ARIANI DI FELIPPO

Delimitação e Alinhamento de Conceitos Lexicalizados no Inglês Norte-americano e no Português Brasileiro

Araraquara
2008

ARIANI DI FELIPPO

Delimitação e Alinhamento de Conceitos Lexicalizados no Inglês Norte-americano e no Português Brasileiro

Tese apresentada à Faculdade de Ciências e Letras, Universidade Estadual Paulista – Campus de Araraquara, como parte dos requisitos para a obtenção do título de Doutor em Letras (Área de Concentração: Lingüística e Língua Portuguesa).

Linhas de pesquisa: Estudos do léxico; Análise fonológica, mofossintática, semântica e pragmática.

Orientador: *Prof. Dr. Bento Carlos Dias da Silva*

Bolsa: CNPq

Araraquara
2008

Di Felippo, Ariani

Delimitação e alinhamento de conceitos lexicalizados no inglês norte-americano e no português brasileiro / Ariani Di Felippo – 2008.
253 f. : 30 cm

Tese (Doutorado em Lingüística e Língua Portuguesa) – Universidade Estadual Paulista, Faculdade de Ciências e Letras, Campus de Araraquara
Orientador: Bento Carlos Dias-da-Silva

1. Lingüística. 2. Processamento Automático de Língua Natural. 3. Semântica Lexical. 4. Base de dados. 5. Conceito. 4. Lexicalização.

ARIANI DI FELIPPO

DELIMITAÇÃO E ALINHAMENTO DE CONCEITOS LEXICALIZADOS NO INGLÊS NORTE-AMERICANO E NO PORTUGUÊS BRASILEIRO

COMISSÃO JULGADORA

TESE PARA OBTENÇÃO DO GRAU DE DOUTOR

Presidente e Orientador: **Prof. Dr. Bento Carlos Dias da Silva**
1º Examinador: Profa. Dra. Stella Esther Ortweiller Tagnin
2º Examinador: Profa. Dra. Cláudia Zavaglia
3º Examinador: Profa. Dra. Beatriz Nunes de Oliveira Longo
4º Examinador: Profa. Dra. Rosane de Andrade Berlinck

Araraquara, 01 de agosto de 2008

Dedicatória

Dedicada às únicas quatro pessoas vivas que, mesmo não sabendo muito bem o que é Lingüística, ou mesmo Processsamento Automático das Línguas Naturais, acreditam que esta tese iguala-se às obras-primas dessas áreas:

minha mãe, *Luiza*, "a mãe que um milhão de garotas gostariam de ter";

meu pai, *Oscar*, "aquele que perde o amigo, mas não a piada";

minha irmã, *Aline*, "a preferida deles, até que eu vire *doutora*".

meu marido, *Anselmo*, "para quem o cavalo não faz curva".

Agradecimentos

Comumente, o que se encontra nesta seção é a materialização do reconhecimento por vários tipos de apoio recebidos durante os anos em que uma tese fora desenvolvida. Nela, o autor agradece às pessoas e instituições que tiveram papel fundamental na realização do trabalho, enfatizando o que estas fizeram por ele durante o doutorado.

Aqui, no entanto, a ênfase é dada àquilo que as pessoas e instituições, ao atuarem durante o processo, "não" me deixaram fazer ou evitaram que eu fizesse, pois, a meu ver, isso é tão importante quanto aquilo que eu pude realizar com a ajuda delas. Segue, então, meu verdadeiro reconhecimento a:

A **minha família**, que "não permitiu" que eu deixasse de acreditar em mim mesma.

Aos **meus amigos** mais próximos, que "não permitiram" que eu me tornasse uma pessoa sedentária; que eu ficasse ausente das principais sessões de cinema desses anos; que eu faltasse nem mesmo a uma *happy hour* ou que eu ficasse sem "ouvidos" nas horas das dúvidas.

Ao **NILC**, cujos coordenadores e pesquisadores "não permitiram" que eu ficasse sem recursos dos mais variados tipos para a realização da tese; que eu ficasse isolada, sem a chance de participar dos seminários internos do grupo e dos principais eventos de pesquisa e de fazer contato com vários pesquisadores.

Ao **meu orientador**, que "não permitiu" em momento algum que eu ficasse "abandonada", sem uma boa orientação.

À agência de fomento **CNPq**, que "não permitiu" que eu ficasse desprovida de recursos financeiros para a realização desta tese.

**“The only real voyage of discovery consists
not in seeking new landscapes,
but in having new eyes.”**

- Michael Proust -

Resumo

Devido a vários fatores, como saliência perceptual e relevância semiótica, as línguas apresentam repertórios diferentes de conceitos lexicalizados (isto é, conceitos expressos por unidades lexicais). As divergências léxico-conceituais dificultam o tratamento computacional das línguas naturais em tarefas como tradução automática e recuperação de informação multilíngüe. Assim, a construção de base de dados lexicais bilíngües e multilíngües em que as unidades de línguas distintas estão inter-relacionadas por meio do conceito a elas subjacente tem recebido muita atenção no Processamento Automático das Línguas Naturais (PLN). Para o português brasileiro (PB), faz-se urgente a construção desse tipo de recurso. Nesse cenário, esta tese visa a investigar os padrões de lexicalização do PB e a construir um recurso léxico-conceitual, ainda que de extensões reduzidas, que possa auxiliar o processamento automático dessa língua em meio escrito. Assumindo-se a concepção de PLN enquanto "uma engenharia da linguagem humana", utilizou-se uma metodologia tripartida que divide as atividades nos domínios: lingüístico, lingüístico-computacional e computacional. Este trabalho, em especial, não realizou as atividades previstas no terceiro domínio, pois estas não fazem parte do escopo desta pesquisa. No domínio lingüístico, um conjunto de conceitos lexicalizados no inglês norte-americano (AmE), extraído da WordNet de Princeton (WN.Pr), foi delimitado por meio da análise manual de recursos estruturados (base de dados e dicionários) e não-estruturados (*corpora* textuais). Na seqüência, as expressões do PB (em especial, as unidades lexicais) que materializam tais conceitos foram manualmente extraídas de dicionários bilíngües (AmE-PB), dicionários monolíngües e *thesaurus* e de *corpora* textuais do PB. No domínio lingüístico-computacional, os conceitos lexicalizados no AmE e PB, identificados na fase anterior, foram alinhados por meio de uma interlíngua estruturada (ou ontologia). Essa interlíngua segue os construtos do modelo de representação do conhecimento MultiNet, o qual fornece meios de representação robustos para formalizar a semântica das línguas naturais. O alinhamento foi feito no editor Protégé-OWL, um dos mais utilizados para se criar e editar ontologias, o que resultou em uma base de dados léxico-conceituais, denominada REBECA. Nessa base, uma porção do léxico do PB está diretamente alinhada a uma porção da WN.Pr. Conseqüentemente, os padrões de lexicalização apresentados por essas línguas podem ser comparados. A base REBECA tem potencial para ser utilizada em várias aplicações de PLN, por exemplo, na recuperação de informação multilíngüe. Nessa tarefa, a base pode expandir as unidades lexicais de uma língua para unidades relacionadas na outra língua via a interlíngua. Além disso, essa base pode gerar automaticamente um dicionário bilíngüe do domínio dos "veículos com rodas". Ainda no domínio lingüístico-computacional, o alinhamento dos conceitos lexicalizados no AmE e PB nos moldes do projeto EuroWordNet também foi realizado. Essa tarefa contou com o auxílio do *plug-in* de visualização do Protégé-OWL denominado TGVizTab. Tais alinhamentos podem ser diretamente utilizados no projeto em andamento que visa ao mapeamento da *wordnet* do português brasileiro (a WordNet.Br) à WN.Pr. A base bilíngüe resultante desse alinhamento será um importante recurso para os lingüistas e investigadores do PLN. Dessa forma, esta tese é uma tentativa de contribuir para a identificação das diferenças léxico-conceituais entre o AmE e o PB e para o processamento automático do PB, promovendo a visão lingüisticamente motivada das pesquisas de PLN e fortificando o trabalho colaborativo entre lingüistas e cientistas da computação.

Palavras-chave: conceitos; unidades lexicais; alinhamento léxico-conceitual; representação do conhecimento; semântica.

Abstract

Because of several factors, including, for instance, perceptual salience and semiotic relevance, languages have different inventories of lexicalized concepts (i.e. concepts expressed by lexical units). The lexical-conceptual divergences are a hindrance to computational treatment of natural languages in tasks such as machine translation and cross-language information retrieval. Therefore, the construction of bilingual and multilingual lexical databases, in which the lexical units of different languages are aligned by their underlying concepts, has become a very important research topic in Natural Language Processing (NLP). For Brazilian Portuguese (BP), in particular, the construction of such resources is urgent. In this scenario, this thesis aims to investigate lexicalization patterns of BP and to develop a lexical-conceptual resource for the automatic processing of written BP language. Assuming a compromise between NLP and Linguistics, this work follows a three-domain approach methodology, which claims that the research activities should be divided into the linguistic, linguistic-computational, and computational domains. In particular, this research does not perform the last step, since it is not in the scope of this work. Accordingly, in the linguistic domain, a set of lexicalized concepts of North-American English (AmE) extracted from Princeton WordNet (WN.Pr) was selected through manual analysis of the structured (lexical databases and standard dictionaries) and unstructured resources (textual *corpora*). Given those concepts, their lexical and phrasal expressions in BP were manually compiled from bilingual dictionaries, with the help of standard monolingual dictionaries, thesauri, and textual *corpora*. In the linguistic-computational domain, the lexicalized concepts of AmE and BP previously identified were aligned by means of a semantic structured interlingua (or ontology). The interlingua is composed of the same set of concepts extracted from WN.Pr and its structure relies on the MultiNet, a specific knowledge representation formalism. MultiNet provides the semantic representatives for the description of the semantics of natural language expressions. The alignment was done in the Protégé-OWL editor, one of the most popular tools to create and edit ontologies. The alignment result is a bilingual lexical-conceptual database, named REBECA. In this database, part of the BP lexicon is strictly aligned with part of WN.Pr. Thus the different lexicalization patterns can be compared and checked cross-linguistically. REBECA has the potential to be used in several NLP tasks, for instance, in multilingual cross-information retrieval, by expanding words in one language to related words in another language via the interlingua. Besides, even in the linguistic-computational domain, the alignment of the lexicalized concepts in AmE and BP based on the EuroWordNet method was done using the Protégé-OWL visualization plug-in TGVizTab. The results of these alignments can be directly applied to the ongoing work of mapping the Brazilian Portuguese wordnet (WordNet.Br) onto the WN.Pr. The alignment of these two databases will result in an important bilingual resource for NLP and linguistic communities. Consequently, this thesis is an attempt to contribute both to the identification of divergences of lexicalization patterns between AmE and BP and to the automatic processing of BP, fostering the linguistically-motivated conception of NLP and the collaborative work between linguists and computational scientists.

Keywords: concepts; lexical units; lexical-conceptual alignment; knowledge representation; semantics.

Índice de Quadros

Quadro 1: Exemplos de classificação dos conceitos nominais.....82
Quadro 2: Os traços para a caracterização dos objetos. 90
Quadro 3: As expressões no AmE dos conceitos do tipo <wheeled vehicle>.....109
Quadro 4: O *template* conceitual. 114
Quadro 5: O teste para a construção de *synsets*. 122
Quadro 6: O *template* lexical. 128
Quadro 7: O *template* léxico-conceitual. 129
Quadro 8: O *template* léxico-conceitual do conceito <motor vehicle>. 131
Quadro 9: O *template* léxico-conceitual do conceito <car>. 134
Quadro 10: O *template* léxico-conceitual do conceito <touring car>.....136
Quadro 11: Os conceitos lexicalizados no AmE do domínio dos "veículos com rodas" e sua expressão no PB. 152
Quadro 12: As estatísticas das lexicalizações estudadas no PB. 152
Quadro 13: A matriz lexical, polissemia e a sinonímia. 154
Quadro 14: Extensões da WN.Br: *synsets* (e glosas) novos. 184
Quadro 15: Refinamento da WN.Br: *synsets* modificados. 184
Quadro 16: Refinamento da WN.Br: *synsets* confirmados.....184
Quadro 17: Os alinhamentos por meio da relação EQ_SYNONYM. 188
Quadro 18: As relações interlinguais complexas estabelecidas pelo *synset* {carro; vagão}...191
Quadro 19: Os alinhamentos por meio das relações interlinguais complexas. 194

Índice de Figuras

Figura 1: Complexidade e abstração dos níveis de conhecimento no âmbito do PLN. 1
Figura 2: Método de alinhamento por interlíngua não-estruturada. 5
Figura 3: Método de alinhamento por interlíngua estruturada. 6
Figura 4: A arquitetura padrão de um sistema de PLN de Dias-da-Silva (1996). 12
Figura 5: Amostra da organização dos *synsets* na WN.Pr. 17
Figura 6: Informações associadas a um *synset*. 18
Figura 7: A arquitetura da base multilíngüe EuroWordNet. 20
Figura 8: Um exemplo de alinhamento na EuroWordNet por EQ_SYNONYM. 21
Figura 9: Diferentes tipos de alinhamento da EuroWordNet. 22
Figura 10: A interlíngua estruturada do sistema NADIA. 24
Figura 11: Acepções de *river*. 24
Figura 12: Acepções de *rivière* e *fleuve*. 25
Figura 13: Interlíngua e alinhamentos do NADIA. 25
Figura 14: Representação da denotação de *bicicleta* em termos de conjuntos. 45
Figura 15: Descrição formal da intensão de *bicicleta*. 46
Figura 16: Representação da denotação de *A bicicleta quebrou* em termos de conjuntos. 46
Figura 17: Relação entre língua, mente e mundo. 47
Figura 18: Níveis de categorização. 49
Figura 19: Elementos da estrutura interna de um *frame*. 55
Figura 20: Conexões entre *frames*. 56
Figura 21: Ilustração dos construtos básicos de uma rede semântica. 57
Figura 22: Exemplo de um fragmento de uma rede semântica. 59
Figura 23: Rede semântica que representa parte do significado expresso pela frase *Mariana quer a bicicleta*.. 60
Figura 24: Fragmento da rede de hierarquia estruturada KL-ONE. 61
Figura 25: Fragmento de uma rede semântica nos moldes do MultiNet. 63
Figura 26: O lugar da representação do conhecimento nos diferentes “mundos”. 64
Figura 27: Os meios de representação semântica do MultiNet. 68
Figura 28: Exemplo das combinações do tipo [SORT=*co*] e dos traços dos objetos concretos (*con-object*). 73
Figura 29: Os três tipos de conhecimento no MultiNet. 75
Figura 30: Os valores do atributo K-TYPE. 77
Figura 31: Os atributos multidimensionais do MultiNet e seus valores. 78
Figura 32: Organização dos tipos [SORT=*ent*] em função do atributo complexo LAY. 83
Figura 33: O meios de representação da estrutura conceitual. 85
Figura 34: Entrada lexical enriquecida com informação enciclopédica. 89
Figura 35: Exemplos de dependências entre tipos e traços. 92
Figura 36: Relações e funções para a caracterização semântica dos objetos concretos. 95
Figura 37: Um exemplo de parte de uma hierarquia de objetos conceituais. 96
Figura 38: Dois mecanismos de herança. 97
Figura 39: As relações e funções do nível pré-extensional. 101

Figura 40: Esquema do processo de identificação e compilação das unidades lexicais do PB. 126
Figura 41: Os conceitos como classes. 160
Figura 42: A inserção das glosas como comentários de classes. 162
Figura 43: A inserção das relações como `Object Properties` e do tipo conceitual, traços semânticos e atributos multidimensionais como `Datatype Properties`. 163
Figura 44: As relações e os conceitos por elas relacionados. 164
Figura 45: A herança de informações. 166
Figura 46: Gerenciamento das relações como propriedades. 167
Figura 47: A especificação das relações inversas. 168
Figura 48: A inserção das expressões lingüísticas como instâncias. 169
Figura 49: A especificação das expressões lingüísticas nos campos do módulo `Individual Editor`. 170
Figura 50: A *interface* do *plug-in* TGVizTab, exibindo o conceito WheeledVehicle no centro do grafo. 172
Figura 51: Painel de Configurações do TGVizTab: abas `Slots` e `Classes`. 173
Figura 52: Exemplo do grafo gerado pelo TGVizTab. 174
Figura 53: A exibição dos rótulos dos arcos do grafo. 175
Figura 54: *Menu* do TGVizTab que fornece opções de controle para os nós do grafo. 175
Figura 55: Distinção gráfica que representa os conceitos lexicalizados e os não-lexicalizados no PB. 176
Figura 56: Estrutura ontológica da base REBECA. 178
Figura 57:Um exemplo de identificação de expressões alternativas para conceitos não-lexicalizados. 179
Figura 58: A base REBECA no formato OWL. 180
Figura 59: As relações de hierarquia estabelecidas por <railcar>. 189
Figura 60: Identificação das relações interlinguais complexas com o TGVizTab. 190
Figura 61: Interface gráfica do editor da WN.Br para o alinhamento léxico-conceitual. 195
Figura 62: O editor da WN.Br e o alinhamento por EQ_HAS_HYPERONYM. 197
Figura 63: Um exemplo de diferentes estruturações léxico-conceituais. 201

SUMÁRIO

Seção I 1
Introdução 1
1.1.Contexto – motivação e justificativa 1
1.2. Objetivos 7
1.3. Metodologia 8
1.4. Estrutura do texto 9

