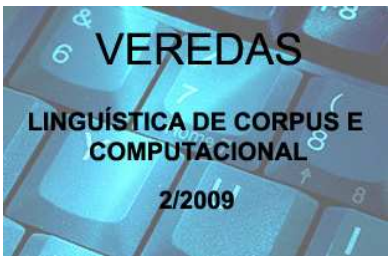

A interlíngua da base lexical bilíngue REBECA

Ariani Di Felippo (DL/UFSCar)
Bento Carlos Dias-da-Silva (FLC/UNESP/Ar.)

RESUMO: No Processamento Automático das Línguas Naturais (PLN), as bases de dados lexicais desempenham papel central em diversos sistemas que processam língua natural. Neste trabalho, apresentamos a interlíngua que foi utilizada na construção da base de dados lexicais bilíngue REBECA, uma das poucas que englobam o português do Brasil. A interlíngua dessa base é composta por um conjunto de conceitos que funciona como elo entre as bases monolíngues do inglês norte-americano e do português brasileiro. Especificamente, apresentamos: a (i) composição, a (ii) estrutura, o (iii) formalismo de representação dos conceitos e a (iv) ferramenta computacional de auxílio à construção da referida base. Como resultado, obtivemos uma interlíngua formal e hirarquicamente organizada, que garantiu o alinhamento eficaz das bases monolíngues da REBECA.

Palavras-chave: Processamento automático das línguas naturais; Base de dados lexicais; Interlíngua; Alinhamento léxico-conceitual; Conceito.

Introdução

No âmbito do Processamento Automático das Línguas Naturais (PLN), os sistemas computacionais que processam (interpretam/ geram) língua natural registrada em meio escrito necessitam manipular conhecimentos linguísticos de vários tipos, dependendo da aplicação para a qual são construídos.[1] De acordo com alguns manuais de PLN, tais tipos estão hierarquizados com base em uma escala de abstração e complexidade, ou seja, quanto mais alto for o nível nessa escala, mais complexos serão a modelagem e o tratamento computacional do conhecimento (Figura 1). No nível mais inferior dessa escala, está o conhecimento morfológico, seguido pelos conhecimentos sintático, semântico e pragmático-discursivo.

[1] O processamento computacional da fala, ou melhor, das línguas naturais em meio sonoro, tem ficado a cargo da área denominada Reconhecimento e Síntese de Fala (do inglês, *Speech Recognition and Synthesis*). Assim, o termo PLN aplica-se ao processamento de língua natural, tanto nas modalidades escrita e oral, "desde que registrada em meio escrito".

VEREDAS ON-LINE – LINGUÍSTICA DE CORPUS E COMPUTACIONAL – 2/2009, P. 50-67 – PPG LINGUÍSTICA/UFJF – JUIZ DE FOR A – ISSN 1982-2243

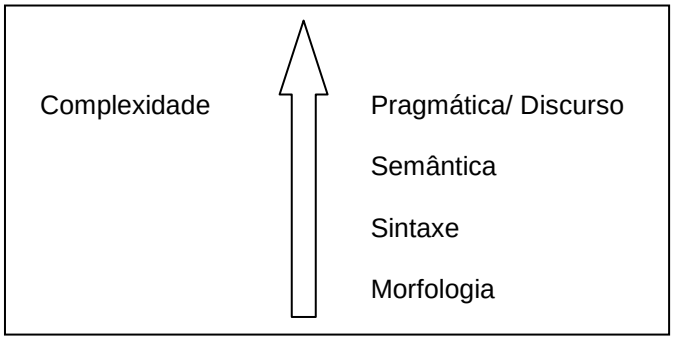

Figura 1: Complexidade e abstração dos níveis de conhecimento no âmbito do PLN.

Atualmente, embora os sistemas de PLN sejam comumente capazes de manipular com certa competência os conhecimentos linguísticos mais concretos, eles não são capazes de fazer o mesmo com os conhecimentos linguísticos mais abstratos. Em outras palavras, isso quer dizer que, apesar de os sistemas realizarem satisfatoriamente os passos básicos de processamento da língua, eles não são capazes de "entender" o que os usuários dizem ou fazem (PALMER, 2001).

A compreensão ou interpretação das línguas naturais, entretanto, tem se tornado essencial para alguns sistemas que processam língua, mais particularmente, para aqueles que processam duas ou mais línguas, como os sistemas de "tradução automática" (do inglês, *machine translation*) e "recuperação de informação multilíngue" (do inglês, *cross-language information retrieval* ou *multilingual information retrieval*).

Tal compreensão requer, naturalmente, a manipulação do conhecimento linguístico nos níveis semântico e pragmático-discursivo, os mais abstratos segundo a hierarquia da Figura 1. Por exemplo, na tentativa de criar sistemas de busca capazes de processar o conteúdo semântico de suas bases textuais, propôs-se[2] um novo padrão para a *web*: a *Web* Semântica (do Inglês, *Semantic Web*). Essa concepção da *web* pauta-se na construção de ontologias[3] para tornar acessível aos sistemas de PLN o conteúdo semântico de documentos.

No caso do tratamento computacional do conhecimento semântico, salientamos que, para "entender" ou "interpretar" o significado de expressões linguísticas simples ou complexas (sintagmas e sentenças) de um texto, é notória a necessidade de recursos bilíngues e/ou multilíngues que armazenam informação semântica sobre as unidades lexicais (SAINT-DIZIER, VIEGAS, 1995; PALMER, 2001, HANKS, 2004).

Nesse cenário, destacamos os recursos multilíngues em que bases monolíngues de línguas distintas estão alinhadas por meio de uma interlíngua, ou seja, uma coleção única de conceitos. O tipo de interlíngua utilizado deve ser capaz de lidar tanto com os casos mais simples, em que há conceitos equivalentes nas línguas envolvidas, quanto com os casos mais complexos, em que há divergências léxico-conceituais entre as línguas. Comumente, são

[2] A *Web Semântica* foi proposta pelo *World Wide Web Consortium* (W3C) http://www.amtechs.com/w3c/w3c7points.html

[3] O termo *ontologia* pode ser entendido como "uma especificação formal de uma conceitualização compartilhada"; "formal", porque deve englobar uma representação formal (ou formalismo) ou explícita; "compartilhada", porque deve registrar uma visão consensual sobre o conhecimento em questão (NIRENBURG, RASKIN, 2004).

utilizados dois tipos de alinhamento: (i) por meio de interlíngua estruturada ou (ii) por meio de interlíngua não-estruturada.

