA, B.C.; DI FELIPPO, A., NUNES, M.G.V. The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. In: LREC, 6, 2008, Marrakech, Morocco. *Proceedings...* Marrakech, 2008. p. 335-342.

DI FELIPPO, A; DIAS-DA-SILVA, B. C. REBECA: uma base de dados léxico-conceituais bilíngue inglês-português. In: Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA/SBIA´08), 4, 2008, Salvador-BA, Brazil. *Proceedings...* Salvador, 2008. p. 1-10.

FELLBAUM, C. *WordNet*: an electronic lexical database. Cambridge: MIT Press, 1998. 446 p.

FERNANDES, F. *Dicionário de sinônimos e antônimos da língua portuguesa*. São Paulo: Globo, 2001. 872 p.

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

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

HELBIG, H. *Knowledge representation and semantics for natural language*. Berlin/Heidelberg: Springer-Verlag, 2006. 647 p.

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

HOUAISS, A.; VILLAR, M. S. *Dicionário eletrônico Houaiss da língua portuguesa. (versão 1.0)*. Rio de Janeiro: Editora Objetiva, 2001. 1 CD-ROM.

LANDAU, S. I. (Ed.). *Cambridge dictionary of American English*. Cambridge: CUP, 2000. 1087 p.

LYONS, J. *Semantics*. Cambridge: CUP, 1977. v. 2, 897 p.

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

SAINT-DIZIER, P.; VIEGAS, E. *Computational lexical semantics*. Cambridge: CUP, 1995. 457 p.

SUMMERS, D. (Ed.). *Longman dictionary of contemporary English online*. Longman Group Ltda, 2005. Disponível em: <http://www.ldoceonline.com/>. Acesso em: mai. 2008.

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

WEISZFLOG, W. *Michaelis moderno Dicionário Inglês*. São Paulo: Ed. Melhoramentos, 2000. Disponível em:<http://michaelis.uol.com.br/moderno/ingles/index.php>. Acesso em: mai. 2009.

WILKS Y. A.; SLATOR, B. M.; GUTHRIE, L. M. *Electric words*: dictionaries, computers, and meanings. Cambridge: ACL-MIT Press, 1996. 289 p.