Seção II 11
Bases de Dados Lexicais Bilíngües e Multilíngües 11
2.1. A arquitetura de um sistema de PLN padrão 11
2.2. As bases de dados multilíngües: dois exemplos 13
2.2.1. A base EuroWordNet 14
2.2.1.1. A WordNet de Princeton 14
2.2.1.2. A WN.Pr como uma ontologia lingüística 18
2.2.1.3. Características da EuroWordNet 19
2.2.2. A interlíngua do sistema NADIA 23
2.3. Síntese da Seção II 26

Seção III 27
O Processamento Automático das Línguas Naturais 27
3.1. Visão geral 27
3.2. O PLN e as representações de conhecimento (RC) 36
3.2.1. A RC como um substituto do conhecimento 37
3.2.2. A RC como um conjunto de compromissos ontológicos 37
3.2.3. RC como teoria do raciocínio 39
3.2.4. A RC como um meio de processamento computacional eficiente 40
3.2.5. A RC como um meio de expressão humana 40
3.3. Três paradigmas de RC 40
3.3.1. As RCs baseadas na lógica 41
3.3.1.1. Abordagem teórica do significado: referencial 41
3.3.1.2. A metalinguagem formal dos modelos baseados na lógica 44
3.3.2. As RCs baseadas em *frames* e em redes semânticas 47
3.3.2.1. Abordagem teórica do significado: cognitiva 47
3.3.2.2. A organização dos conceitos 50
3.3.2.3. Outras abordagens do significado: estruturalista e pragmática 52
3.3.2.4. A metalinguagem formal dos modelos baseados em frames 55
3.3.2.5. A metalinguagem formal dos modelos baseados em redes semânticas 57
3.4. Síntese da Seção III 62

Seção IV 63
O paradigma MultiNet: uma linguagem de representação do conhecimento lingüisticamente codificado 63
4.1. Introdução 63
4.2. Os meios de representação dos conceitos no MultiNet 67

4.2.1. A classificação dos conceitos 68
4.2.1.1. Tipos conceituais e traços semânticos 68
4.2.1.2. Tipos de conhecimento classificados no MultiNet 75
4.2.2. Os atributos multidimensionais 77
4.2.3. Os meios de estruturação dos conceitos 84
4.2.3.1. Relações e funções 84
4.2.3.2. Encapsulamento de conceitos 86
4.2.3.3. Relacionamentos axiomáticos 86
4.3. Representação do conhecimento e léxico: a relação entre conhecimento de mundo e conhecimento lingüístico no MultiNet 87
4.3.1. Abordagem em um nível 88
4.3.2. Abordagem em dois níveis 88
4.3.3. O MultiNet e o componente semântico do léxico 90
4.4. Síntese da Seção IV 92

Seção V 94
O MultiNet e a representação dos objetos 94
5.1. A caracterização semântica dos objetos 94
5.1.1. SUB – relação de subordinação 96
5.1.2. Caracterização intraobjetiva 97
5.1.2.1. Caracterização material 97
5.1.2.2. Caracterização qualitativa 98
5.1.3. Caracterização interobjetiva 99
5.1.3.1. Atribuição 99
5.1.3.2. Comparação 99
5.1.4. Caracterização espaço-temporal 100
5.1.4.1. Caracterização espacial 100
5.1.4.2. Caracterização temporal 100
5.1.5. Caracterização télica 100
5.2. As relações e funções do nível pré-extensional 101
5.2.1. As relações e funções no domínio dos conjuntos 101
5.2.2. As relações de conexão entre os níveis intensional e pré-extensional 102
5.3. Síntese da Seção V 102

Seção VI 104
A identificação de parte dos "objetos concretos discretos" e sua expressão do PB 104
6.1. Delimitação do tipo conceitual 104
6.2. Delimitação do domínio conceitual 104
6.3. Seleção do conjunto de conceitos expressos no AmE 105
6.4. Especificação dos conceitos lexicalizados no AmE 110
6.4.1. Critérios teóricos 110
6.4.1.1. Nível de significado 110
6.4.1.2. Tipo de significado 112
6.4.1.3. O conjunto de relações consideradas 112
6.4.2. Proposição do *template conceitual* 113
6.5. Identificação e compilação das expressões lingüísticas do PB 117
6.5.1. Critérios teóricos 117
6.5.1.1. A expressão lexical dos objetos 117
6.5.1.2. A montagem de synsets 122

6.5.2. O método de identificação e compilação das expressões lingüísticas e os recursos lexicais 123
6.5.3. Etapas da identificação das expressões lingüísticas do PB 124
6.5.4. Proposição e preenchimento do *template lexical* 127
6.6. Exemplo de preenchimento dos *templates* léxico-conceituais 130
6.7. Os conceitos e as expressões lingüísticas do PB (unidades lexicais e SLRs) 137
6.7.1. Dados estatíticos 152
6.7.2. Observações sob a perspectiva semasiológica 153
6.8. Síntese da Seção VI 154

Seção VII 156
Construção da base léxico-conceitual bilíngüe e o alinhamento das WN.Pr e WN.Br 156
7.1. O editor Protégé 157
7.1.1. O editor Protégé-OWL: noções preliminares 158
7.2. Adaptações do editor para o alinhamento e para a criação da base bilíngüe 159
7.2.1. Inserção das informações no editor e a criação da base bilíngüe 159
7.2.1.1. Os conceitos como classes 159
7.2.1.2. As relações como propriedades 162
7.2.1.3. As expressões lingüísticas como instâncias 169
7.2.2. A visualização gráfica da interlíngua e das expressões lingüísticas de seus conceitos constitutivos 171
7.3. A base léxico-conceitual bilíngüe REBECA 177
7.4. Contribuições para o desenvolvimento da WordNet.Br 181
7.4.1. A WN.Br 181
7.4.2. Contribuições para o refinamento e extensão dos *synsets* da WN.Br 182
7.4.3. Contribuições para o alinhamento das bases WN.Br e WN.Pr 185
7.4.3.1. A tarefa de alinhamento léxico-conceitual 185
7.4.3.2. A identificação das relações interlinguais por meio do plug-in TGVizTab 188
7.4.4. O editor da WN.Br e a tarefa de alinhamento léxico-conceitual 194
7.5. Síntese da Seção VII 198

Seção VIII 199
Considerações finais e etapas futuras 199

Referências bibliográficas 207

APÊNDICE 1: 223
APÊNDICE 2: 232

Convenções tipográficas

Fonte Arial entre os sinais de maior e menor < >
Para conceitos.
Exemplos: <casa> e <bater as botas>

Fonte Arial com a letra inicial maiúscula
Para a nomeação dos conceitos no Protégé-OWL.
Exemplos: Car, WheeledVehicle

Fonte Times New Roman em itálico
Para expressões lingüísticas e para termos em língua estrangeira.
Exemplos: *car, carro/ graphical interface.*

Fonte Times New Roman em negrito
Para termos técnicos acompanhados (ou não) por definição.
Exemplos: **lexicalização** e **conceito.**

Fonte Times New Roman entre aspas duplas
Para ênfase.
Exemplo: "engenheiros da linguagem"

Fonte Times New Roman entre aspas simples
Para citações.
Exemplo: '[...] qualquer questão em Lingüística nunca será resolvida para a satisfação de todos'.

Fonte Times New Roman em itálico e entre aspas simples
Para títulos de livros, artigos, etc., ao longo do texto.
Exemplo: '*A interface léxico-enciclopédia*'

Fonte Courier New
Para construtos específicos do editor Protégé-OWL.
Exemplos: `Object Property` e `Datatype Property`

Fonte Times New Roman em maiúsculo
Para categorias conceituais
Exemplos: VEÍCULO e AVE

Seção I

Introdução

1.1.Contexto – motivação e justificativa

No âmbito do Processamento Automático das Línguas Naturais (doravante, PLN), os sistemas computacionais que processam (interpretam/ geram) língua natural registrada em meio escrito necessitam de conhecimento lingüístico em suas várias dimensões, dependendo da aplicação para a qual são construídos. Os manuais de PLN baseiam-se em uma hierarquia de tipos de conhecimento lingüístico, elaborada 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.

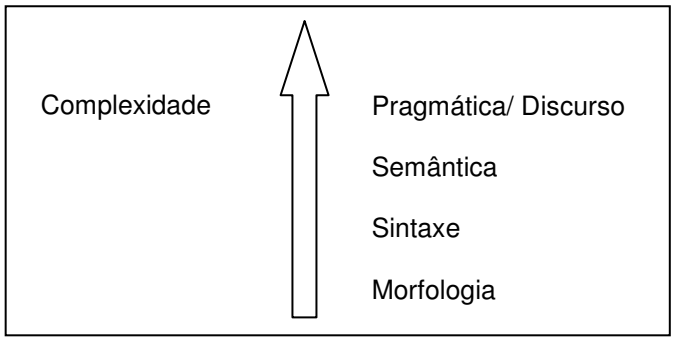


Figura 1: Complexidade e abstração dos níveis de conhecimento no âmbito do PLN.

Apesar de os sistemas desenvolvidos realizarem satisfatoriamente os passos básicos de processamento da língua, eles não são capazes de "entender" o que os usuários dizem ou fazem (PALMER, 2001). Essa compreensão tem se tornado essencial para alguns sistemas que processam língua, mais particularmente, para aqueles que processam duas ou mais línguas, como os sistemas de "tradução automática" (do inglês, *machine translation*) e "recuperação de informação multilíngüe" (do inglês, *cross-language information retrieval* ou *multilingual information retrieval*). Para tanto, é notória a necessidade de se tratar o conhecimento de nível semântico, um dos mais abstratos e complexos de acordo com a escala da Figura1.

Quanto mais abstrato e complexo for o tipo de conhecimento lingüístico, mais complexo é também o desenvolvimento de recursos lexicais que armazenam esse conhecimento. No caso do tratamento computacional do conhecimento semântico, salienta-se que, para "entender" ou "interpretar" as expressões lingüísticas simples ou complexas (sintagmas e sentenças) de um texto, é notória a necessidade de recursos bilíngües e/ou multilíngües que armazenam informação semântica sobre as unidades lexicais (SAINT-DIZIER, VIEGAS, 1995; PALMER, 2001, HANKS, 2004).

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 fica 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 idéia (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 extralingüísticas 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)[1].

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,

[1] Dessa forma, as comunidades lingüísticas 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 lingüístico, totalmente independente da realidade extralingüística (TAYLOR, 1985).

tais como proeminência perceptual, convenção social e lingüística 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 português brasileiro[2] (doravante, PB), do inglês norte-americano[3] (doravante, AmE; do inglês, *American English*), do francês, do alemão, etc. Em todas essas línguas, há uma unidade lexical (ou mais) que expressa tal conceito: no PB, *bicicleta;* no AmE, *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 AmE por *nut,* não é conhecido pelos falantes do francês e do alemão (CRUSE, 2004). O mesmo acontece 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 AmE (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, diz-se 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 AmE; mais precisamente, o AmE 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, diz-se que o conceito do PB é "subespecificado" e os do AmE são "superespecificados". As divergências do segundo tipo também ocorrem

[2] O emprego do termo "português brasileiro" indica que se considera as variantes brasileira e européia do português. A variante brasileira é caracterizada neste trabalho pelas obras lexicográficas e textuais utilizadas como fontes para a identificação dos conceitos lexicalizados.

[3] Ao se empregar o termo "inglês norte-americano", reconhece-se a existência das variedades norte-americana e britânica da língua inglesa. A caracterização do chamado "inglês norte-americano" é dada pela WordNet de Princeton (doravante, WN.Pr) (FELLBAUM, 1998a,b), construída para essa variante e considerada ponto de partida do trabalho empírico desta tese.

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, ressaltam-se as chamadas **divergências pragmáticas**, que ocorrem quando um conceito lexicalizado na língua *x* não é lexicalizado na língua *y* e sim expresso por meio de combinações livres[4] (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 (ALLAN, 2001; CRUSE, 2004).

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

No caso do desenvolvimento das bases de dados lexicais multilíngües, 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.

Quanto às bases que utilizam o método de alinhamento baseado em **interlíngua**, em que a ligação entre os conceitos lexicalizados por línguas distintas é feita mediante uma coleção única de conceitos, salientam-se a base multilíngüe EuroWordNet[5] (VOSSEN, 1998),

[4] 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.

[5] O potencial de uso da base da EuroWordNet no âmbito do PLN tem sido testado, por exemplo, na tarefa de recuperação de informação multilíngüe (PALMER, 2001; GONZALO et al., 1998). 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).

desenvolvida para línguas européias, e a estratégia, baseada em interlíngua, do NADIA (SÉRASSET, 1994a,b), sistema de gerenciamento de bases de dados lexicais multilíngües.

Na EuroWordNet, o alinhamento é feito por meio de uma interlíngua **não-estruturada**, ou seja, uma coleção de conceitos em que não há inter-relações. 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.

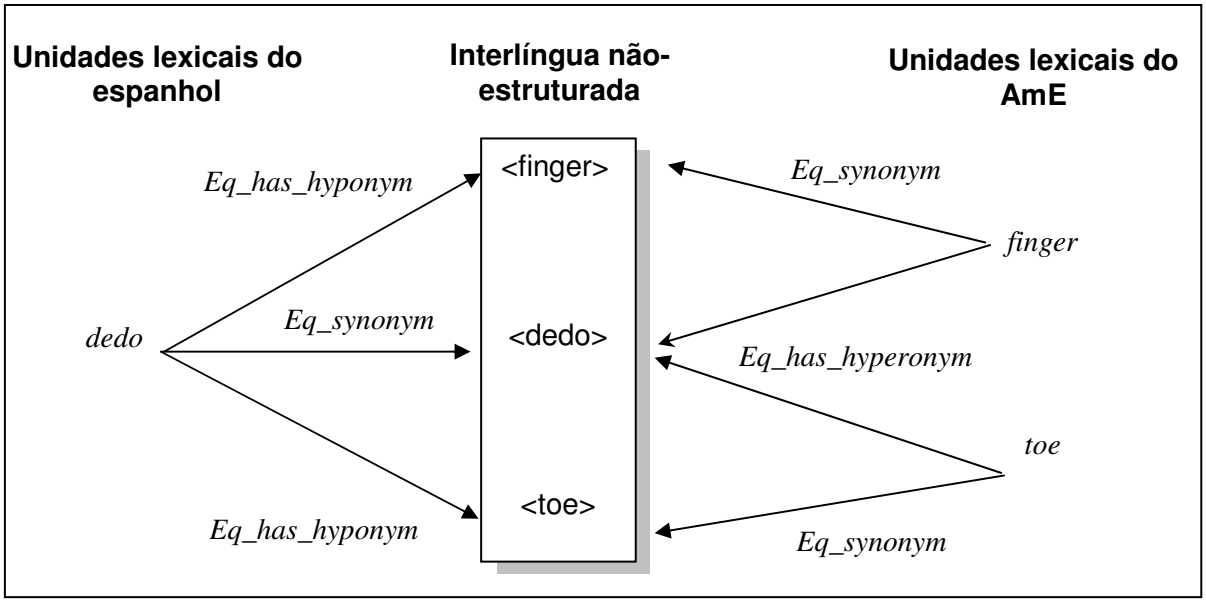


Figura 2: Método de alinhamento por interlíngua não-estruturada.

Na Figura 2, a interlíngua não-estruturada (denominada *Inter-lingual-index,* ILI) é composta pelos conceitos <finger> e <toe>, advindos do AmE, e <dedo>, advindo do espanhol, e entre os quais não há nenhuma relação.

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.

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 liga 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.

No sistema de gerenciamento NADIA há pouco mencionado, bases multilíngües 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. Diz-se que há "certa relação" entre os conceitos porque a estruturação dos mesmos só é estabelecida diante de casos em que há divergências léxico-conceituais.

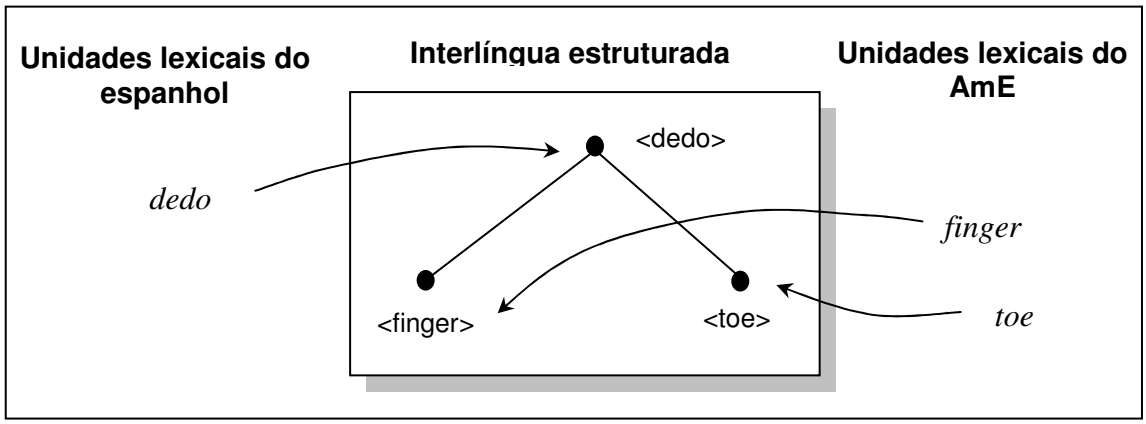


Figura 3: Método de alinhamento por interlíngua estruturada.