No âmbito do processamento automático do português do Brasil, destacamos (i) o projeto em desenvolvimento que visa ao alinhamento das bases da WordNet de Princeton (WN.Pr) (FELLBAUM, 1998) e da *wordnet* para o PB, a WordNet.Br (WN.Br) (DIAS-DA-SILVA et al., 2008), e (ii) a base bilíngue REBECA (DI FELIPPO, 2008; DI FELIPPO, DIAS-DA-SILVA, 2008). Desenvolvida para o par de línguas inglês norte-americano/português brasileiro (*Ingl*-PB), a REBECA é responsável pelo alinhamento de um conjunto de conceitos lexicalizados (isto é, expressos por unidades lexicais[4]) no *Ingl* a um conjunto de conceitos lexicalizados no PB por meio de uma interlíngua estruturada, a qual é enfatizada neste trabalho.

Para tanto, apresentamos, na Seção 1, os vários tipos de divergências que dificultam o alinhamento léxico-conceitual e, consequentemente, o desenvolvimento de bases lexicais bilíngues e/ou multilíngues. Na Seção 2, apresentamos alguns trabalhos relacionados, enfatizando o tipo de interlíngua utilizado em cada um deles. Na Seção 3, destacamos a interlíngua e as bases monolíngues que compõem a REBECA. Especificamente, apresentamos as principais características de tal interlíngua, como: (i) composição, (ii) estrutura e (iii) formalismo de representação dos conceitos. Na Seção 4, damos destaque à ferramenta computacional que auxiliou a construção da REBECA e, consequentemente, a sua interface de consulta e edição. Por fim, apresentamos os principais problemas e vantagens da adoção de uma interlíngua estruturada na construção de uma base bilíngue.

1. As divergências léxico-conceituais e as lacunas lexicais

No caso do desenvolvimento das bases de dados lexicais em que as unidades de várias línguas estão inter-relacionadas por meio do conceito que elas expressam, a complexidade de se lidar com o nível do conhecimento léxico-semântico torna-se bastante evidente. Tal complexidade deve-se não só à abstração e complexidade do tipo de conhecimento, mas principalmente às divergências léxico-conceituais que existem entre as línguas.

Um conceito pode ser entendido como uma "descrição mental", uma ideia (compartilhada pelos falantes) de um tipo de coisa (p.ex.: objeto, evento ou fenômeno do mundo real ou imaginário) que permite (aos falantes) "discriminar entidades desse tipo das entidades dos demais tipos", ou seja, categorizar (ROSCH, 1973; TAYLOR, 1985; LÔBNER, 2002 CROFT, CRUSE, 2004, CRUSE, 2006).

O processo responsável pela formação dos conceitos recebe o nome de conceitualização. Esse processo opera sobre informações extralinguísticas provenientes de fontes diversas (visual, motora, auditiva, etc.) e tem como norte princípios gerais de organização conceitual, incluindo uma ontologia do senso comum, conceitualizações do espaço e tempo e condições gerais subjacentes ao conhecimento enciclopédico e a sistemas de crenças (BOCK, 1982, LEVELT, 1992, BIERWISCH, SCHREUDER, 1992; HANDKE, 1995).[5]

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

[5] Dessa forma, as comunidades linguísticas apresentam diferentes repertórios conceituais, que revelam diferentes categorizações ou perspectivas de mundo. Essa abordagem, aliás, afasta-se das abordagens denotacional e estrutural do significado. Do ponto de vista denotacional, os significados refletem propriedades e coisas existentes no mundo real ou imaginário e, do ponto de vista do estruturalismo, o significado é arbitrário e interno ao sistema linguístico, totalmente independente da realidade extralinguística (TAYLOR, 1985).

Do ponto de vista ideal, qualquer conceito pode, em princípio, ser expresso no sistema lexical de qualquer língua. No entanto, segundo a abordagem cognitiva, a lexicalização, ou seja, o processo pelo qual um conteúdo semântico é expresso por uma unidade lexical (TALMY, 1985), seja ela simples, como *casa*, composta, como *guarda-roupa*, ou mesmo complexa, como *nota fiscal*, resulta da interação entre convenção e motivação (TAYLOR, 1985; LAKOFF, 1987). Diversos fatores parecem intervir na lexicalização de um conceito, tais como proeminência perceptual, convenção social e linguística e relevância semiótica. A união da maioria desses fatores provavelmente leva à lexicalização de um conceito.

Dessa forma, muitos conceitos são lexicalizados em várias línguas naturais. Por exemplo, o conceito <bicicleta> é comum ao sistema léxico-conceitual do PB,[6] do *Ingl*,[7] do francês, do alemão, etc. Em todas essas línguas, há uma unidade lexical (ou mais) que expressa tal conceito: no PB, *bicicleta;* no *Ingl*, *bicycle*; no francês, *bicyclette*; no alemão, *fahrrad*. Se esse fosse sempre o caso, construir bases de dados lexicais mutilíngues seria uma tarefa relativamente simples. Entretanto, isso nem sempre ocorre, posto que as línguas apresentam divergências no nível léxico-conceitual.

Há, por exemplo, as chamadas divergências conceituais (VOSSEN et al., 1998; CRUSE, 2004), que ocorrem quando um conceito lexicalizado na língua x (ou língua-fonte) não faz parte do repertório geral dos conceitos da língua y (ou língua-alvo). O conceito <frutos secos>, por exemplo, lexicalizado no *Ingl* por *nut*, não é conhecido pelos falantes do francês e do alemão (CRUSE, 2004). O mesmo ocorre com o conceito <um tipo de gim feito da casca do limão>, lexicalizado no holandês por *citroenjenever*, que não é conhecido, por exemplo, pelos falantes do PB e do *Ingl* (PETERS et al., 1998). Tais diferenças geram as chamadas lacunas culturais ou conceituais.