Na Figura 3, vê-se que as unidades lexicais das diferentes línguas relacionam-se a um único conceito da interlíngua, posto que os conceitos que a constituem estão organizados. Se, por um lado, a inserção de novos conceitos torna-se mais complexa, pois requer uma reestruturação da interlíngua, por outro lado, o problema do número elevado de *links* entre as línguas e a interlíngua é solucionado.

Por fim, salienta-se que, tanto na EuroWordNet como no NADIA, 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[6] (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.

Para o português do Brasil, a base bilíngüe, resultante do alinhamento (nos moldes da EuroWordNet) das bases da WN.Pr e da *wordnet* para o português brasileiro, a WordNet.Br (doravante, WN.Br) (DIAS-DA-SILVA et al., 2002; DIAS-DA-SILVA et al., 2008), será o único recurso lexical para o par AmE-PB em que as unidades lexicais relacionam-se em função de seu significado (cf. Subseção 7.4.3, p. 185).

[6] Metalinguagem pode ser definida, no caso, como a linguagem em que o significado é descrito ou traduzido.

1.2. Objetivos

Diante desse cenário, objetivam-se nesta tese:

(i) delimitar um conjunto específico de conceitos lexicalizados no AmE;

(ii) identificar as lexicalizações desses conceitos no PB (ou seja, as unidades lexicais que os expressam); ou melhor, identificar os chamados **padrões de lexicalização**, ou seja, as associações entre unidade lexical e conceito (TALMY, 1985).

(iii) descrever esse conjunto de conceitos em um modelo formal de representação, de tal forma que esses conceitos funcionem como uma interlíngua estruturada e formal entre o AmE e o PB.

Dessa forma, o AmE é a língua-fonte e o PB, a alvo. Além disso, vê-se que este trabalho segue a abordagem onomasiológica, pois parte de conceitos para a identificação das diferentes expressões lingüísticas que o expressam[7].

Os conceitos selecionados para a investigação neste trabalho são aqueles classificados como **objetos (conceituais) concretos discretos**, os quais intuitivamente categorizam referentes perceptíveis pelos sentidos, localizados no tempo e no espaço, que são contáveis e indivisíveis (LYONS, 1977). A escolha dessa classe de conceitos justifica-se pelo fato de que eles, devido a sua natureza hierárquica, são passíveis de uma sistematização formal e, segundo Jackendoff (2002), são os conceitos que prototipicamente são categorizados como nominais. O domínio conceitual no qual tais conceitos foram delimitados é o dos "veículos com rodas" (no inglês, *wheeled vehicles*). A escolha desse domínio não se justifica por questões teóricas, mas práticas. No caso, a delimitação bem-definida e a extensão reduzida desse domínio facilitaram as tarefas de análise e representação léxico-conceitual.

Dessa forma, objetiva-se, ao final, produzir uma base de dados lexicais do domínio dos veículos sobre rodas em que os conceitos lexicalizados no AmE e no PB estejam alinhados por meio de uma interlíngua estruturada e formal.

Além disso, objetiva-se secundariamente identificar as relações interlinguais entre esses conceitos nos moldes da EuroWordNet. Essa identificação busca auxiliar a indexação das bases WN.Br e WN.Pr (DIAS-DA-SILVA et al, 2006; DI FELIPPO, DIAS-DA-SILVA, 2007). Mais especificamente, auxilia o alinhamento dos conceitos pertencentes ao domínio dos veículos com rodas.

---

[7] Devido à adoção da abordagem onomasiológica, as questões relacionadas à variação contextual do significado, como a polissemia e a homonímia, não são tratadas neste trabalho. Uma discussão sobre essas questões seria relevante se o trabalho seguisse a abordagem semasiológica, já que partiria das expressões lingüísticas, naturalmente ambíguas, para a identificação dos conceitos a elas subjacentes.

Com a construção da base bilíngüe e com as contribuições para o desenvolvimento da WN.Br, objetiva-se contribuir, ainda que modestamente, para o avanço do PLN no Brasil e, ideologicamente, para fortificar tanto (i) a ponte entre cientistas e engenheiros da linguagem quanto (ii) a concepção lingüisticamente motivada de PLN.

1.3. Metodologia

Para tanto, toma-se por base Dias-da-Silva (1996, 2006), que fornece os passos essenciais para o desenvolvimento de projetos na área do PLN. Em outras palavras, Dias-da-Silva fornece a concepção de PLN e o equacionamento metodológico que guiaram a realização deste trabalho.

Para Dias-da-Silva, os sistemas de PLN são vistos como "sistemas especialistas" (GRISHMAN, 1986) ou "sistemas baseados em conhecimento" (do inglês, *knowledge-based systems*). Segundo essa concepção, as pesquisas realizadas no PLN - como esta - envolvem uma "engenharia do conhecimento[8] lingüístico" e podem, conseqüentemente, beneficiar-se das estratégias desenvolvidas na área da Engenharia do Conhecimento, mais precisamente, dos modelos de representação do conhecimento (doravante, RC) (DIAS-DA-SILVA, 2006). Seguindo essa concepção, adotou-se o MultiNet (HELBIG, 2006), que se enquadra no paradigma dos modelos de representação do conhecimento baseados em redes semânticas. Neste trabalho, o MultiNet desempenhou duas funções essenciais: (i) a de suporte teórico-metodológico para a descrição do significado e (ii) a de formalismo para o registro da representação semântica. Os conceitos, representados segundo o MultiNet, serviram como interlíngua estruturada para o alinhamento dos conceitos lexicalizados no AmE e no PB.

Ainda segundo Dias-da-Silva (2006), ao conceber um sistema de PLN como um tipo especial de sistema especialista, as pesquisas no âmbito do PLN não só se beneficiam dos modelos de representação do conhecimento, mas também podem seguir as etapas descritas por Hayes-Roth (1990) para o desenvolvimento dos sistemas especialistas, são elas: "extração do solo" (isto é, explicitação dos conhecimentos e habilidades, "lapidação" (isto é, representação formal desses conhecimentos e habilidades) e "incrustação" (isto é, o programa de computador que codifica essa representação). Dias-da-Silva reinterpreta tais etapas e propõe que as pesquisas em PLN sejam divididas em três domínios, a saber: **lingüístico, lingüístico-computacional** e **computacional**. No domínio lingüístico, as atividades ficam concentradas na investigação dos fatos da língua natural em diferentes dimensões

[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 (NIRENBURG et al., 1992).

(morfológica, sintática, semântico-conceitual e até mesmo pragmático-discursiva) de acordo com a especificidade do sistema, ferramenta ou recurso que se queira desenvolver. No domínio lingüístico-computacional, por sua vez, estudam-se modelos formais de representação para os conhecimentos reunidos no domínio lingüístico que sejam computacionalmente tratáveis. E, por fim, no domínio computacional, as atividades ficam concentradas nas questões relativas à implementação do sistema de PLN, como: identificação dos módulos do sistema, do papel de cada um desses módulos, do fluxo de informações a ser gerenciado, da linguagem de programação a ser utilizada, entre outras. Neste trabalho, as atividades de pesquisa ficaram restritas aos domínios lingüístico e lingüístico-computacional.

As atividades relativas ao domínio lingüístico ficaram especialmente concentradas nas tarefas de: (i) delimitação do tipo conceitual e investigação teórica de suas características; (ii) delimitação do domínio conceitual; (iii) seleção do conjunto de conceitos lexicalizados no AmE pertencente ao domínio escolhido em (ii); (iv) delimitação dos conceitos selecionados em (iii); (v) identificação e aquisição das lexicalizações no PB dos conceitos delimitados em (iv); (vi) verificação da adequação das lexicalizações do PB; (vii) identificação das relações interlinguais nos moldes da EuroWordNet.

No domínio lingüístico-computacional, as atividades desenvolvidas foram: (i) seleção de um editor, ou seja, uma ferramenta computacional para auxiliar a tarefa de alinhamento dos conceitos lexicalizados; de um modo geral, esse editor é capaz de registrar a representação dos conceitos da interlíngua no formato *multinet* e suas respectivas lexicalizações no AmE e no PB; (ii) inserção das informações no editor e a subseqüente construção da base léxico-conceitual bilíngüe; (iii) proposição de novas funcionalidades para o editor da WN.Br que permitem o alinhamento das bases da WN.Pr e WN.Br; por meio da extensão do editor, o alinhamento das bases, até então manual, passará a seguir uma estratégia "assistida por computador" (do inglês, *computer-aided*), o que auxilia o trabalho do lingüista.

1.4. Estrutura do texto

Na Seção II, apresentam-se mais detalhadamente a base de dados mutilíngüe EuroWordNet e a concepção de interlíngua do sistema NADIA. Nessa descrição, focalizam-se as principais características do tipo de interlíngua utilizada nesses dois projetos.

Na Seção III, três questões principais são abordadas. As duas primeiras dizem respeito à definição de PLN como uma espécie de "engenharia do conhecimento lingüístico" e sua conseqüente relação com os modelos de representação do conhecimento. A terceira questão

diz respeito à concepção de representação do conhecimento como teoria e formalismo (metalinguagem ou linguagem formal) para a descrição da semântica das línguas naturais. Ainda sobre as representações do conhecimento, três paradigmas são descritos: o baseado na lógica, o baseado em *frames*[9] e o baseado em redes semânticas.

Na Seção IV, apresenta-se o MultiNet, paradigma de RC adotado neste trabalho. Mais especificamente, apresentam-se os recursos gerais que esse paradigma oferece para a representação dos conceitos.

Na Seção V, apresentam-se os meios representacionais do MultiNet específicos para a descrição dos objetos conceituais concretos discretos.

Na Seção VI, descrevem-se as várias atividades desenvolvidas no domínio lingüístico, focalizando a análise dos padrões de lexicalização, os vários tipos de fonte (lingüístico-computacionais, lexicográficas e textuais) e os critérios utilizados nas várias atividades relacionadas à referida análise.

Na Seção VII, descrevem-se as atividades desenvolvidas no domínio lingüístico-computacional. Em especial, descrevem-se as principais características do editor utilizado no alinhamento, as etapas de inserção dos dados no mesmo e a base bilíngüe resultante do trabalho. Descreve-se também a tarefa de identificação das relações interlinguais por meio da interface gráfica do editor e as extensões propostas especificamente para o editor da WN.Br.

Por fim, na Seção VIII, apresentam-se algumas considerações finais e etapas futuras deste trabalho.

[9] O termo *frame* em itálico indica a tecnologia de representação do conhecimento. Tal grafia tem a função de garantir a distinção entre a tecnologia representacional dos *frames* e o conceito lingüístico de "frame semântico", apesar de haver certa ligação entre eles (cf. Subseção 3.3.2.4, p. 55).

Seção II

Bases de Dados Lexicais Bilíngües e Multilíngües

2.1. A arquitetura de um sistema de PLN padrão

Teoricamente, as arquiteturas propostas para sistemas de PLN acabam por espelhar a arquitetura proposta para o sistema lingüístico (ALLEN, 1995). Como decorrência, um sistema de PLN deve possuir (i) módulos autômatos, que realizam tarefas específicas, e (ii) módulos que armazenam um modelo de conhecimento proposicional, os quais visam a criar simulacros de parcelas de mundo que lhe servem de referencial para interpretar os enunciados lingüísticos. Apesar de a arquitetura de um sistema de PLN variar de acordo com as especificidades da aplicação, dois grupos de componentes são imprescindíveis para a implementação de qualquer sistema desse tipo: as bases de conhecimento e os módulos de processamento que atuam sobre essas bases (DIAS-DA-SILVA, 1996). A Figura 4 ilustra esses dois grupos de componentes.

Os módulos de conhecimento podem ser divididos em três: o de análise, o especializado e o de síntese. As bases de conhecimento podem ser dividas em três bases: gramatical, conceitual e lexical. Com exceção do módulo especializado, os demais módulos de processamento e as bases de conhecimentos possuem estrutura e funcionamento semelhantes, embora os conteúdos possam variar em função da especificidade do sistema. Toda a especificação dos módulos descrita a seguir foi extraída de Dias-da-Silva (1996).

O **Módulo de Análise** (MA) é geralmente formado pelo analisador morfológico e pelo analisador sintático (ou *parser* sintático), além dos interpretadores semântico e pragmático-discursivo. Esse módulo é responsável pela construção de uma representação interna do significado das sentenças de entrada (no caso, digitadas via teclado).

O **Módulo de Síntese** (MS), por sua vez, transforma a representação abstrata gerada pelo MA em uma seqüência de "frases contextualizadas". Ao realizar a tarefa de construção de uma representação semântica, por exemplo, o MA utiliza-se, dependendo da sofisticação do sistema de que é parte, das bases gramatical, conceitual e lexical para executar todas ou parte das análises: morfológica, sintática, semântica e, até mesmo, pragmática. Assim, cada base de

conhecimento, por sua vez, fornece ao MA informações de natureza diferente (cf. HUTCHINS, SOMERS, 1997).

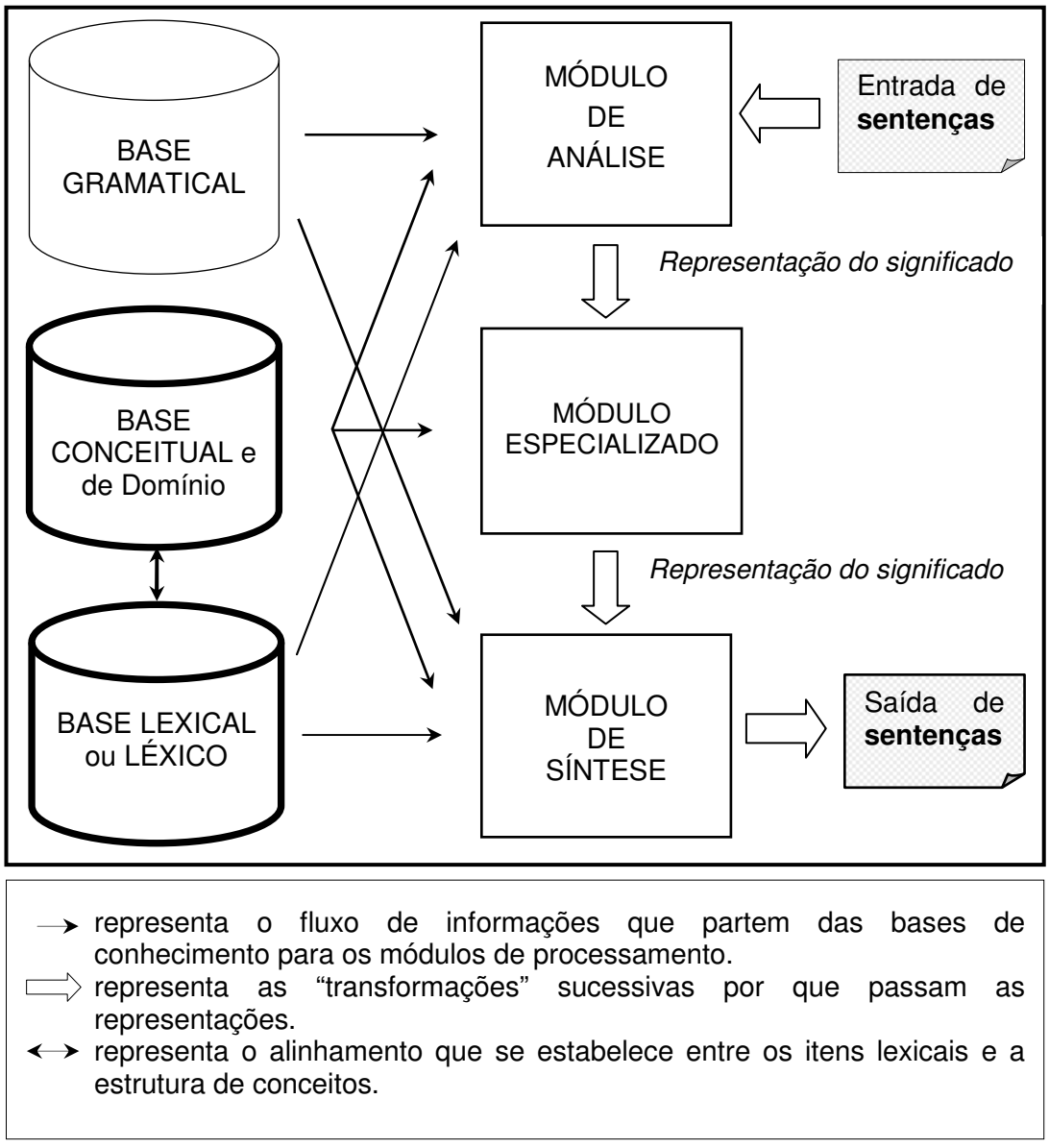


Figura 4: A arquitetura padrão de um sistema de PLN de Dias-da-Silva (1996).

A **Base Gramatical** fornece a representação das regras sintáticas da língua, que podem ser vistas como condições de admissibilidade de estruturas sintáticas bem-formadas; condições que servirão de referência para o módulo de análise – responsável pela construção das representações sintáticas, semânticas e pragmático-discursivas.

A **Base Conceitual e de domínio** fornece um modelo do mundo físico e conceitual, descrevendo tipos básicos de objetos, eventos, propriedades, relações e atributos em termos de representações hierarquicamente estruturadas, isto é, a sua estrutura consiste em uma rede

de unidades conceituais interligadas em termos de relações de hiponímia/ hiperonímia, entre outras. Essa base também pode fornecer conceitos mais específicos, ou seja, conceitos referentes a domínios particulares do conhecimento ou conceitos relacionados a tarefas específicas para a qual o módulo esteja sendo projetado.