Há também as divergências denotativas e conotativas (ALONGE et al., 1998; VOSSEN et al., 1998, BENTIVOGLI et al., 2000).

O primeiro tipo de divergência ocorre quando, para um conceito lexicalizado na língua x, há um ou mais conceitos aproximados (mais geral ou específico) lexicalizados na língua y. Em outras palavras, dizemos que as unidades de y englobam parcialmente a denotação da unidade de x. Por exemplo, o conceito <cada um dos cinco prolongamentos articulados que terminam as mãos e os pés do homem>, lexicalizado no PB por *dedo*, possui conceitos aproximados lexicalizados no *Ingl*; mais precisamente, o *Ingl* lexicaliza os conceitos <cada um dos cinco prolongamentos articulados que terminam as mãos do homem> (*finger*) e <cada um dos cinco prolongamentos articulados que terminam os pés do homem> (*toe*). Aqui, dizemos que o conceito do PB é "subespecificado" e os do *Ingl* são "superespecificados".

As divergências do segundo tipo também ocorrem quando, para um conceito lexicalizado na língua x, há um ou mais conceitos aproximados lexicalizados na língua y. Nesse caso, no entanto, a equivalência aproximada relaciona-se ao fato de que o conceito lexicalizado em y não carrega os matizes conotativos do conceito lexicalizado em x. Esse é o caso, por exemplo, do conceito <menino na terceira infância e na puberdade; aproximadamente dos sete aos treze anos>, lexicalizado no italiano por *fanciullo*, e que, no PB, por exemplo, aproxima-se do conceito <criança do sexo masculino>, lexicalizado por *menino* e *garoto.*

Além das diferenças conceituais, denotativas e conotativas, ressaltamos as chamadas divergências pragmáticas, que ocorrem quando um conceito lexicalizado na língua *x* não é

[6] A variante brasileira é caracterizada neste trabalho pelas obras lexicográficas e textuais utilizadas como fontes para a identificação dos conceitos lexicalizados.

[7] Ao empregarmos o termo "inglês norte-americano", reconhecemos a existência das variedades norte-americana e britânica da língua inglesa. A caracterização do "inglês norte-americano" é dada pela WN.Pr (FELLBAUM, 1998), construída para essa variante e considerada ponto de partida do trabalho empírico aqui apresentado.

lexicalizado na língua *y* e sim expresso por meio de combinações livres[8] (VOSSEN et al., 1998; CRUSE, 2004; HELBIG, 2006). Um exemplo de divergência pragmática é a expressão do conceito <deter o curso das águas por meio de represas> no PB e no italiano. No PB, esse conceito é expresso pela unidade simples *represar* e no italiano é expresso por meio de uma combinação livre, a saber: *sbarrare con una diga* (no PB, *barrar com um dique*).

Tanto as diferenças conceituais como as pragmáticas geram as chamadas lacunas lexicais (do inglês, *lexical gaps*), ou seja, casos em que não há unidades lexicais em uma dada língua que expressam um conceito lexicalizado em outra língua (CRUSE, 2004).

Diante de tais divergências, observamos que diferentes línguas, em diferentes momentos de sua história, podem dividir de modo diferente um mesmo campo conceitual entre suas unidades lexicais.

A seguir, descrevemos como o problema das divergências léxico-conceituais é tratado no desenvolvimento de algumas bases de dados.

2. Trabalhos relacionados

No caso do desenvolvimento das bases de dados lexicais multilíngues, o método de alinhamento utilizado para relacionar as línguas que fazem parte da base deve ser capaz de lidar tanto com os casos mais simples, em que há conceitos equivalentes lexicalizados nas línguas envolvidas, quanto com os casos mais complexos, em que há divergências léxico-conceituais entre as línguas.

Comumente, são utilizados dois tipos de alinhamento: (i) por meio de interlíngua estruturada ou (ii) por meio de interlíngua não-estruturada. O termo "interlíngua" está sendo empregado aqui como sinônimo de "uma coleção única de conceitos". Quanto ao método por interlíngua não-estruturada, a base multilíngue EuroWordNet[9] (VOSSEN, 1998), desenvolvida para línguas europeias, é um exemplo paradigmático. Quanto ao método por interlíngua estruturada, salientamos a interlíngua do NADIA (SÉRASSET, 1994) e do SIMuLLDA (JANSSEN, 2004), ambos sistemas de gerenciamento de bases de dados lexicais multilíngues.

2.1. A base multilíngue EuroWordNet e o ILI

Na EuroWordNet, o alinhamento é feito por meio de uma interlíngua não-estruturada, ou seja, uma coleção de conceitos em que não há nenhuma espécie de relação entre os mesmos. Tal interlíngua é formada pelo conjunto comum dos conceitos lexicalizados nas línguas que estão armazenadas na base e pelos conceitos lexicalizados em cada uma das línguas. A Figura 2 ilustra tal interlíngua.

Na Figura 2, observamos que o fragmento da interlíngua não-estruturada (denominada *Inter-lingual-index*, ILI) é composto pelos conceitos <finger> e <toe>, advindos do *Ingl*, e <dedo>, advindo do espanhol, e entre os quais não há nenhuma relação.

[8] São combinações que seguem somente regras gerais de sintaxe. Os elementos constituintes dessas combinações podem ocorrer livremente com outros elementos da língua. Além disso, o significado das combinações livres é composicional e os seus constituintes podem ser substituídos por sinônimos.

[9] O potencial de uso da base da EuroWordNet no âmbito do PLN tem sido testado, por exemplo, na tarefa de recuperação de informação multilíngue (PALMER, 2001). Além do potencial tecnológico dessa base, a EuroWordNet permite estudos semânticos comparativos das línguas, posto que cada *wordnet* armazenada revela especificidades do léxico de cada língua (PETERS et al., 1998).

O alinhamento na EuroWordNet, por ser baseado em uma interlíngua não-estruturada, como ilustrado na Figura 2, apresenta vantagens e desvantagens. A principal vantagem reside no fato de que a expansão da interlíngua por meio do acréscimo de conceitos lexicalizados específicos de uma nova língua é relativamente simples.

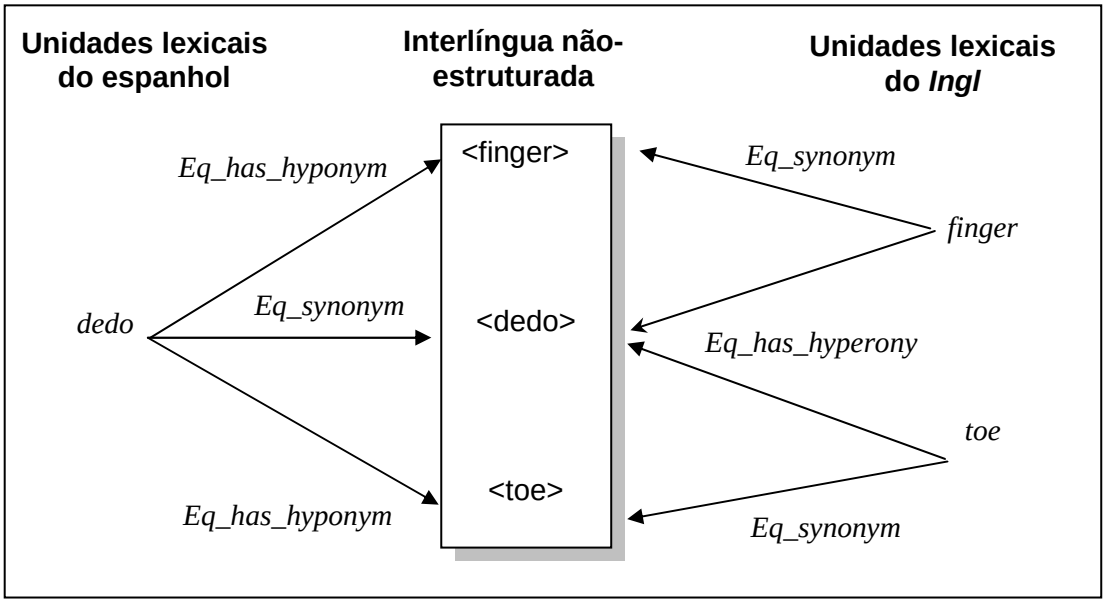

Figura 2: Método de alinhamento por interlíngua não-estruturada.

A principal desvantagem resulta do fato de que um único conceito lexicalizado em uma determinada língua pode ligar-se a vários elementos da interlíngua. Esse é o caso, por exemplo, do conceito lexicalizado em espanhol <dedo>, que se relaciona a três elementos distintos da interlíngua, como ilustrado na Figura 2. Com o acréscimo de novas línguas à base, o número de *links* pode crescer consideravelmente. Vale ressaltar que o alinhamento das bases da WN.Br e WN.Pr está sendo feito nos moldes da EuroWordNet.

2.2. A interlíngua do sistema NADIA

No sistema de gerenciamento NADIA, bases multilíngues são desenvolvidas por meio de um alinhamento baseado em interlíngua estruturada, pois há "certa relação" entre os conceitos que a constituem, como a ilustrada na Figura 3. A interlíngua do NADIA também é formada pelo conjunto comum dos conceitos lexicalizados nas línguas que estão armazenadas na base e pelos conceitos particulares lexicalizados em cada uma das línguas. Dizemos que há "certa relação" entre os conceitos porque a estruturação dos mesmos só é estabelecida diante de casos em que há divergências léxico-conceituais, como o caso ilustrado na Figura 3.

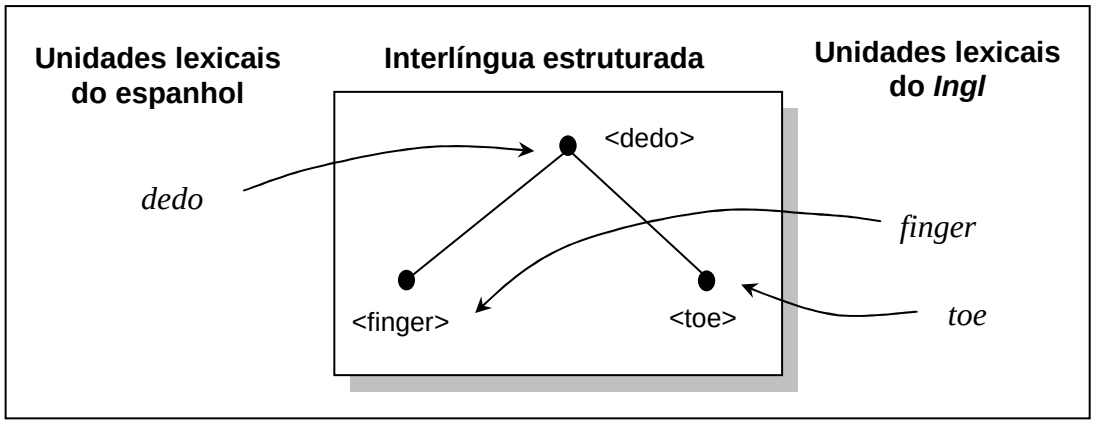

Figura 3: Método de alinhamento por interlíngua estruturada do sistema NADIA.

Na Figura 3, é possível observar que as unidades lexicais das diferentes línguas relacionam-se a um único conceito da interlíngua, posto que os conceitos que a constituem estão organizados.

2.3. A interlíngua do sistema SIMuLLDA

A interlíngua utilizada no sistema de gerenciamento de bases de dados SiMuLLDA é do tipo estruturada, concebida de forma semelhante à interlíngua do NADIA. A arquitetura geral da interlíngua do SIMuLLDA está ilustrada na Figura 4.