Em particular, à **Base Lexical**, cabe a tarefa de fornecer, aos MA e MS, a coleção de unidades lexicais, para as quais se faz necessária a especificação de conjuntos de traços morfológicos, sintáticos, semânticos e pragmático-discursivos (BOGURAEV, BRISCOE; 1989; BRISCOE, 1991; SANFILIPPO, 1995; PALMER, 2001). Esse tipo de base de dados, no domínio do PLN, é definido como o **léxico** do sistema (HANDKE, 1995; GRISHMAN, CALZOLARI, 1997). Tais léxicos também são denominados **dicionários tratáveis por máquina** (do inglês, *machine tractable dictionaries*) (WILKS et al., 1996).

Sabe-se que a construção desse tipo de recurso é uma tarefa bastante custosa, principalmente devido à quantidade e especificadade das informações a serem descritas e representadas. Diante disso, surgem, pautadas na noção de "reutilização"[10], as chamadas **bases de dados lexicais** (BDLs) (do inglês, *lexical databases*). As BDLs são entendidas como grandes repositórios de informação lexical, um "mega arquivo". Nas BDLs, a informação lexical está armazenada de forma estruturada, podendo fornecer diferentes tipos de informação a diferentes sistemas de processamento automático de língua natural.

Neste trabalho, focaliza-se um tipo especial de BDL, em que as unidades lexicais de diferentes línguas estão indexadas por meio do significado que elas expressam. Mais especificamente, focalizam-se aqui as BDLs construídas com base em um alinhamento por interlíngua. Grosso modo, uma interlíngua pode ser entendida como um conjunto de significados aos quais as unidades lexicais que os expressam em diferentes línguas são alinhadas (NIRENBURG, 1989; COPELAND et al., 1991; JANSSEN, 2004).

2.2. As bases de dados multilíngües: dois exemplos

A seguir, focalizando-se as principais características do tipo de interlíngua utilizada, apresentam-se a base multilíngüe EuroWordNet (VOSSEN, 1998), desenvolvida para línguas européias e baseada no formato *wordnet*, e a interlíngua do NADIA, sistema de gerenciamento de bases de dados lexicais multilíngües (SÉRASSET, 1994a,b).

[10] Na Ciência Computacional, utiliza-se freqüentemente o termo "reusabilidade" (do inglês, *reusability*).

2.2.1. A base EuroWordNet

Na EuroWordNet, as bases lexicais em formato *wordnet* desenvolvidas para o inglês (britânico), holandês, espanhol, italiano, alemão, francês, tcheco e estônio[11] estão interligadas por meio de uma coleção de conceitos inicialmente oriundos da WN.Pr (1.5). Antes de se apresentar a estratégia de alinhamento adotada nessa base, apresenta-se o referido formato *wordnet*. Esse formato de BDL provém da base desenvolvida para o AmE, a WN.Pr (FELLBAUM, 1998a,b). A seguir, descrevem-se as principais características da WN.Pr, posto que algumas delas são revelantes para a compreensão da arquitetura da EuroWordNet.

2.2.1.1. A WordNet de Princeton

A WN.Pr é uma rede em que as unidades lexicais, pertencentes às categorias dos substantivos, verbos, adjetivos e advérbios, organizam-se sob a forma de ***synsets***[12] (do inglês, *synonym set*). Em outras palavras, pode-se dizer que o *synset* é um conjunto de unidades lexicais de uma mesma categoria sintática que podem ser intercambiáveis em um determinado contexto, p.ex.: {bicycle, bike, wheel, cycle}. O *synset*, por definição, é construído de modo a codificar um único conceito lexicalizado por suas unidades constituintes. Vale ressaltar, no entanto, que a WN.Pr também armazena conceitos não-lexicalizados no AmE, ou seja, conceitos para os quais não há uma expressão lexical (isto é, expressão que se espera encontrar como entrada em um dicionário monolíngüe). Incluem-se nesse grupo os codificados pelos *"synsets"* {natural object}, {external body parts}, etc. A principal razão da inclusão desses conceitos é auxiliar a organização da hierarquia conceitual (VOSSEN, 1998).

O emprego do *synset* como construto representacional pressupõe que o falante tem acesso aos conceitos expressos pelos itens lexicais de sua língua. A WN.Pr adota a noção de **sinonímia contextual** para a montagem de *synsets*. De acordo com essa noção de sinonímia, "duas unidades lexicais são sinônimas em um contexto C, se a substituição de uma pela outra em C não altera o valor de verdade denotado por C" (ULLMANN, 1963; LYONS, 1981; MILLER; FELLBAUM, 1991). A sinonímia contextual contrapõe-se à **sinonímia absoluta**, segundo a qual "duas unidades lexicais são totalmente sinônimas quando são substituíveis, uma pela outra, em todos os contextos, sem que haja mudança do valor de verdade da proposição expressa pelas sentenças em que as substituições são feitas" (ULLMANN, 1963;

[11] A extensão da EuroWordNet para outras línguas vem se consolidando com a integração de línguas como português europeu, sueco, basco, catalão, russo, grego e dinamarquês (VOSSEN et al., 1999).

[12] 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.

LYONS, 1981; MILLER; FELLBAUM, 1991). A sinonímia total é raramente encontrada na língua geral, sendo mais comum nas línguas de especialidades.

Assim, se o falante não conhece o significado de uma determinada forma lexical, uma forma sinônima é suficiente para que ele identifique o conceito apropriado. Por exemplo, se o falante desconhece a forma x e essa forma é parte do *synset* y e o falante conhece as formas z e k desse *synset*, então, porque a forma desconhecida x é parte de y, o falante passa a ter acesso ao significado da forma x. Por exemplo, se o falante não conhece o significado da forma lexical *abafo*, exemplificada em (1), ele pode acessar esse significado a partir do *synset* {abafo, agasalho}, se ele conhecer o significado da forma *agasalho*.

(1) O capote, sendo um abafo, foi concebido para proteger o seu utilizador dos rigores do frio da planície.

Os *synsets*, então, são construídos a partir da possibilidade de intersubstituição de itens lexicais em contextos mínimos. Por exemplo, verbo *conduzir* pode ser substituído, em (2a), por *comandar* e, em (2b), por *levar*. A relação de sinonímia contextual estabelecida entre *conduzir* e *comandar* em (2a) e entre *conduzir* e *levar* em (2b) autoriza a criação dos dois *synsets*, exemplificados em (3) (MORAES, 2008).

(2) a. O presidente Fernando Henrique vai precisar, além do talento, de muita sorte para conduzir a política econômica, este ano.
b. O assalto ao prospector bancário começou na manhã do crime, quando um dos indivíduos conduziu outros dois até junto ao restaurante do Galanta.

(3) a. {conduzir; comandar}
b. {conduzir; levar}

Entre os *synsets*, codificam-se cinco principais relações lógico-conceituais: antonímia, hiponímia, meronímia, acarretamento e causa (LYONS, 1979; CRUSE, 1986; FELLBAUM, 1998a,b):

a) **Hiperonímia/ Hiponímia:** relação entre um conceito mais generalizante (o hiperônimo) e um conceito mais específico (o hipônimo). Um item lexical x é hipônimo de outro item lexical y se o falante aceita frases construídas a partir da seguinte fórmula: um x é um (tipo de) y. Por exemplo, a aceitação das frases *a limusine é um tipo de carro* e *um carro é*

um tipo de veículo identifica o possível *synset* {limusine} como hipônimo do *synset* {carro} e {carro} como hipônimo de {veículo}.

b) **Antonímia**: relação que engloba diferentes tipos de oposição semântica. São elas: *antonímia complementar*: relaciona pares de itens lexicais contraditórios em que a afirmação do primeiro acarreta a negação do segundo e vice-versa, por exemplo: {vivo} e {morto}; *antonímia gradual*, que relaciona itens lexicais que denotam valores opostos em uma escala como, por exemplo, {pequeno} e {grande}; e "antonímia recíproca", que relaciona pares de itens lexicais que se pressupõem mutuamente, sendo que a ocorrência do primeiro pressupõe a ocorrência do segundo como, por exemplo, {comprar} e {vender}.

c) **Meronímia/ Holonímia:** relação entre um *synset* que expressa um "todo", o holônimo, por exemplo, o *synset* hipotético {carro}, e outros *synsets* que expressam partes do todo, os merônimos, por exemplo: {pára-choque}, {pneu}, {direção}, {câmbio}, etc.

d) **Acarretamento:** relação que se estabelece entre uma ação A1 e uma ação A2; a ação A1 denotada pelo verbo x acarreta a ação A2 denotada pelo verbo y se A1 não puder ser feita sem que A2 também o seja. Esse é o caso, por exemplo, da relação entre os verbos *correr* e *deslocar-se*, já que a ação de correr (A1) acarreta a ação de deslocar-se (A2); assim, estabalece-se a relação de acarretamento entre os possíveis *synsets* {correr} e {deslocar-se}. Vale salientar que o acarretamento é uma relação unilateral, isto é, por um lado *correr* acarreta *deslocar-se,* mas, por outro, o inverso não ocorre, *deslocar-se* não necessariamente acarreta *correr*.

e) **Causa:** relação que se estabelece entre uma ação A1 e uma ação A2 quando a ação A1 denotada pelo verbo x causa a ação A2 denotada pelo verbo y. Esse é o caso, por exemplo, da relação que se estabelece entre a ação denotada por *matar* e a ação denotada pelo verbo *morrer*.

Na Figura 5, cujo exemplo foi extraído da WN.Pr (version 2.1), exemplificam-se dois tipos de relação: a hiperonímia e a meronímia. Vê-se nessa Figura que o *synset* {car; auto; automobile; machine; motorcar} está relacionado, por exemplo, a:

a) o conceito mais geral ou *synset* hiperônimo: {motor vehicle; automotive vehicle};
b) os conceitos mais específicos ou *synsets* hipônimos, p.ex.: {bus; jalopy; heap } e {cab; taxi; hack; taxicab};

c) os conceitos que indicam partes ou *synsets* merônimos, p.ex.: {bumper}, {car door}, {car mirror} e {car window}.

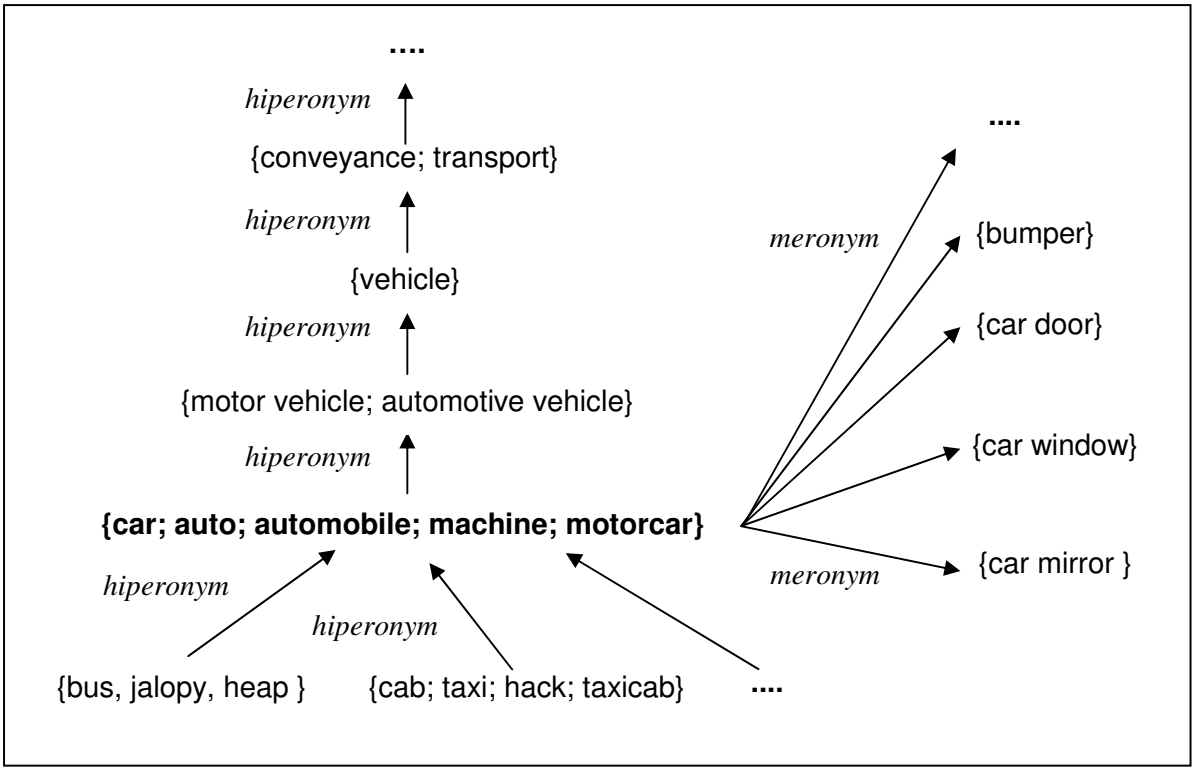


Figura 5: Amostra da organização dos *synsets* na WN.Pr.

Observa-se ainda que cada *synset* relaciona-se novamente a outros *synsets*, por exemplo, o *synset* {motor vehicle; automotive vehicle} está relacionado à {vehicle} e {conveyance; transport}.

A WN.Pr armazena ainda uma série de informações associadas a cada *synset*:

a) um número que identifica o *synset*; por exemplo, para {bicycle; bike; wheel; cycle}, tem-se o número 02834778;
b) o tipo semântico do conceito representado no *synset*; p.ex.: o *synset* {bicycle; bike; wheel; cycle} é do tipo semântico <noun.artifact>;
c) uma glosa, isto é, uma definição informal do conceito representado no *synset*; p.ex.: <a wheeled vehicle that has two wheels and is moved by foot pedals>;
d) frases-exemplo extraídos de *corpora*[13]*;*
e) um conjunto de indexadores (do inglês, *pointers*), que estabelecem as relações semântico-conceituais entre os *synsets*.

[13] No caso do *synset* {bicycle; bike; wheel; cycle}, usado como exemplo, ressalta-se que não há frases-exemplo armazenadas na WN.Pr para as unidades que o constituem.

A Figura 6 ilustra essas informações.

Tipos de informação	**Valores**
Número de identificação do *synset*	(01941830}
Tipo semântico do conceito expresso pelo *synset*	<verb.motion>
Lista de itens lexicais que constituem o *synset*	{lead; take; conduct; guide}
Glosa	take somebody somewhere (“levar alguém a algum lugar”)
Frases-exemplo para cada item constitutivo	We lead him to our chief. (“Nós o levamos até nosso chefe”) Can you take me to the main entrance? (“Você pode me levar até a entrada principal?”) He conducted us to the palace. (“Ele nos conduziu ao palácio”)

Figura 6: Informações associadas a um *synset*.

2.2.1.2. A WN.Pr como uma ontologia lingüística

A WN.Pr tem sido considerada uma **ontologia lingüística**. Ontologias lingüísticas são consideradas recursos de larga escala e definem-se, por um lado, como estoques de unidades lexicais de uma língua e, por outro, como estruturas ontológicas ou inventários dos conceitos (lexicalizados) compartilhados por uma comunidade lingüística. Com base nessa definição, pode-se dizer que a WN.Pr é, ao mesmo tempo, uma Base Lexical e uma Base Conceitual da arquitetura apresentada na Figura 4.

Diz-se que as ontologias lingüísticas são um tipo especial de ontologia porque elas armazenam conceitos lexicalizados (em dada língua) e não são objetos formais (MAGNINI, SPERANZA, 2002).

Esse tipo de ontologia opõe-se às chamadas **ontologias formais**, em que os conceitos e relações entre conceitos estão explicitamente descritos por meio de um formalismo ou modelo de representação formal (NIRENBURG, RASKIN, 2004) e extrapolam a representação de conceitos lexicalizados.

A WN.Pr, enquanto ontologia lingüística, tem sido utilizada em várias aplicações de PLN como recuperação e extração de informação, desambiguação lexical de sentido (do inglês, *word sense disambiguation*), categorização e estruturação de documentos, entre outras (MORATO et al, 2004; DI FELIPPO, 2006).

2.2.1.3. Características da EuroWordNet

Na EuroWordNet, a interligação das *wordnets* é feita por meio da indexação dos *synsets* de cada uma delas ao Índice Inter-Lingual (ILI) (do inglês, *Inter-Lingual-Index*), que funciona como uma espécie de interlíngua. Por questões pragmáticas, como manutenção, expansão e reutilização da base de dados léxico-conceitual, o ILI é uma interlíngua **não-estruturada** de conceitos, a qual foi identificada a partir de *synsets* da WN.Pr[14]. Os conceitos que compõem o ILI são chamados registros ILI. Cada registro ILI é, inicialmente, composto de três informações retiradas dos *synsets* da WN.Pr:

(i) o número de identificação do *synset*;

(ii) a lista de itens lexicais;

(iii) a glosa.

Além dessas informações, os registros ILI estão associados a duas ontologias: a Ontologia de Conceitos Gerais (do inglês, *Top-concept Ontology*) e a Ontologia de Rótulos de Domínio (do inglês, *Domain Label Ontology*).