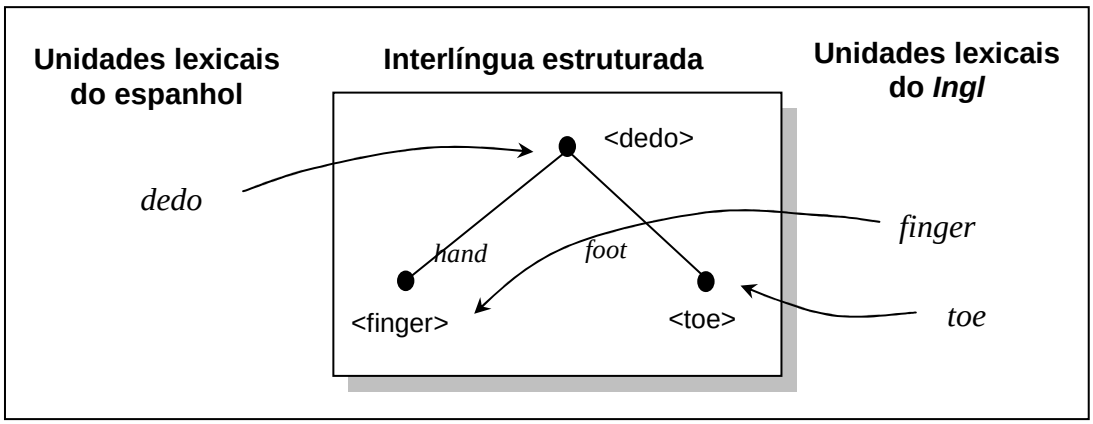

Figura 4: Método de alinhamento por interlíngua estruturada do sistema SIMuLLDA.

Assim como acontece no NADIA, a interlíngua do SiMuLLDA permite que unidades lexicais das diferentes línguas relacionem-se a um único conceito da interlíngua, posto que os conceitos que a constituem estão hierarquicamente organizados.

Por meio da Figura 4, é possível observar que os conceitos <dedo>, <finger> e <toe>, constitutivos da interlíngua, estão hierarquicamente organizados; no caso, <dedo> é o hiperônimo dos conceitos <finger> e <toe>. A diferença entre a interlíngua do NADIA e a do SIMuLLDA reside no fato de que, no SIMuLLDA, os conceitos da interlíngua podem ser especificados com traços semânticos adicionais, denominados “atributos definicionais”. Na

Figura 4, por exemplo, o conceito <finger> possui o atributo definicional *hand* e o conceito <toe>, *foot.*

Quanto à interlíngua utilizada nos sistemas NADIA e SIMuLLDA, salientamos que a principal desvantagem desse método de alinhamento reside no fato de que a inserção de novos conceitos torna-se mais complexa, pois requer uma reestruturação da interlíngua. Por outro lado, a principal vantagem diz respeito ao fato de que o problema do número elevado de *links* entre as línguas e a interlíngua é solucionado.

Por fim, salientamos também que, tanto na EuroWordNet como nos sistemas NADIA e SIMuLLDA, os conceitos integrantes da interlíngua são descritos por meio de uma metalinguagem informal. No entanto, para o processamento automático das línguas, o emprego de uma metalinguagem[10] (semântica) formal para a descrição dos conceitos é essencial, posto que, quanto mais explícito for o conhecimento (no caso, o semântico-conceitual) contido em uma base, mais ela se torna manipulável pelo sistema computacional do qual é parte.

Na próxima Seção, destacamos a (i) composição, a (ii) estrutura e o (iii) formalismo de representação dos conceitos constituintes da interlíngua na base REBECA.

3. A base REBECA

Como mencionado, a base REBECA possui uma interlíngua estruturada, responsável por alinhar ou indexar os *synsets* do *Ingl* aos *synsets* do PB. Esta nada mais é do que o próprio conjunto de conceitos extraídos da WN.Pr que fora representado pelo modelo de representação do conhecimento (RC) MultiNet (HELBIG, 2006). [11]

3.1. O conjunto de conceitos da interlíngua: composição

O conjunto dos conceitos constitutivos da interlíngua pertence ao domínio dos "veículos com rodas" e são todos do tipo "objeto concreto discreto". A escolha desse domínio não se justifica por questões teóricas, mas sim práticas; no caso: delimitação bem-definida e extensão reduzida. Quanto ao tipo conceitual, salientamos que, segundo (LYONS, 1977), 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 linguística, 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.

Além disso, o conjunto de conceitos que constituem a interlíngua da base REBECA 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

[10] Metalinguagem pode ser definida, no caso, como a linguagem em que o significado é descrito ou traduzido.
[11] Vale ressaltar aqui que, uma vez representada 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).

uma glosa (ou seja, uma definição informal) em PB com base principalmente nos dicionários monolíngues do *Ingl* (LANDAU, 2001; SUMMERS, 2005).

3.2. O MultiNet e a interlíngua da base REBECA: estrutura e formalismo

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, adotamos o modelo de RC MultiNet (do inglês, *Multilayered Extended Semantic Networks*), que se baseia na metalinguagem formal das redes semânticas e cujos construtos básicos estão ilustrados na Figura 5.

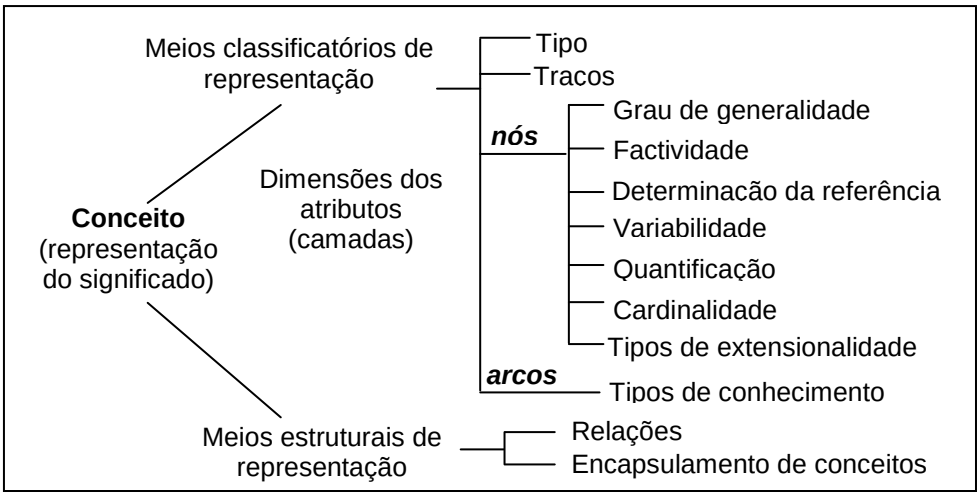

Figura 5: Os construtos de representação do MultiNet.