A primeira é independente de língua e formada por uma coleção de 1.024 registros ILI que representam os conceitos básicos lexicalizados pelas *wordnets* individuais. Essa ontologia segue os tipos de entidades caracterizados em Lyons (1977). As entidades de 1ª ordem denotam entidades concretas e são canonicamente expressas por nomes concretos. Tais entidades são subclassificadas em 33 subtipos, que incluem, por exemplo: origem, forma, composição, substância, entre outros. As entidades de 2ª ordem denotam propriedades, ações, processos, relações e eventos e são canonicamente expressas por verbos, nomes e adjetivos. As entidades de 2ª ordem estão classificadas em 30 subtipos, que incluem, por exemplo: propriedade, relação, evento delimitado, etc. Por fim, as entidades de 3ª ordem denotam proposições e são canonicamente expressas por nomes abstratos e frases. A Ontologia de Conceitos Gerais reflete, então, distinções semânticas importantes, p.ex.: objeto e substância, lugar, dinâmico e estático, etc.

A segunda ontologia, a de Rótulos de Domínio, pode ser vista como uma hierarquia de rótulos que indicam tópicos ou *scripts* (cf. Subseção 3.3.2.2, p. 50), p.ex.: esporte, hospital, restaurante, tráfego, etc.

---

[14] A coleção de ILIs era inicialmente composta por todos os *synsets* da WN.Pr 1.5. No entanto, ao longo do processo de indexação das *wordnets* européias à base de ILIs, tornou-se necessária a ampliação dessa coleção inicial com a adição de índices que representam lexicalizações de conceitos específicas de uma língua e que não estavam contempladas no inventário inicial. Dessa forma, a coleção de ILIs passou a ser o "superconjunto" de todos os conceitos lexicalizados nas diferentes *wordnets* européias (PETERS et al., 1998).

A Figura 7, elaborada com base em Vossen (1998), ilustra a arquitetura da base EuroWordNet[15].

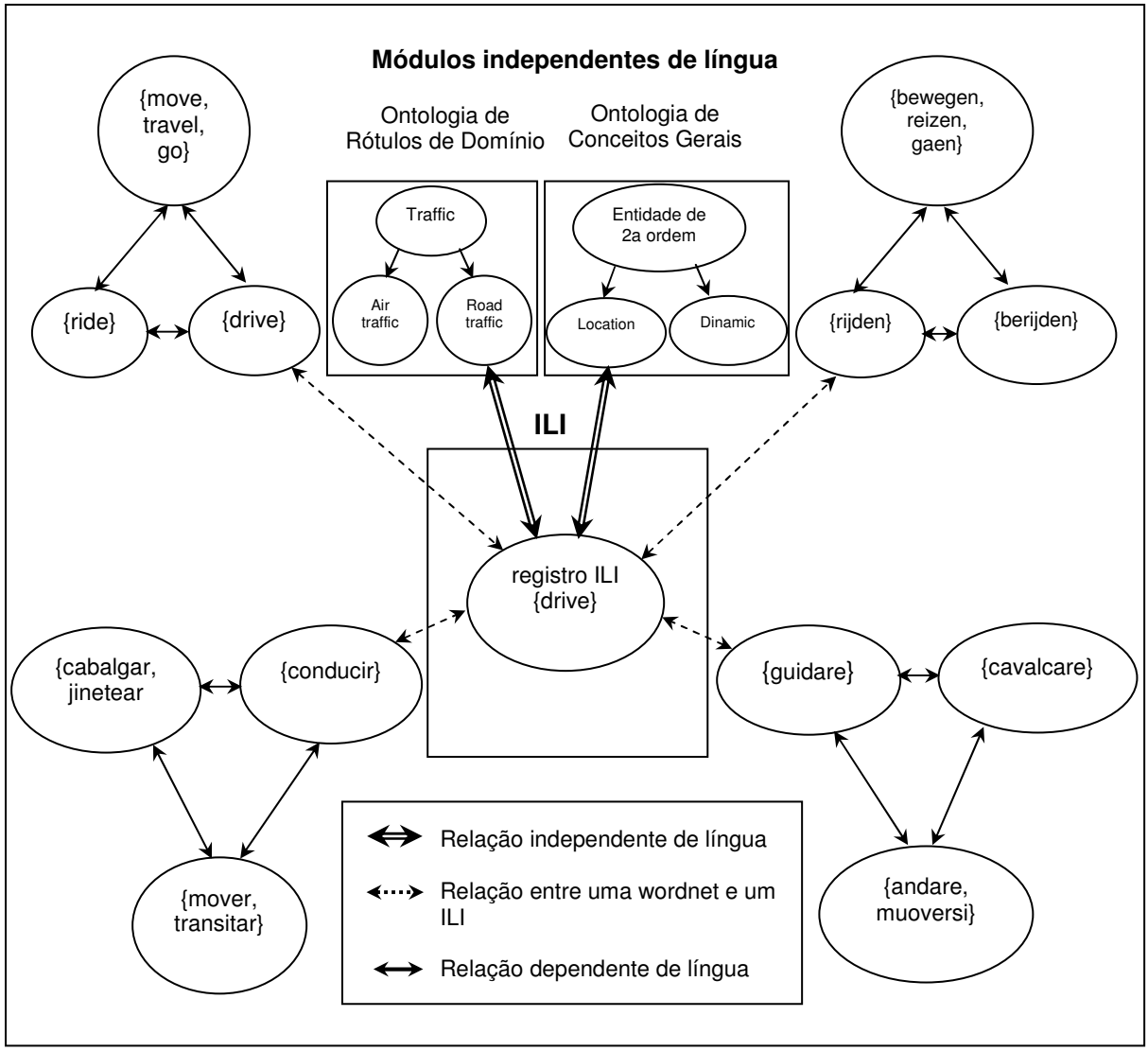


Figura 7: A arquitetura da base multilíngüe EuroWordNet.

Nessa base, devido ao fato de o ILI ser uma interlíngua não estruturada, a ligação dos *synsets* de *wordnets* de línguas diferentes aos registros ILI é feita por meio de um conjunto de relações interlinguais de equivalência, propostas à semelhança das relações conceituais internas às *wordnets*. Dentre tais relações, destacam-se quatro principais: EQ_SYNONYM, EQ_NEAR_SYNONYM, EQ_HAS_HYPERONYM e EQ_HAS_HYPONYM (VOSSEN, 1998; PETERS et al., 1998). Quando um *synset* de uma *wordnet* específica está indexado a um ILI por meio da relação EQ_SYNONYM, diz-se que tal *synset* lexicaliza o conceito representado pelo ILI.

[15] Por motivo de simplificação, os ILIs, na Figura 7, estão descritos apenas por meio de um rótulo simples.

Na Figura 8[16], por exemplo, vê-se que o conceito identificado pelo ILI {violoncello; cello}/ "a large stringed instrument; a seated player holds it upright while playing" é lexicalizado, no holandês, pelas formas {violoncel; cel; cello}; no espanhol, pelas formas {chelo; violoncelista; violoncelo; violonchelo; cello}; no italiano, pela forma {violoncello} e, no inglês, pelas formas {violoncello; cello} (ALONGE et al, 1998).

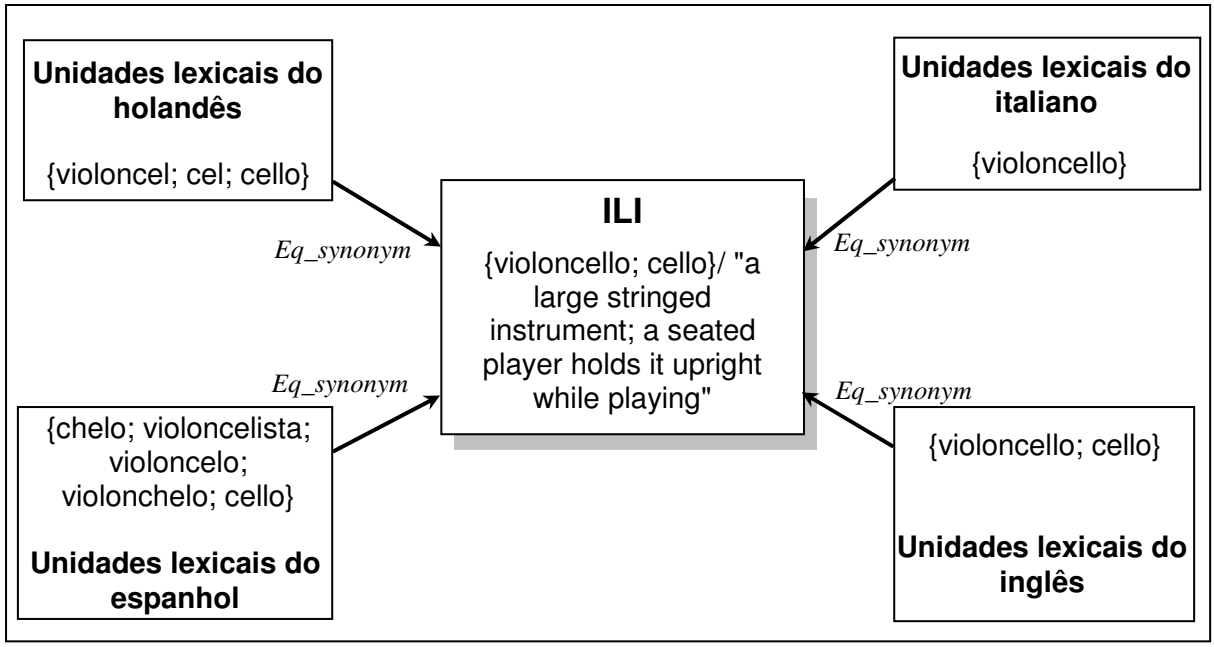


Figura 8: Um exemplo de alinhamento na EuroWordNet por EQ_SYNONYM.

As demais relações são utilizadas exatamente quando há divergências léxico-conceituais entre os ILIs e as *wordnets* envolvidas no projeto. A relação EQ_NEAR_SYNONYM rotula divergências de granularidade entre um conceito lexicalizado em uma língua x e um ILI. A relação EQ_HYPERONYM rotula equivalências em que um *synset* lexicaliza um conceito mais específico que o conceito representado por um ILI e a relação EQ_HAS_HYPONYM rotula relações de equivalência em que o *synset* lexicaliza um conceito mais genérico que o conceito representado por um ILI. O emprego de tais relações resulta em uma arquitetura elegante para a base multilíngüe. Na Figura 9, elaborada com base em Peters et al. (1998), ilustra-se o emprego das relações EQ_SYNONYM, EQ_HAS_HYPERONYM e EQ_HAS_HYPONYM.

[16] Por motivo de simplificação, os ILIs estão descritos apenas por meio da lista de unidades lexicais e da glosa.

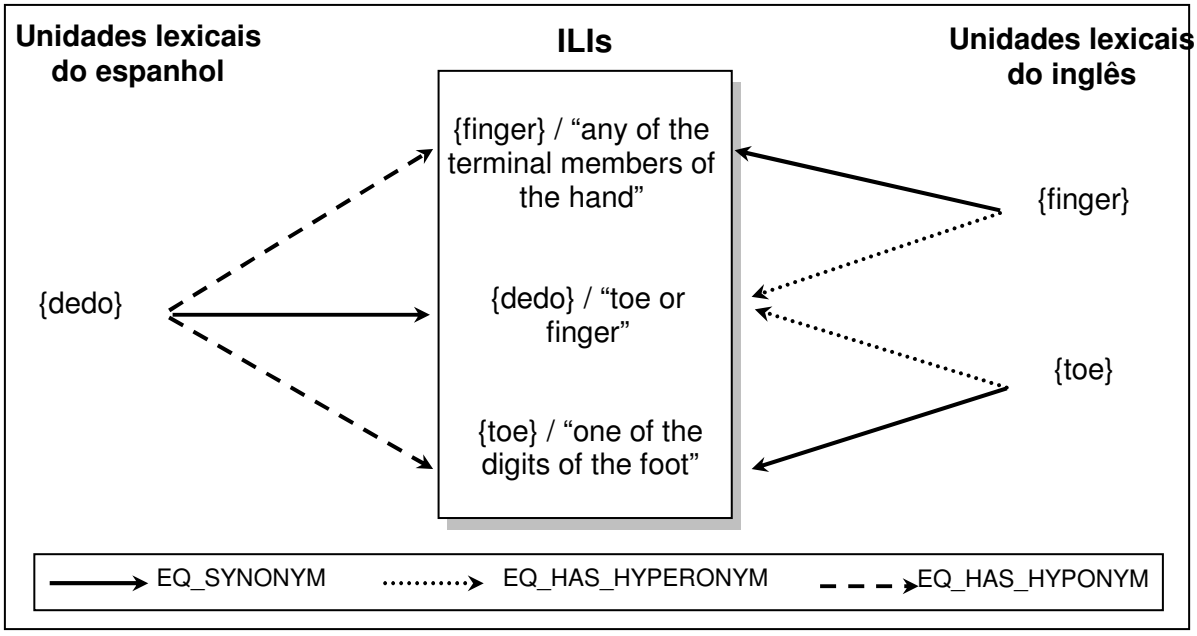


Figura 9: Diferentes tipos de alinhamento da EuroWordNet.

Na Figura 9, três ILIs compõem a interlíngua: {finger}/ "any of the terminal members of the hand" e {toe}/ "one of the digits of the foot". {dedo}, provenientes da WN.Pr, e "toe or finger", provenientes da *wordnet* do espanhol.

No espanhol, o *synset* {dedo} lexicaliza o conceito expresso pelo ILI {dedo}/ "toe or finger" e, por isso, está vinculado a esse ILI por meio da relação EQ_SYNONYM. Devido ao fato de o espanhol não lexicalizar os conceitos expressos pelos ILIs {finger} / 'any of the terminal members of the hand" e {toe} / "one of the digits of the foot", o *synset* {dedo} está relacionado a esses dois ILIs por meio da relação EQ_HAS_HYPONYM. Quanto ao AmE, vê-se que os *synsets* {finger} e {toe} estão indexados, respectivamente, aos ILIs {finger} / 'any of the terminal members of the hand" e {toe} / "one of the digits of the foot" por meio da relação EQ_SYNONYM, indicando, assim, que as unidades *finger* e *toe* são lexicalizações de tais conceitos. Como o AmE não lexicaliza o conceito expresso pelo ILI {dedo}/ "toe or finger", os *synsets* {finger} e {toe} estão ligados a esse ILI por meio da relação EQ_HAS_HYPERONYM.

Apesar da elegância, a arquitetura de indexação empregada no projeto EuroWordNet apresenta vantagens e desvantagens, como já mencionado na Seção I.

A principal vantagem do método de alinhamento baseado em interlíngua não-estruturada reside na facilidade de manutenção e expansão da base de dados léxico-conceitual, posto que a inserção de conceitos específicos de uma nova língua à interlíngua é tarefa relativamente simples.

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>, cujo *synset* {dedo} está relacionado a três elementos distintos da interlíngua, como ilustrado na Figura 9. Com o acréscimo de novas línguas à base, o número de *links* pode crescer consideravelmente.

Outra limitação da interlíngua empregada na EuroWordNet surge da falta de uma metalinguagem semântica formal para a descrição dos elementos/ conceitos que compõem a interlíngua. Um elemento que compõe o ILI é identificado em função de três tipos de informação: (i) o *synset*, (ii) o número de identificação do *synset* e (iii) a glosa. Dessa forma, tal interlíngua não emprega uma linguagem semântica formal (também entendida como formalismo semântico). Para o PLN, no entanto, é relevante que o conhecimento semântico-conceitual contido em uma BDL seja o mais explícito possível, o que o torna manipulável pelo sistema de PLN do qual a base fará parte. Vale ressaltar, no entanto, que, apesar de seu "baixo nível" de formalização, a EuroWordNet e, principalmente a WN.Pr, têm sido amplamente utilizadas no PLN.

2.2.2. A interlíngua do sistema NADIA

Nesta Subseção, apresenta-se a concepção de interlíngua do sistema de gerenciamento de BDLs multilíngües denominado NADIA (SÉRASSET, 1994a,b).

Nesse sistema, para a construção de uma BDL multilíngüe, o método utilizado para o alinhamento das unidades lexicais baseia-se em uma interlíngua estruturada, denominada **dicionário de acepções** (do inglês, *acceptions dictionary*).

Os elementos da interlíngua são chamados **acepções** (do inglês, *acceptions*), sendo provenientes das diferentes línguas envolvidas no processo de alinhamento. Ou melhor, dos léxicos monolíngües ou **dicionários** dessas língua. Dessa forma, pode-se dizer que um ILI da EuroWordNet equivale a uma acepção do dicionário de acepções do NADIA.

Quando duas ou mais línguas lexicalizam um mesmo conceito, ou melhor, quando duas ou mais línguas possuem unidades lexicais que expressam a mesma acepção, o alinhamento no NADIA é feito de modo similar ao da EuroWordNet (cf. Figura 8, p. 21). A principal diferença entre os métodos de alinhamento utilizados na EuroWordNet e no NADIA diz respeito aos casos em que há divergências léxico-conceituais entre as línguas que fazem parte da BDL multilíngüe. No caso da EuroWordNet, como visto na Seção anterior, as tais divergências podem causar mais de um alinhamento ou *link* entre um único conceito

lexicalizado em uma determinada língua e a interlíngua. No caso do NADIA, isso não acontece, posto que a interlíngua é de certa forma estruturada (Figura 10).

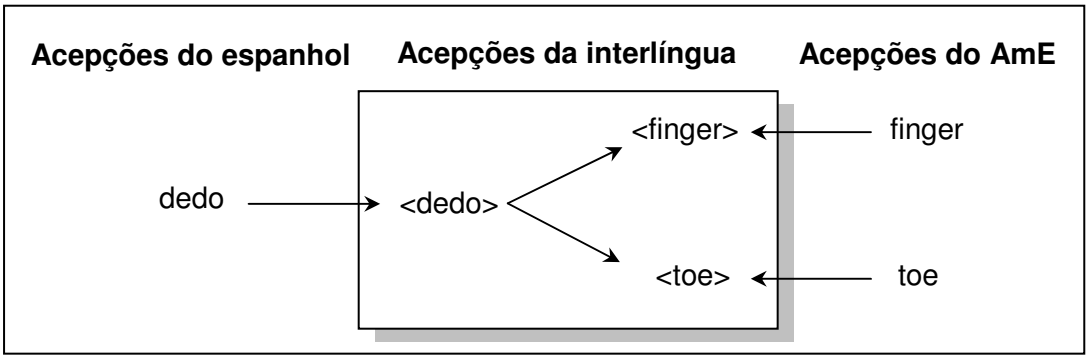


Figura 10: A interlíngua estruturada do sistema NADIA.

Diz-se "de certa forma" porque a estruturação dos elementos da interlíngua só é estabelecida quando divergências léxico-conceituais são encontradas. Quando há uma correspondência direta entre duas acepções provenientes de línguas distintas, estas são fundidas em uma única acepção da interlíngua. Quando não há uma correspondência direta, as acepções de uma língua são mantidas no dicionário de acepções ou interlíngua. Por exemplo, considera-se que no dicionário monolíngüe do AmE, a entrada de *river* tem duas acepções, as quais estão ilustradas na Figura 11, elaborada com base em Sérasset (1994b).

Dicionário do inglês	
Entrada	**Acepções**
river	river_1: natural stream of water flowing in a channel to the sea or to the lake, etc. or joining another~.
	river_2: great flow, a ~*of lava; ~s of blood,* great bloodshed (in war)

Figura 11: Acepções de *river.*

Suponha-se agora que, para *rivière* and *fleuve*, haja as seguintes acepções, as quais estão ilustradas na Figura 12, também elaborada com base em Sérasset (1994b).

Dicionário do francês	
Entradas	**Acepções**
rivière	rivière_1: cours d'eau naturel de moyenne importance
	rivière_2: *Par anal. Sport.* Fosse rempli d'eau que doit sauter le cheval (steeple-chase) ou le coureur (steeple)
	rivière_3: *Fig.* Flots, ruisseau. *"dês rivières de sang"*
	rivière_4: *Fig.* rivière de diamants: collier de diamants montés en chatons
fleuve	fleuve_1: Grande rivière (remarquable par le nombre de sés affluents, l'importance de son débit, la longueur de son cours), spécialement lorsqu'elle aboutit à la mer
	fleuve_1: *Littér.* Ce qui coule, ce qui est répandu en abondance

Figura 12: Acepções de *rivière* e *fleuve*.

Como a unidade do AmE *river* pode ser traduzida para *rivière* ou *fleuve*, o dicionário de acepções ou interlíngua deve explicitar as acepções rivière_1 e fleuve_1 como subacepções da acepção river_1, resultando, assim, no alinhamento ilustrado na Figura 13.

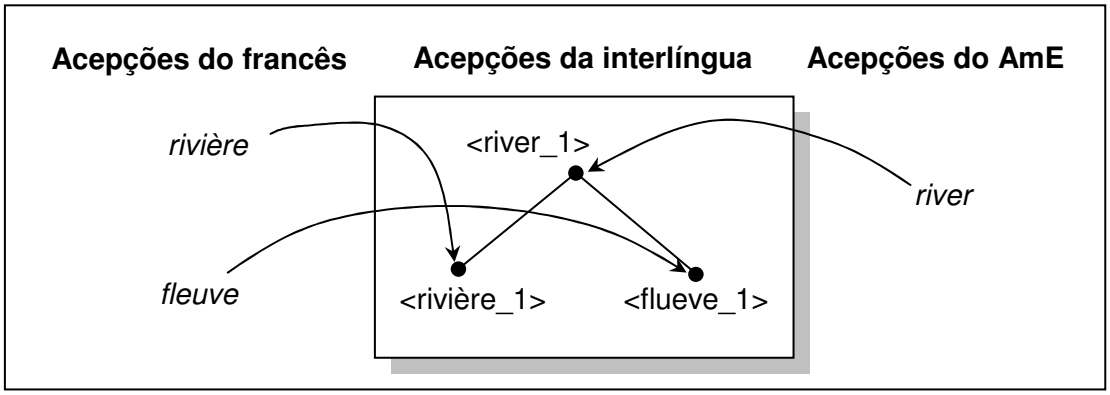


Figura 13: Interlíngua e alinhamentos do NADIA.

Vale salientar mais uma vez que as relações entre as acepções da interlíngua são introduzidas apenas quando há a necessidade de se preencher uma lacuna lexical. Assim, a interlíngua de acepções, como um todo, não é estruturada; somente as acepções que participam da resolução de uma lacuna lexical são estruturadas. Dessa forma, a questão relacionada ao número de *links* parece ser solucionada com a adoção do método baseado em interlíngua estruturada.

2.3. Síntese da Seção II

Nesta Seção, foram enfatizados os seguintes tópicos:

a) a definição e o papel de um léxico e de uma BDL no âmbito do PLN;
b) as BDLs multilíngües em que se adota o método de indexação baseado em interlíngua não-estruturada e as em que se adota o método baseado em interlíngua estuturada;
c) as vantagens e desvantagens de cada método de alinhamento;

A partir da discussão desses tópicos, opta-se pela utilização do método baseado em uma interlígua estruturada para alinhar os conceitos lexicalizados no AmE e no PB, apesar de esse método dificultar a tarefa de inclusão de novos elementos à interlíngua. Opta-se por tal método para evitar o número excessivo de *links* entre as línguas e a interlíngua. Outra razão para a escolha de uma interlíngua estruturada é a possibilidade de contrastá-la com o método empregado no projeto EuroWordNet e que está sendo utilizado para o alinhamento das bases da WN.Pr e WN.Br.

Como mencionado da Seção I, busca-se, ao final, construir uma BDL bilíngüe em que os conceitos lexicalizados nas referidas línguas estejam alinhados por meio de uma interlíngua estruturada, e os elementos constitutivos dessa interlíngua (os conceitos) estejam representados por um formalismo semântico.

Para alcançar os objetivos ora propostos, parte-se de uma concepção específica de PLN, a qual é apresentada da próxima Seção.

Seção III

O Processamento Automático das Línguas Naturais

3.1. Visão geral

O surgimento

Conforme observa Dias-da-Silva (1996), desde que os computadores foram introduzidos em nossa cultura, na década de 40, fazê-los "entender" instruções necessárias para a execução de tarefas tem sido um desafio para os "engenheiros da linguagem". A primeira solução encontrada foi a criação das "linguagens de programação" (do inglês, *programming languages*). Com o tempo, linguagens cada vez mais inteligíveis foram criadas, como LISP, PROLOG, etc. Mesmo assim, instruções nessas linguagens são inevitavelmente rígidas, pois precisam ser descritas exatamente como o previsto.

Com a introdução dos primeiros computadores pessoais, que começaram a fazer sua história na década de 70, a questão da comunicação entre o homem e a máquina ganhou ainda mais importância. Desenvolveu-se, como conseqüência, o conceito *user-friendly* – ou seja, "amigável" ou "fácil de aprender e usar" (MICROSOFT PRESS, 1998, p. 742). Esse conceito revelava a preocupação dos engenheiros da linguagem em fazer dos computadores instrumentos cada vez mais amigáveis, já que eles passavam a ser utilizados por pessoas comuns, isto é, não-especialistas. Mais especificamente, esses engenheiros buscavam tornar a comunicação entre o homem e a máquina mais natural e intuitiva, pois, a partir do momento em que a maioria dos usuários definitivamente deixava de ser especialista em Informática, os problemas de comunicação e de significação se tornavam mais importantes.

Assim, os engenheiros da linguagem passaram a pensar em possíveis linguagens que pudessem intermediar uma comunicação mais amistosa entre os computadores e seus usuários comuns. Uma das soluções encontradas (e que atualmente está presente em todos os computadores), pautada na utilização da linguagem visual, foram as chamadas "interfaces gráficas com o usuário" (do inglês, *graphical user interfaces* - GUIs), ou apenas, "interfaces gráficas". Nessa linguagem icônica, programas, arquivos e opções são representados por meio

de imagens e objetos gráficos como menus, janelas, caixas de diálogos, etc. O usuário pode selecionar e ativar essas opções com o *mouse* ou, em geral, através do teclado (MICROSOFT PRESS, 1998, p. 386). Outra possibilidade seria a utilização/ adaptação da linguagem humana, ou seja, a criação de programas que pudessem, ainda que de modo rudimentar, emular o conhecimento e o desempenho lingüísticos humanos. Em outras palavras, ensinar o computador a falar a língua dos homens[17].

Segundo Dias-da-Silva (2006), a possibilidade de interação homem/máquina por meio da língua dos homens e o surgimento dos primeiros sistemas de tradução automática impulsionaram os estudos ou investigações que receberam o nome **Processamento Automático de Línguas Naturais** (do inglês, *Automatic Natural Language Processing* ou *Natural Language Processing*).

O lugar

De modo geral, no PLN, buscam-se soluções para questões computacionais que requerem o tratamento computacional de línguas naturais, sejam escritas ou faladas[18,19]. Entretanto, o processamento computacional da fala, ou melhor, das línguas naturais em forma oral, tem ficado a cargo de uma outra área, denominada Reconhecimento e Síntese de Fala (do inglês, *Speech Recognition and Synthesis*) (JURAFSKY, MARTIN, 2000). Esta, por questões tecnológicas e pela especificidade dos sons, tem sido investigada pela Engenharia Elétrica, mais precisamente, na área de Processamento de Sinais. Assim, ressalta-se – e será importante dizê-lo agora – que o termo PLN aplica-se ao processamento computacional de língua natural, tanto nas modalidades escrita e oral, "desde que ambas sejam registradas em meio escrito".

Mais precisamente, o PLN dedica-se a investigar, propor e desenvolver sistemas computacionais que têm a língua natural na forma escrita como objeto primário (GRISHMAN, 1986). Para tanto, os pesquisadores buscam fundamentos em várias disciplinas matrizes: Filosofia da Linguagem, Psicologia, Lógica, Inteligência Artificial, Matemática, Ciência da Computação, Lingüística Computacional e Lingüística (DIAS-DA-SILVA, 1996).

---

[17] Na era pré-computador pessoal, a possibilidade do uso das línguas naturais na comunicação com a máquina já estava entre as questões sob investigação. No entanto, os engenheiros da linguagem visavam apenas à simplificação da vida dos programadores e técnicos que lidavam diretamente com os computadores, sem atentarem para as necessidades do usuário comum.

[18] Linguagens alternativas (p.ex.: a de sinais, para os deficientes auditivos) têm sido igualmente alvo de estudos que visam à sua automatização.

[19] Como salienta Nugues (2006), o processamento de língua (escrita) e o processamento de fala são, por vezes, considerados "processamento de língua natural". Isso acontece, segundo o autor, sob o ponto de vista aplicado ou industrial.

Vê-se, assim, que o PLN é um domínio duplamente heterogêneo. O primeiro aspecto dessa heterogeneidade está ligado aos objetivos, que vão desde a proposição e desenvolvimento de programas que auxiliam a investigar material lingüístico (p.ex.: programas que calculam a freqüência de ocorrências de palavras em textos) até a meta de criar supercomputadores, dotados de inteligência artificial (JURAFSKY, MARTIN, 2000; DIAS-DA-SILVA, 2006). O segundo aspecto heterogêneo está ligado ao fato de que, para concretizar a pluralidade de objetivos, os pesquisadores necessitam percorrer as várias disciplinas matrizes, o que caracteriza esse domínio como multidisciplinar.

O objetivo abrangente e principalmente o caráter multidisciplinar do PLN dificultam delimitar o seu lugar dentre as várias disciplinas correlatas. Para as Ciências da Computação, por exemplo, o PLN é visto como uma subárea da Inteligência Artificial[20]. Isso se deve ao fato de as primeiras indagações sobre o processamento automático das línguas naturais terem sido motivadas por uma das preocupações da Inteligência Artificial, a saber: a interação homem-máquina via "língua dos homens". Muitas vezes, PLN também é usado como sinônimo de Lingüística Computacional. Aliás, "lingüística computacional" comumente nomeia grandes conferências e revistas internacionais que abrangem os estudos de PLN[21]. Entretanto, a Lingüística Computacional, segundo Klavans (1989), Kay (1985) e outros, é o domínio que investiga questões bastante específicas do PLN, a saber: os algoritmos para as análises morfológica e gramatical. Alguns autores, por sua vez, como Bolshakov e Gelbukh (2004), consideram o PLN uma área "mais lingüística que computacional" e, conseqüentemente, uma subárea da Lingüística Aplicada. Já outros, como Nugues (2006), vêem-no como uma legítima interseção entre a Lingüística e as Ciências da Computação.

Além da dificuldade de delimitar o lugar desse campo dentre as disciplinas correlatas, muitos enfatizam que o corpo de conhecimento do PLN é controverso e fragmentado ou, em outras palavras, um conjunto de experiências acumuladas. Na verdade, o PLN não pertence a esta ou àquela área do conhecimento; ele é, como bem salienta Dias-da-Silva (2006), uma área de investigação científica complexa e multifacetada por natureza, sobrepondo-se, por conseguinte, a parcelas das várias áreas correlatas e já consagradas.

---

[20] As disciplinas Inteligência Artificial e Lingüística Computacional são tidas como ramificações das Ciências da Computação. A disciplina IA passou a ser reconhecida pela comunidade científica a partir da chamada *Dartmouth Summer Research Project on Artificial Intelligence* em 1956. A LC, cuja denominação foi cunhada por David Hays em 1967, tinha, em seus primeiros anos, o objetivo de investigar as linguagens de programação e as linguagens formais (DIAS-DA-SILVA, 2006).

[21] Autores como Klavans (1989), Bolshakov e Gelbukh (2004), Mitkov (2004) e vários outros adotam a dominação Lingüística Computacional e não PLN.

Os objetivos

Nestes 50 anos de pesquisas, o PLN tem demonstrado ser um campo fértil em que os pesquisadores têm conseguido desenvolver tecnologias, com graus diferentes de sofisticação e de níveis de desempenho (BOLSHAKOV, GELBUKH, 2004; MARTINS, 2004; MITKOV, 2004):

a) como a elaboração de certos dicionários, *thesaurus* e enciclopédias eletrônicas; essas obras lexicográficas são geralmente compiladas por lexicógrafos e concebidas para uso humano, sendo armazenadas e comercializadas em CD-ROM. A microestrutura dessas obras é, na essência, a dos dicionários impressos. O fato de serem armazenados em formato digital contribui para potencializar toda uma rede de relações morfológicas, sintagmáticas, semânticas e paradigmáticas entre diferentes unidades lexicais e possibilitar, conseqüentemente, o acesso imediato à informação por outras vias que não apenas a **entrada** – único meio para a sua localização nos dicionários impressos (BELIAEVA et al., 1990);
b) como a construção de (i) sistemas de recuperação de informação (do inglês, *automatic information retrieval systems*), que buscam ou encontram textos (ou parte de textos) relevantes a uma dada "consulta" (do inglês, *query*) em uma coleção de textos ou documentos (TZOUKERMAN et al, 2004); nesses sistemas, documentos representam um tipo de informação, cuja recuperação, em outras palavras, pode ser definida como a seleção de documentos, caracterizados por um conjunto de descritores (palavras-chave ou outros símbolos), como resposta a uma consulta; e de (ii) sistemas de extração de informação (do inglês, *information extraction systems*), que buscam encontrar certa informação, ou seja, uma resposta, a dada pergunta de entrada em um ou mais documentos (GRISHMAN, 2004);
c) como a proposição e implementação de (i) sistemas de tradução automática (do inglês, *automatic translation systems*), que partem de um texto-fonte, escrito em uma língua natural *x*, e produzem um texto-alvo, ou seja, uma versão do texto-fonte em uma língua *y*; alguns desses sistemas são ditos "completos", pois funcionam totalmente sem a intervenção de humanos, outros, por sua vez, são ditos "de auxílio à tradução" (SLOCUM, 1985; NIRENBURG, 1989; HUTCHINS, 2004; SOMERS, 2004); e de (ii) sistemas de sumarização automática (do inglês, *automatic summarization systems*): esses sistemas caracterizam-se por gerar "extratos" (justaposição de porções do texto fonte) ou "resumos"

(texto gerado a partir de um plano de resumo) de um ou mais textos de acordo, por exemplo, com uma determinada taxa de compressão (HOVY, 2004).

A construção desses sistemas nem sempre é o foco das investigações. Muitas vezes, busca-se pesquisar questões relativas a processos, métodos e recursos necessários à construção dos sistemas de PLN.

Quanto aos processos, Mitkov (2004) salienta, por exemplo: a etiquetação morfossintática (do inglês, *part-of-speech tagging),* a segmentação textual (do inglês, *text segmentation),* análise sintática (do inglês, *parsing),* a resolução da anáfora (do inglês, *anaphora resolution),* a desambiguação de sentido lexical (do inglês, *word-sense desambiguation)*, entre outros. Por vezes, a investigação desses processos resulta na construção de ferramentas (ou instrumentos) de PLN. Por exemplo, a investigação das questões relacionadas à etiquetação morfossintática pode levar à construção de um etiquetador morfossintático[22] (do inglês, *part-of-speech tagger*) e a investigação dos problemas relativos à análise sintática automática pode gerar um analisador sintático[23] (do inglês, *parser*). Além dessas ferramentas, há também as seguintes: lematizador[24] (do inglês, *lemmatizer*), radicalizador[25] (do inglês, *stemmer*), concordanceador[26] (do inglês, *concordancer*), entre outras. Algumas delas são componentes essenciais de vários sistemas.