O MultiNet tem sido empregado principalmente como interlíngua semântica para recuperação de informação na *Web* por meio de interfaces em língua natural. 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 5, os quais são responsável pela macro e microestruturação da interlíngua.

(a) 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 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.[12] 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.

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.

(b) 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 "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 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. 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, ressaltamos 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 6, o conceito <cart> (no PB, *carroça*), elemento constitutivo da interlíngua, é representado pelo MultiNet.

[12] 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.

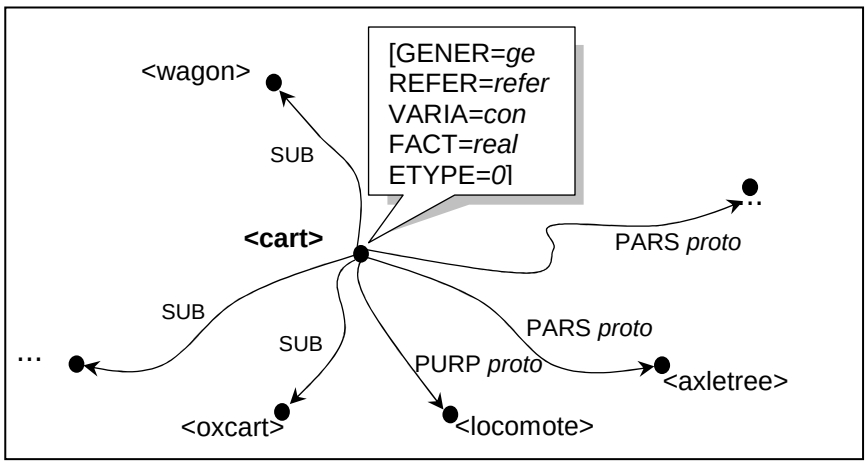

Figura 6: Representação de um conceito lexicalizado segundo o MultiNet.

3.3. As bases léxico-conceituais monolíngues

(a) A base do inglês americano

Com base nos dicionários monolíngues do *Ingl* (LANDAU, 2001; SUMMERS, 2005), 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. 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. Ressaltamos 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*, utilizamos o portal WebCorp[13].

(a) A base do português brasileiro

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 e VILLAR, 2001) e de sinônimos (BARBOSA, 2000; FERNANDES, 1997) foram manualmente consultados para a identificação de unidades sinônimas às compiladas nos dicionários bilíngues e subsequente montagem dos *synsets* do PB. Em uma terceira etapa, verificamos 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[14] e textos disponíveis na *Web*. Os textos em PB disponíveis na *Web* foram consultados através do motor de busca Google,[15] 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.

[13] http://www.webcorp.org.uk/index.html
[14] 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.
[15] http://www.google.com.br/

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 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, obtivemos uma base monolíngue do PB composta por 84 conceitos organizados em *synsets*.
A seguir, focalizamos brevemente a ferramenta computacional utilizada para a construção efetiva da base REBECA.

4. O editor Protégé e a interface da REBECA

Para a construção da base REBECA, utilizamos um dos editores de ontologia mais difundidos na literatura, o Protégé (3.3.1).[16] Especificamente, utilizamos a versão desenvolvida com base na linguagem OWL.[17] 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 e representadas no domínio linguístico-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 monolíngue do *Ingl* e os *synsets* e *phrasets* que compõem a base do PB foram inseridos 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).

Na Figura 7, 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.

[16] http://protege.stanford.edu/
[17] A abreviatura OWL é de *Web Ontology Language*. Mais informações em http://www.w3.org.

Essa exibição é possível devido ao *plug-in*[18] 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 7), baseia-se na tecnologia denominada TouchGraph, que oferece vários recursos de visualização de uma rede conceitual, como alto grau de interação, rápida renderização,[19] visão panorâmica e *zoom*, entre outros.[20]

Na Figura 7, 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 7).

No caso da Figura 7, 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.

[18] 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

[19] 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).

[20] 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.