Quanto aos métodos ou técnicas, os pesquisadores têm investigado a viabilidade de diferentes abordagens para a construção de sistemas de PLN. Atualmente, co-existem pesquisas realizadas segundo abordagens lingüísticas, não-lingüísticas (ou estatísticas) e híbridas (MARTINS, 2004). As abordagens lingüísticas pautam-se na especificação explícita e declarativa de propriedades e de regras ou padrões regulares de comportamento lingüístico. As abordagens não-lingüísticas, por sua vez, pautam-se na recuperação/identificação, induzida automaticamente, de regularidades subjacentes aos dados lingüísticos e, por isso, necessitam de extensos *corpora* para que os padrões possam ser identificados. As estratégias

---

22 Ferramenta computacional responsável pela marcação de um texto com etiquetas morfossintáticas. Esses etiquetadores podem ser construídos manualmente, por lingüistas, ou automaticamente, extraídos de *corpus*. (VOUTILAINEN, 2004).

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

24 Ferramentas que reduzem cada palavra de um texto ao seu lema ou forma canônica, ou seja, formas não-marcadas, desprovidas de flexões (SPARCK-JONES, WILLET, 1997). Na lematização, os verbos são reduzidos ao *infinitivo* (p.ex.: casamos > casar) e os substantivos e adjetivos ao *masculino singular* (p.ex.: latas > lata/ feias > feio).

25 Ferramentas computacionais que reduzem as palavras de um texto ao seu radical (SPARCK-JONES, WILLET, 1997).

26 Ferramentas computacionais que produzem concordâncias ou listagens das ocorrências de um item específico acompanhado do texto ao seu redor (co-texto) (BERBER SARDINHA, 2004).

híbridas, por fim, reúnem as características das lingüísticas e das não-lingüísticas (DORR et al, 1999).

A construção de certas ferramentas e a aplicação de determinados métodos ou técnicas, ambos importantes para o subseqüente desenvolvimento de sistemas de PLN, necessitam, quase sempre, dos chamados recursos lingüístico-computacionais, cujo planejamento (e construção) constitui tarefa nada trivial. Exemplos desses recursos são:

a) **corpora** (textuais): coleções de textos úteis para o levantamento de conhecimento lingüístico (lexical, sintático, semântico, etc.). Esse levantamento pode ser feito por lingüistas, com a ajuda de programas de manipulação de *corpus*, ou por meio da aplicação de métodos estatísticos. A extração do conhecimento exige que a quantidade de textos seja grande, variada e representativa e que os textos estejam em formato adequado para que a extração possa ser automática (BERBER SARDINHA, 2004);
b) **léxicos**: estoques de unidades lexicais descritas juntamente com seus traços morfológicos, sintáticos, semânticos e/ou pragmático-discursivos e sistematicamente organizadas de acordo com algum critério. Tanto as unidades quanto as propriedades a elas associadas podem ser representadas por formalismos altamente estruturados (HANDKE, 1995);
c) **ontologias** e **bases de conhecimento**: inventários de conceitos, propriedades e relações entre conceitos que representam "uma interpretação da realidade", ou seja, o conhecimento de mundo compartilhado pelos membros de uma comunidade lingüística. A representação de uma ontologia pode variar segundo o grau de formalização. Uma ontologia formal, em especial, apresenta os conceitos e as relações (entre conceitos) explicitamente definidas, ou seja, "legíveis pela máquina" (GRUBER, 1995);
d) **gramáticas***:* sistemas de regras expressos segundo sistemas formais, que (i) descrevem as estruturas das sentenças de uma dada língua e (ii) permitem, juntamente com o léxico, reconhecer e gerar sentenças dessa língua (KAPLAN, 2004).

Por fim, salienta-se que o PLN também possui um viés acadêmico e não somente científico-tecnológico. Dentre os objetivos dos pesquisadores, estão (i) a investigação da adequação formal, pragmática e psicossocial de teorias lingüísticas por meio da implementação dos modelos de gramática e de processamento lingüístico especificados por essas teorias e a própria (ii) proposição de sofisticados modelos computacionais capazes, por exemplo, de extrair informações específicas de bases de textos (VARELI, ZAMPOLLI, 1997).

A motivação lingüística, o desafio e o compromisso

Segundo Dias-da-Silva (2006), que se inspira em Winograd (1972), um sistema de PLN pode ser visto como um tipo especial de "sistema especialista" na medida em que requer uma parcela específica do conhecimento humano – o conhecimento lingüístico – para realizar tarefas específicas como correção ortográfica, tradução automática, etc.

No âmbito da Inteligência Artificial, um **sistema especialista** (do inglês, *expert system*) é um sistema computacional inteligente, que toma decisões e resolve problemas referentes a um determinado campo de atuação, como finanças e medicina, utilizando conhecimento e regras analíticas definidas por especialistas no assunto (JACKSON, 1990; HAYES-ROTH, 1990; MICROSOFT PRESS, 1998; GIARRATAMO, RILEY, 2004). Um sistema de diagnóstico, por exemplo, necessita saber quais as características das doenças a serem diagnosticadas, pois, sem elas, é impossível elaborar um diagnóstico automaticamente. Dentre os sistemas especialistas descritos na literatura, destacam-se o (i) Dendral, primeiro sistema especialista, criado para ajudar os químicos a determinar a estrutura molecular, (ii) o Mycin, que diagnostica doenças sangüíneas infecciosas, e o (iii) Dipmeter Advisor, que auxilia na análise de dados recolhidos durante a exploração de petróleo.

Projetar, então, um sistema de PLN, ou seja, um sistema que simule parcelas da competência e do desempenho lingüístico humanos, pressupõe a especificação de vários conhecimentos e habilidades que os falantes – especialistas nesse domínio – possuem. Mais precisamente, Dias-da-Silva (1996) resume que, para simular uma língua natural de modo satisfatório, um sistema de PLN deve conter vários sistemas de "conhecimento" e "realizar" uma série de atividades cognitivas, tais como: (i) possuir um "modelo simples de sua própria mentalidade"; (ii) possuir um "modelo detalhado do domínio específico do discurso"; (iii) possuir um modelo que represente "informações morfológicas, sintáticas, semânticas, (iv) contextuais e do conhecimento de mundo físico"; (v) "compreender o assunto que está em discussão"; (vi) "lembrar, discutir, executar seus planos e ações"; (vii) participar de um diálogo e responder, com ações e frases, às frases digitadas pelo usuário; (viii) solicitar esclarecimentos quando seus programas heurísticos não conseguirem; (ix) compreender uma frase.

Em outras palavras, para as pesquisas que adotam a "concepção lingüisticamente motivada de PLN" – na qual se acredita neste trabalho –, o computador não poderá satisfatoriamente emular uma língua natural se não conseguir, em alguma medida, compreender o assunto que está em discussão. Logo, é preciso fornecer à máquina descrições e formalizações de dados lingüísticos nas dimensões: morfológica, sintática, semântico-conceitual e pragmático-

discursiva (ROCA, 2000). E aí a Lingüística tem um papel imprescindível, pois, apesar dos aspectos problemáticos comumente apontados pelos engenheiros da liguagem, ela apresenta os parâmetros norteadores essenciais a respeito das características e funções das línguas naturais a que os investigadores do PLN podem recorrer.

Para o desenvolvimento de uma pesquisa "linguisticamente motivada de PLN", acredita-se ser necessário, como defende Dias-da-Silva (1996, 1998, 2006), o trabalho colaborativo entre os cientistas e os engenheiros da linguagem.

Essa colaboração, apesar de certa evolução nos últimos anos, continua longe de ser a ideal, conforme aponta Dias-da-Silva (2006). Há, ainda, o distanciamento entre essas duas comunidades, o que dificulta e/ou atrasa a descoberta de soluções e o conseqüente avanço no desenvolvimento dos recursos, ferramentas e, logo, dos sistemas. Tal distanciamento tem sido justificado por razões técnicas fornecidas por ambos os lados. Os engenheiros criticam, por exemplo, a pluralidade, a incompletude e a pouca formalização das descrições lingüísticas, o linguajar técnico muitas vezes hermético e a preocupação dos lingüistas em estudar a linguagem humana *per se.* Os lingüistas, por sua vez, enfatizam que os engenheiros – tidos como indivíduos com pouca intuição sobre os fatos da língua – concentram-se no desenvolvimento de sistemas rudimentares e desprovidos de qualquer fundamentação lingüística[27]. A falta de contato entre essas duas comunidades, aliás, também é regada por imagens estereotipadas e distorcidas que os pesquisadores de uma área têm do trabalho realizado na outra, principalmente quando as áreas de conhecimento são tão distintas, como Lingüística e Ciências da Computação.

No entanto, adota-se neste trabalho o mote "cooperar é preciso" (DIAS-DA-SILVA, 2006). Nos casos em que o distanciamento foi vencido, a colaboração entre lingüistas e cientistas da computação mostrou-se não somente benéfica para o PLN, mas também para a Lingüística e Ciência da Computação. A Lingüística, por exemplo, tem se beneficiado, do ponto de vista prático, com vários recursos que auxiliam na análise de material lingüístico. Do ponto de vista teórico, tem se beneficiado também com a formulação de modelos descritivos mais completos (ou seja, modelos de análise e descrição de cada um dos estratos da gramática e do inter-relacionamento entre os módulos da competência e do desempenho) e explícitos (ou seja, descritos em termos de linguagens formais). Exemplos paradigmáticos desse tipo de contribuição do PLN são: (i) implementação, teste e avaliação de gramáticas propostas pela

[27] Por exemplo, os próprios dicionários eletrônicos, em que o material lingüístico é apenas manipulado por meio de técnicas de indexação, e os chamados "tradutores de bolso", que são limitados a manipular listas de palavras, expressões e fragmentos de frases por meio principalmente de comparações e substituições com o objetivo de montar/ completar frases com as palavras e/ou expressões armazenadas (DIAS-DA-SILVA, 2006).

Lingüística Teórica (GRISHMAN, 1986), como a gramática funcional de Dik (1997) (SIEWIERSKA, 1991; ATKINS, ZAMPOLLI, 1994) e parcela da gramática funcional de Halliday (1985) (BUTLER, 1985); (ii) desenvolvimento de modelos gramaticais, como a HPSG (POLLARD, SAG, 1994); (iii) proposição de modelos diversos, p.ex.: modelos computacionais dos atos de fala, modelos computacionais da teoria da referência (DIAS-DA-SILVA, 2006).

Conclui-se esta "apresentação-advertência", alertando que o PLN é uma área complexa e multifacetada e que, mesmo aparentemente caótica, tem se mostrado produtiva. Além dos sistemas, ferramentas e recursos citados ao longo deste texto, citam-se ainda as seguintes tecnologias desenvolvidas pelo PLN:

a) **sistemas de correção ortográfica** (do inglês, *spelling checker systems*): processam um texto em uma dada língua natural com os objetivos de (i) identificar os erros cometidos quanto à ortografia (palavras que não constam do léxico dessa língua ou usadas em contexto impróprio) e (ii) sugerir alternativas prováveis e ortograficamente corretas a cada erro identificado;
b) **sistemas de correção gramatical** (do inglês, *grammar checker systems*): detectam, embora de modo rudimentar, desvios gramaticais em um texto, como os de concordância nominal ou verbal, pontuação, regência nominal e outros;
c) **analisadores semânticos** (do inglês, *semantic parsers*): extraem automaticamente parcelas do conhecimento semântico subjacente a um texto;
d) **analisadores discursivos** (do inglês, *discourse parsers*): extraem automaticamente parcelas do conhecimento discursivo (isto é, do nível da pragmática e do discurso) subjacente a um texto;
e) **sistemas de categorização de textos** (do inglês, *text categorization systems*): classificam, mesmo que de modo parcial, textos de acordo com algum critério (p. ex.: domínio, gênero, estilo, retórica, autoria, etc.);
f) **sistemas de diálogos** (do inglês, *dialog systems*): englobam os sistemas de interpretação de diálogos e os sistemas que participam de um diálogo, geralmente travado com o usuário (p.ex. um sistema de reserva/compra de passagens.);
g) **sistemas de auxílio à escrita** (do inglês, *computer-assisted writing system*): auxiliam a produção de texto, em que o usuário pode encontrar recursos para construir textos bem estruturados, de um gênero e/ou domínio específicos, entre outros.

Vale ressaltar que, por aspectos econômicos, as tecnologias em PLN são, na maioria das vezes, desenvolvidas para a língua inglesa, sendo que tais tecnologias não são diretamente transportáveis para outras línguas. Dessa forma, o processamento de uma língua natural requer o desenvolvimento de recursos e ferramentas de base que dêem suporte para o desenvolvimento de sistemas voltados para a língua em questão.

O PLN no Brasil, isto é, o processamento computacional do português do Brasil, ainda escasso em tecnologias em relação ao AmE, mostrou significativa evolução nos últimos anos, mas muito ainda há para ser feito. Como resultado dessa evolução, a comunidade do PLN no Brasil dispõe hoje, por exemplo, de:

a) **sistemas de sumarização** (extrativa monodocumento) (PARDO et al., 2002; RINO et al., 2004; LEITE et al. 2007) e de correção ortográfica e gramatical (MARTINS et al., 1998);
b) **ambiente de auxílio à escrita** (científica) (FELTRIM, 2004);
c) **recursos de base**, como léxicos (NUNES et al., 1996; MUNIZ, 2004), *thesaurus* (DIAS-DA-SILVA et al., 2000), *corpora* (ALUÍSIO et al., 2003) e bases lexicais (ZAVAGLIA, 2002, DI FELIPPO, DIAS-DA-SILVA, 2007) e ontológicas (ZAVAGLIA, 2005);
d) **ferramentas**, como segmentador sentencial (PARDO, 2006), alinhador lexical (isto é, ferramenta que alinha unidades lexicais de *copora* paralelos) (CASELI et al., 2005); etiquetadores morfossintáticos (AIRES, 2000), *parsers* (BICK, 2000; MARTINS et al., 1998) e analisador retórico (monodocumento) (PARDO, NUNES, 2006), entre vários outros.

3.2. O PLN e as representações de conhecimento (RC)

Ao se assumir a concepção de Dias-da-Silva (2006), segundo a qual a construção de um sistema de PLN requer uma espécie de "engenharia do conhecimento lingüístico", assume-se, por conseguinte, que as pesquisas em PLN podem se beneficiar das estratégias desenvolvidas na área da Engenharia do Conhecimento. Mais precisamente, tais pesquisas podem se beneficiar dos chamados **modelos de representação de conhecimento** (DIAS-DA-SILVA, 2006).

Davis et al (1993) buscam definir RC em função dos vários papéis distintos e essenciais que ela desempenha. Segundo esses autores, uma RC é: (i) um substituto, (do inglês, *surrogate*) (ii) um conjunto de compromissos ontológicos (do inglês, *ontological*

commitments), (iii) uma teoria do raciocínio (do inglês, *intelligent reasoning*), (iv) um meio para uma computação eficiente e (v) um meio de expressão humana.

3.2.1. A RC como um substituto do conhecimento

Uma representação do conhecimento funciona como um substituto para o conhecimento representado (isto é, "a coisa ou mundo original"). Entender as representações do conhecimento dessa forma implica algumas questões, p.ex.: (i) o grau de fidelidade do substituto em relação ao elemento a ser substituído e (ii) a explicitação ou a omissão por parte do substituto de atributos do elemento a ser substituído. Naturalmente, uma fidelidade total é praticamente impossível, tanto na prática como na teoria (WOODS, 1985). Teoricamente, essa fidelidade é impossível porque qualquer coisa, com exceção do original, será necessariamente diferente da coisa representada. Em outras palavras, a representação mais adequada de um conceito, por exemplo, seria o próprio conceito, sendo que todas as outras representações seriam inadequadas sob esse ponto de vista; elas inevitavelmente contêm simplificações da coisa substituída. Dessa forma, a imperfeição dos substitutos é inevitável, sendo que, mesmo o substituto mais adequado não impede que se lide com o fato de que as representações são imperfeitas e, por isso, geradoras de possíveis erros. Naturalmente, a "arte" de se selecionar uma representação está em encontrar uma que mais se aproxime do original e, com isso, minimize os possíveis erros.

3.2.2. A RC como um conjunto de compromissos ontológicos

Todas as representações, tidas como substitutos, são aproximações do conhecimento a ser representado. Isso ocorre porque cada aproximação focaliza alguns aspectos da coisa original e ignora outros. Conseqüentemente, selecionar uma determinada representação do conhecimento é, ao mesmo tempo, adotar um conjunto de decisões sobre o que "ver" no original. Nas palavras de Davis et al. (1993), selecionar uma representação do conhecimento significa assumir um conjunto de "compromissos ontológicos" (do inglês, *ontological commitments*), que pode ser considerado a contribuição mais importante de uma representação do conhecimento.