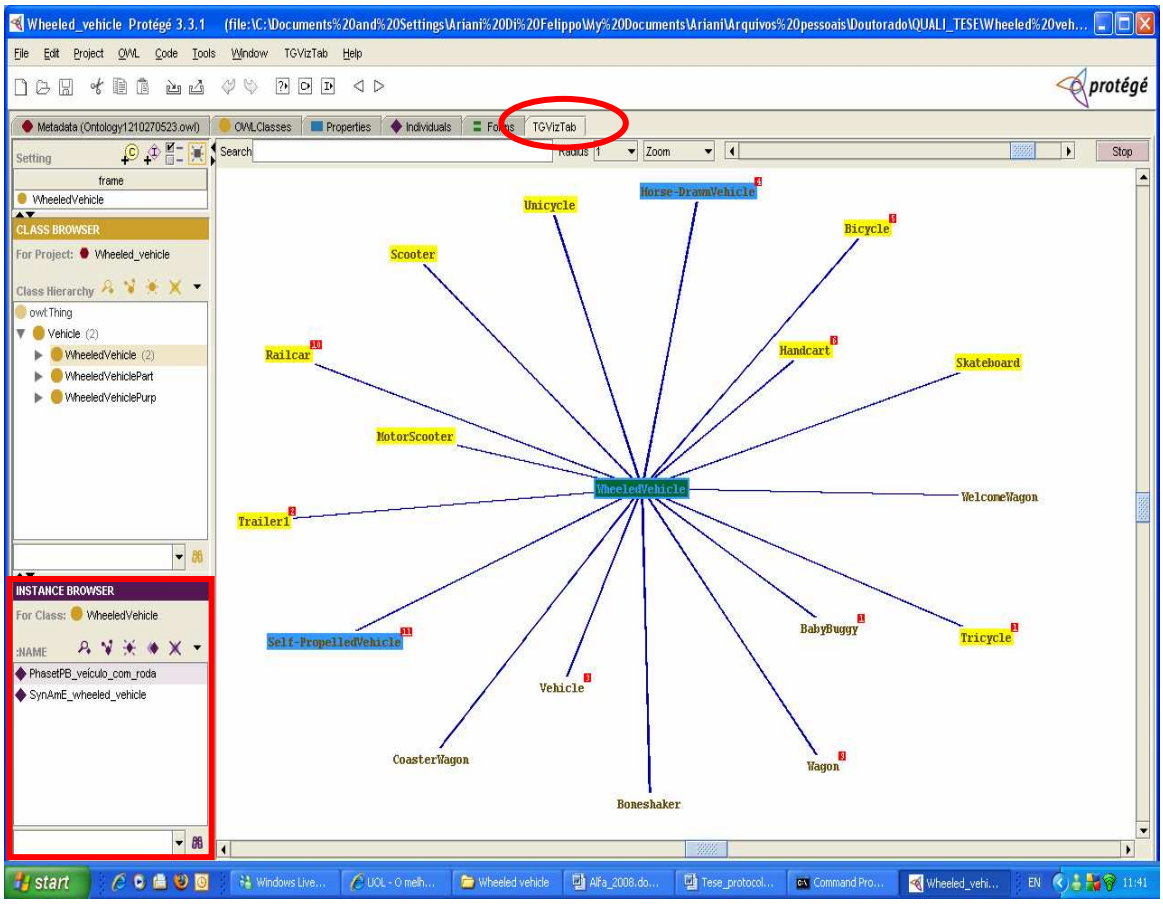

Figura 7: A *interface* do *plug-in* TGVizTab.

Considerações finais

De um modo geral, a base REBECA caracteriza-se, nos moldes da EuroWordNet, 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; (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; (ii) englobar apenas conceitos do tipo "objeto concreto discreto" e pertencentes ao domínio dos "veículos com rodas".

Quanto ao alinhamento, em especial, ressaltamos 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.

Ressaltamos ainda que, nos casos em que há lacunas no PB, a base REBECA é capaz de fornecer, por meio de sua interlíngua estruturada, dois tipos de expressões linguísticas alternativas: os SLRs e a(s) unidades lexicais (ou SLRs) que expressam um conceito hiperônimo.

Na Figura 8, por exemplo, observamos 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 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. 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*.

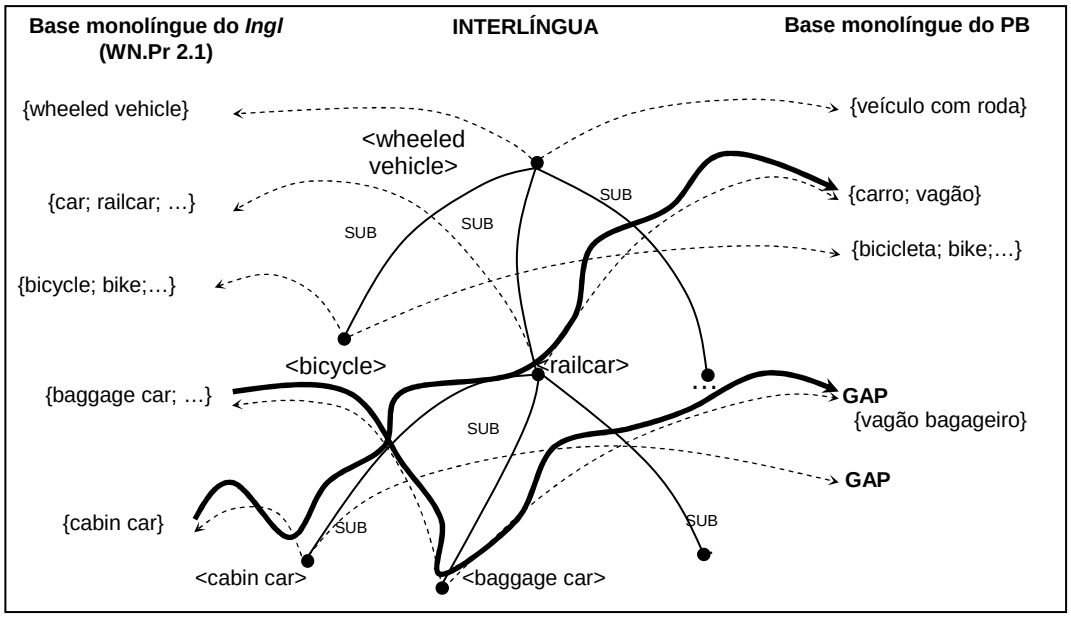

Figura 8: Os alinhamentos léxico-conceituais na base de dados REBECA.

Dessa forma, sob o ponto de vista linguístico, observamos 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 8). Consequentemente, sob o ponto de vista tecnológico, a base REBECA tem 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.

Atualmente, a base REBECA encontra-se disponível no portal de ontologias OntoLP (http://www.inf.pucrs.br/~ontolp/index.php), que divulga principalmente vários tipos de recursos lexicais disponíveis em língua portuguesa (p.ex.: bases de dados lexicais de língua geral, bases terminológicas, vocabulários controlados e ontologias).

Agradecimento

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro à realização da tese de doutorado da qual este trabalho é parte integrante.

ABSTRACT: In Natural Language Processing (NLP) domain, a lexical database, i.e. a systematic stock of words of a natural language, is a crucial component of a wide variety of NLP applications. In this paper, we present the interlingua used in the construction of the bilingual lexical database called REBECA, one of the few lexical resources that encompasses Brazilian Portuguese (BP) language. The interlingua is the core feature of REBECA and it is composed of a set of concepts which allows the alignment of the English and BP monolingual databases. Precisely, we present: (i) the composition; (ii) the structure, (iv) the formalism, and (v) the editing tool used to develop REBECA. . As a result, we obtained a formal and hierarchical interlingua, which allowed a proper link between the monolingual databases.

KEYWORDS: Natural language processing; Lexical database; Interlingua; Lexical-conceptual alignment; Concept.

Referências

ALANI, H. TGVizTab: an ontology visualisation extension for Protégé. In: WORKSHOP KNOWLEDGE CAPTURE (K-Cap'03), WORKSHOP ON VISUALIZATION INFORMATION IN KNOWLEDGE ENGINEERING, Sanibel Island, Florida, USA. *Proceedings...* Sanibel Island, Florida (USA), 2003.

ALONGE, A.; CALZOLARI, N.; VOSSEN, P.; BLOKSMA, L.; CASTELLON, I.; ANTONIA, M.; MARTI, M. A.; PETERS, W. The linguistic design of the EuroWorNet database. *Computers and the Humanities*, Dordrecht: Kluwer Academic Publichers, v.32, p.91-115, 1998.

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

BENTIVOGLI, L.; PIANTA, E.; PIANESI, F. Coping with lexical gaps when building aligned multilingual wordnets. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION - LREC, 2, 2000, Athens. *Proceedings...* Disponível em: <http://multiwordnet.itc.it/english/pubb.php>. Acesso em: 25 out. 2000.

BIERWISCH, M., SCHREUDER, R. From concepts to lexical items. *Cognition*, Amsterdam: Elsevier, v. 42, p. 23-60, 1992.

BOCK, J.K. Towards a cognitive psychology of syntax. *Psychological Review*, n.89, p.1-47, 1982.

CROFT, W., CRUSE, A. *Cognitive linguistics.* Cambridge: Cambridge University Press, 2004.

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

___. *A Glossary of semantics and pragmatics.* United Kingdom: Edinburgh University Press, 2006.

DI FELIPPO, A. *Delimitação e alinhamento de conceitos lexicalizados no inglês norte-americano e no português brasileiro*. 2008. 253f. Tese (Doutorado em Lingüística e Língua Portuguesa), Faculdade de Ciências e Letras (FCL), Universidade Estadual Paulista (UNESP). Araraquara, 2008.

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, 4, 2008. Salvador-BA, Brazil. *Proceedings…* Salvador, 2008.

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.

FELLBAUM, C. (Ed.). *WordNet: an electronic lexical database*. Cambridge, MA: 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.
GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, 43 (5-6), P. 907-928, 1995.
HANDKE, J. *The structure of the lexicon: human vs machine.* Berlin: Mouton de Gruyter, 1995.
HANKS, P. Lexicography. In: MITKOV, R. (Ed.). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press, 2004, cap. 3, p.48-69.
HELBIG, H. *Knowledge representation and semantics for natural language.* Berlin, Heidelberg: Springer-Verlag, 2006.
HOUAISS, A., CARDIM, I. (Orgs.) *Dicionário eletrônico Webster's inglês-português/ português-inglês*. Rio de Janeiro: Ed. Record, 1982. 1 CD-ROM
___, VILLAR, M. de S. *Dicionário eletrônico Houaiss da língua portuguesa*. (versão 1.0). Rio de Janeiro: Editora Objetiva, 2001. 1 CD-ROM
JANSSEN, M. Multilingual lexical databases, lexical gaps, and SIMuLLLDA. *International Journal of Lexicography*, 17, p.136 – 154, 2004.
LAKOFF, G. *Women, fire and dangerous things: what categories reveal about mind.* Chicago: University of Chicago Press, 1987.
LANDAU, S.I. *Cambridge dictionary of American English*. Cambridge: Cambridge University Press, 2001.
LEVELT, W.J.M. *Speaking: to intention to articulation*. Cambridge: The MIT Press, 1992.
LÔBNER, S. *Understanding semantics.* Oxford: Oxford University Press, 2002.
LYONS, J. *Semantics*. Cambridge: Cambridge University Press, 2, 1977.
MICROSOFT PRESS. *Microsoft press dicionário de informática*. Rio de Janeiro: Editora Campus, 805 p., 1998.

NIRENBURG, S.; RASKIN, V. *Ontological semantics.* Cambridge, MA: MIT Press, 2004.
PALMER, M. Multilingual resources, multilingual information management: current levels and future abilities. *Linguistica Computazionale*, v. XIV-XV, p.1-33, 2001.
PETERS, W., VOSSEN, P., DÍEZ-ORZAS, P., ADRIAENS, G. Cross-linguistic alignment of wordnets with an inter-lingual-index. *Computers and the Humanities*, Dordrecht: Kluwer Academic Publichers, v. 32, p. 221-251, 1998.
ROSCH, E. Natural categories. *Cognitive Psychology*, 4, p.328-350, 1973.
SAINT-DIZIER, P., VIEGAS, E. *Computational lexical semantics*. Cambridge: Cambridge University Press, 1995.
SÉRASSET, G. Interlingual lexical organisation for multilingual lexical database in NADIA. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING), 15, 1994, Kyoto, Japan. *Proceedings* ... Kyoto, Japan, 1994, p. 278-282.
SUMMERS, D. (Ed.). *Longman dictionary of contemporary English online*. Longman Group Ltda, 2005. Disponível em: <http//www.ldoceonline.com>
TALMY, L. Lexicalization patterns: semantic structure in lexical forms. In: T. SHOPEN (Ed.) *Language typology and syntactic description: grammatical categories and the lexicon*. (v.3). Cambridge: Cambridge University Press, 1985, p.57-149.
TAYLOR, J. R. *Linguistic categorization: prototypes in linguistic theory.* Oxford: Clarendon Press, 1985.
VOSSEN, P. Introduction to EuroWordNet. *Computers and the Humanities*, Dordrecht: Kluwer Academic Publichers, v.32, p.73-89, 1998.

VOSSEN, P. et al. Compatibility in interpretation of relations in EuroWordNet. *Computers and the Humanities*, Dordrecht: Kluwer Academic Publichers, v. 32, p. 153-184, 1998.
WEISZFLOG, W., *Michaelis: moderno dicionário inglês* (inglês-português/ português-inglês). Ed. Melhoramentos, 2000. Disponível em: <http://michaelis.uol.com.br/moderno/ingles/index.php>