Quanto ao termo "ontológico", cabe uma ressalva, já que este tem sido definido de várias formas. Historicamente, o termo ontologia origina-se do grego *ontos*, ser, e *logos*, palavra. O termo original é a palavra aristotélica *categoria*, que pode ser usada para classificar alguma coisa. Aristóteles apresenta categorias que servem de base para classificar qualquer entidade e introduz ainda o termo *differentia* para propriedades que distinguem diferentes espécies do

mesmo gênero. Em seu sentido filosófico, ó termo ontologia é assim definido no Dicionário Oxford de Filosofia: '[...] o termo derivado da palavra grega que significa "ser", mas usado desde o século XVII para denominar o ramo da metafísica que diz respeito àquilo que existe' (BLACKBURN, 1997).

No âmbito da Ciência da Informação e das Ciências da Computação (e, por conseguinte, do PLN), esse termo tem um sentido particular, diferente daquele adotado na Filosofia. Talvez a definição mais conhecida nesses domínios seja a de Gruber (1995), segundo a qual uma "ontologia" é a "especificação de uma conceitualização". De acordo com essa definição, dizer que uma representação do conhecimento determina um conjunto de compromissos ontológicos significa dizer que ela determina uma visão simplificada do conhecimento que se deseja representar. Em outras palavras, esses "compromissos" são os óculos ou o conjunto de princípios fundamentais (ou teoria) que determina o que se ver, focalizando certos aspectos e negligenciando outros.

Esses compromissos ontológicos não são efeitos colaterais da escolha de uma representação do conhecimento; eles são, na verdade, a essência. Por um lado, eles são inevitáveis devido à imperfeição dos substitutos e, por outro lado, permitem que os aspectos da coisa original, considerados relevantes, sejam focalizados. O efeito de focalização das representações do conhecimento é fundamental porque fornece as diretrizes para o recorte do complexo mundo do conhecimento a ser representado. Por exemplo, os circuitos elétricos podem ser vistos em termos de componentes, conexões (entre os componentes) e sinais que fluem ao longo das conexões. Essa, entretanto, não é a única visão possível dos circuitos elétricos; a Eletrodinâmica, por exemplo, tem outro ponto de vista sobre tais circuitos.

Borst (1997), ao refinar a definição de Gruber (1995), define ontologia 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).

As ontologias podem ser representadas por meio de várias metalinguagens formais ou formalismos (p.ex.: lógica, redes semânticas e *frames*); a informação relevante, no entanto, não é a forma, mas o conteúdo, ou seja, o conjunto de noções oferecidas como um modo de "ver" o conhecimento. No caso do exemplo dos circuitos elétricos, a informação essencial é o conjunto de noções como componentes, conexões e sinais.

Há, entretanto, uma ligação significativa e inevitável entre a visão de um determinado conhecimento e o modo como ele é formalizado. Em outras palavras, as metalinguagens

utilizadas para a descrição formal de uma ontologia, no sentido aqui descrito, carregam em si um determinado ponto de vista sobre o conhecimento a ser moldado.

Sobre essa questão, Davis et al. (1993) salientam que os primeiros compromissos ontológicos são assumidos quando se escolhe uma metalinguagem de descrição de conhecimento, pois toda linguagem formal tem expressividade limitada e não permite, conseqüentemente, representar alguma coisa sobre o domínio em questão. Por exemplo, ao se adotar a lógica, adota-se, ao mesmo tempo, a visão do conhecimento original em termos de indivíduos e relações entre eles, descritos como externos à própria língua.

Cada metalinguagem, então, fornece seu próprio ponto de vista sobre o que é importante. Conseqüentemente, selecionar certa metalinguagem implica comprometer-se com o ponto de vista por ela determinado.

3.2.3. RC como teoria do raciocínio[28]

Segundo Davis et al. (1993), uma RC, como teoria do raciocínio, provê:

(i) **uma concepção de raciocínio inteligente***:* determina o modo como o raciocínio é visto. Há duas grandes correntes de raciocínio inteligente; em uma delas, dita lógica, o raciocínio é visto como uma operação lógica, realizada tipicamente pela dedução[29]; na outra, dita psicológica, o raciocínio é considerado integrante dos processos cognitivos de formação de conceitos e de resolução de problemas, sendo, portanto, parte do pensamento.

(ii) **um conjunto de inferências sancionadas**: a automatização do raciocínio compreende o estudo de métodos de inferência, pelos quais novos conhecimentos podem ser obtidos, por derivação, a partir do conhecimento disponível. Na corrente lógica, o raciocínio opera apenas por meio de inferências "consistentes" (do inglês, *sound inferences*), ou seja, aquelas que somente geram conclusões que sejam conseqüências lógicas das premissas; na corrente psicológica, o raciocínio opera também por meio de inferência não-monotônica. O tipo mais simples de raciocínio não-monotônico é um raciocínio no qual uma conclusão

[28] Hirst (1992) argumenta que um modelo lingüístico-computacional para o tratamento da semântica das línguas naturais deve, entre outras coisas, ser manipulável por regras de inferência.

[29] Os raciocínios dedutivos caracterizam-se por apresentarem conclusões que devem ser necessariamente verdadeiras, se todas as premissas forem verdadeiras. Por exemplo: Se "Todo mamífero tem coração" e "Todos os cavalos são mamíferos", então "Todos os cavalos têm coração". No exemplo apresentado, sendo verdadeiras as duas premissas, a conclusão é necessariamente verdadeira. Outra característica dos raciocínios dedutivos é que aquilo que é dito na conclusão, de alguma forma, já tinha sito dito nas premissas. Como em todos os raciocínios dedutivos corretos, a conclusão reformula a informação contida nas premissas.

deve ser inferida em primeira instância, mas a conclusão pode ser abandonada se houver evidências contrárias[30].

3.2.4. A RC como um meio de processamento computacional eficiente

Nessa função, uma representação do conhecimento precisa ser aplicável computacionalmente. Quanto a esse papel, Woods (1985) destaca dois aspectos que devem ser considerados ao representar conhecimento: (i) a eficácia notacional, que "se preocupa com a forma real e a estrutura da representação, assim como o impacto desta estrutura nas operações do sistema"; e (ii) a adequação expressiva (do inglês, *expressive adequacy*), que diz respeito ao "poder expressivo da representação", ou seja, o que ela pode representar. Ele ainda subdivide a eficácia notacional em alguns pontos: eficiência computacional, clareza conceitual, concisão da representação e facilidade de modificação; e sustenta que as pesquisas na área de RC deveriam se dedicar simultaneamente à eficácia notacional e à adequação expressiva.

3.2.5. A RC como um meio de expressão humana

Por fim, representações do conhecimento são meios pelos quais os humanos dizem algo ou comunicam algo aos computadores. Desse ponto de vista, a representação deve ser legível não só para as máquinas, mas também para os humanos. Davis et al. (1993) destacam que uma tecnologia de representação é uma linguagem em que os humanos se comunicam e, dessa forma, eles devem ser capazes de se expressar por meio delas sem esforço.

A seguir, apresentam-se os principais paradigmas de representação do conhecimento e discute-se sua adequação para o tratamento da semântica das línguas naturais, em especial, no tratamento dos itens lexicais.

3.3. Três paradigmas de RC

Com base em Cercone e McCalla (1987), Brachman e Levesque (1985, 2004) e Helbig (2006), foi possível identificar três grandes paradigmas de RC: (i) RCs baseadas na lógica (do inglês, *logic-oriented KR*)*;* (ii) RCs baseadas em *frames* (do inglês, *frame-oriented KR)*; (iii) RCs baseadas em redes semânticas (do inglês, *network-oriented KR*). A seguir, cada paradigma é descrito em função da abordagem teórica do significado e da metalinguagem formal que englobam.

[30] Por exemplo, quando se ouve algo sobre uma ave, infere-se que ela pode voar. Mas essa conclusão pode ser revertida quando se sabe que se trata de um pingüim.

3.3.1. As RCs baseadas na lógica

3.3.1.1. Abordagem teórica do significado: referencial

Os modelos de RC baseados na lógica foram os primeiros usados pelos pesquisados em Inteligência Artificial para representar estruturas de conhecimento nos computadores. Inclusive, foram considerados durante muito tempo como a resposta para todas as necessidades dos sistemas artificiais de compreensão das línguas naturais.

Nesses modelos, o significado é tratado sob o ponto de vista da **semântica referencial**, **extensional** ou **denotacional** (cf. DOWTY et al., 1981; BARWISE, PERRY, 1983). Segundo essa abordagem, tal como se constituiu a partir daquele que costuma ser considerado seu fundador, G. Frege (1990)[31], as expressões de uma língua adquirem significado ao denotarem objetos e eventos do mundo.

Segundo esse enfoque, o significado de expressões lingüísticas (referenciais) divide-se em **sentido** e **denotação** (do alemão, *Sinn* e *Bedeutung*). Assim, diz-se que uma expressão lingüística expressa um sentido e apresenta uma denotação ou denota algo.

O sentido, segundo Frege, é a forma de apresentação de um objeto, sendo que a denotação é o próprio objeto. O princípio da substituição define que sentidos diferentes podem remeter a uma mesma denotação. Por exemplo, em *O autor de Memórias Póstumas de Brás Cubas era mulato*, pode-se substituir a descrição definida *o autor de Memórias Póstumas de Brás Cubas* pela expressão *o fundador da Academia Brasileira de Letras*, sem alterar o valor de verdade da proposição expressa pela sentença em questão (MOURA, 2000). Ainda quanto à noção de sentido, diz-se que *O autor de Memórias Póstumas de Brás Cubas era mulato* e *O fundador da Academia Brasileira de Letras era mulato* são modos diferentes de apresentação (ou sentidos) de uma mesma proposição.

A denotação**,** por sua vez, é o conjunto potencial de indivíduos a que uma expressão se refere no contexto de emissão. Categorias de expressões distintas têm denotações também distintas.

Em primeiro lugar, apresenta-se a denotação dos predicados (lógicos e não gramaticais), como nomes comuns, adjetivos e verbos. Por exemplo, um nome comun, como *bicicleta*, denota cada um dos objetos que se classifica como bicicleta, ou seja, o conjunto das bicicletas do mundo. Na verdade, devido ao fato de que o conjunto das bicicletas do mundo muda constantemente, considera-se o conjunto de todas as bicicletas possíveis. Adjetivos, por sua

[31] Tradução do texto original em alemão *'Über Sinn und Bedeutung'*, publicado em 1892.

vez, como *inteligente*, denotam cada um dos indivíduos que pertencem ao conjunto dos seres inteligentes. Verbos transitivos, como *querer*, denotam pares de indivíduos em que x quer y. Por fim, verbos intransitivos também denotam; pode-se dizer que *cantar*, por exemplo, denota os seres que cantam.

Em segundo lugar, apresenta-se a denotação dos nomes próprios, também denominados termos singulares. Os nomes próprios denotam certo indivíduo; por exemplo, *Machado de Assis* denota um indivíduo que escreveu *Memórias Póstumas de Brás Cubas* e que fundou a *Academia Brasileira de Letras*.

Por fim, apresenta-se a denotação de sentenças (ou frases). Nesse caso, diz-se que sentenças denotam valor de verdade, ou seja, a denotação de uma sentença é o seu valor de verdade. Na abordagem lógica, a idéia é a de que as sentenças declarativas denotam um valor de verdade, F(also) ou V(erdadeiro). Naturalmente, a proposição expressa por uma sentença pode ser F ou V, dependendo do contexto em que fora emitida.

No âmbito da lógica intensional, as noções de sentido e denotação foram reformuladas e denominadas, respectivamente, **intensão** e **extensão** (ALLAN, 2001).

A intensão de uma expressão lingüística é o conjunto de propriedades características de seu *denotatum* (aquilo que é denotado). Por exemplo, as expressões *o autor de Memórias Póstumas de Brás Cubas* e *o fundador da Academia Brasileira de Letras* diferem em intensão; pois na primeira está presente a noção de escritor, enquanto que na segunda está presente a noção de fundador. As intensões são muitas vezes descritas como conceitos (p.ex.: CANN, 1993), no entanto, os conceitos são entidades cognitivas e as intensões, abstratas (PARTEE, 1979, ALLAN, 2001).

A extensão, por sua vez, pode ser definida como o objeto ou objetos aos quais a expressão lingüística se aplica. Por exemplo, a extensão do termo singular *o autor de Memórias Póstumas de Brás Cubas* coincide com a extensão de *o fundador da Academia Brasileira de Letras*, pois ambos denotam o mesmo indivíduo, no caso, Machado de Assis. Desse ponto de vista, *pégaso* e *a fonte da vida* têm como extensões conjuntos vazios. Mais precisamente, diz-se que a extensão de um termo singular é o conjunto-unidade desse objeto; no exemplo, a extensão de ambos os termos é o conjunto-unidade Machado de Assis. A extensão de um predicado, como *admirar*, é o conjunto de todos aqueles pares ordenados de entidades x e y tais que x admira y. Vale ressaltar que a intensão de um predicado como esse é a propriedade P atribuível a indivíduos. A extensão de um predicado como *bicicleta*, por sua vez, é o conjunto de todos os objetos classificados como bicicleta, sendo que a intensão é o conjunto de propriedades características de um típico *denotatum* de bicicleta. Aplicando as noções de

extensão e intensão às sentenças, diz-se que a extensão é o valor de verdade da proposição expressa pela sentença e a intensão é a própria proposição.

A intensão (ou significado) de um nome comum, como *bicicleta*, pressupõe o método clássico de categorização. Vale ressaltar, aqui, que o termo "clássico" está sendo empregado no mesmo sentido de Taylor (1985), Cruse (2004) e Croft e Cruse (2004): método que se baseia em Aristóteles e que dominou os estudos da Psicologia, Filosofia e Lingüística durante quase todo o século XX.

Segundo esse método, o significado de uma expressão pode ser exaustivamente decomposto em um conjunto finito de primitivos conceituais que são condições necessárias e suficientes para determinar o sentido e a denotação da expressão. Assim, essa lista de propriedades precisa ser satisfeita para o uso correto de uma expressão. Em outras palavras, a denotação pertence a uma categoria determinada se, e somente se, exibe todos e cada um dos primitivos que a definem; a falta de algum desses primitivos significa a sua exclusão automática da categoria.

Essa concepção de categorização baseia-se na distinção feita por Aristóteles entre a essência (isto é, partes imanentes responsáveis pela individualização) de uma coisa e seus acidentes (isto é, propriedades que não são definitórias). Para Aristóteles, o significado de *bicicleta*, por exemplo, é definido por uma fórmula da essência (TAYLOR, 1985). Em outras palavras, dizer que x é um y implica colocar uma entidade x na categoria y. Para isso, faz-se uma comparação das propriedades de x com os traços que definem a essência da categoria y, sendo que esse conjunto de propriedades caracteriza o significado de uma expressão.

Ilustra-se, aqui, a adoção desse método, tido clássico, com o seguinte exemplo: o significado de *bicicleta*. O significado de *bicicleta* pode ser representado, por meio da análise componencial, pelos traços [veículo], [duas rodas], [quadro], [selim] e [movida por pedais]. Esse conjunto de traços inclui duas características: o "gênero próximo" e a "diferença específica". No caso, *bicicleta* define-se pelo fato de pertencer à classe dos veículos (gênero próximo) e pelo fato de ter características específicas como duas rodas, selim, quadro e ser movida por pedais (diferença específica). Esses traços juntos formam a essência de "bicicletidade" ou de "ser uma bicicleta". Segundo o princípio de que os traços da lista são necessários e essenciais, qualquer entidade no mundo que apresente esses traços pode ser corretamente designada pela expressão *bicicleta*. Caso contrário, se um dos traços não for encontrado ou se o valor de um dos traços for diferente, a entidade não apresenta a propriedade "ser uma bicicleta".

Espera-se que o *denotatum* de um *pássaro*, por exemplo, seja bípede, tenha penas e seja capaz de voar. No entanto, há várias espécies de pássaros que não voam, por exemplo, os pingüins. Assim, vê-se que a noção de uma lista de propriedades necessárias nem sempre é suficiente para identificar ou individualizar um objeto.

Resumidamente, a abordagem referencial privilegia o aspecto informacional da língua, focalizando as conexões entre a língua e o mundo, seja ele real ou imaginário.

3.3.1.2. A metalinguagem formal dos modelos baseados na lógica

Do ponto de vista formal, as RC baseadas na lógica (clássica) utilizam os instrumentos de representação empregados na Lógica, que são os conjuntos e os modelos, buscando, assim, explicitar estruturas formais que descrevam a maneira como as expressões de uma língua natural denotam entidades do mundo.

Um **conjunto** é uma coleção de objetos, originários de um determinado domínio (p.ex.: números, indivíduos, etc.); o domínio de onde são originários os objetos de um conjunto é chamado de **universo de discurso** (U) (CHIERCHIA, 2003).

O termo **modelo**, por sua vez, é, grosso modo, empregado para designar uma representação simplificada de uma realidade complexa que objetiva facilitar a compreensão dessa última. Do ponto de vista formal, entretanto, modelo é uma estrutura (um conjunto de objetos com propriedades e relações definidas sobre esses objetos) construída de tal forma que as expressões traduzidas para a lógica possam ser interpretadas. Por essa razão, a semântica referencial também recebe o nome de **semântica de modelo** (extensional) (ALLWOOD et al., 1977; ALLAN, 2001; MÜLLER, 2003; HELBIG, 2006).

Como salienta Müller (2003), ao se adotar a lógica como metalinguagem, é preciso apresentar uma tradução sintática das expressões lingüísticas para a linguagem formal e explicar como as expressões traduzidas para a lógica são interpretadas no mundo. Em outras palavras, a representação da semântica segundo a tradição lógica requer a tradução da língua natural e a organização das entidades em um modelo de mundo logicamente organizado para que as expressões traduzidas possam ser interpretadas, posto que as expressões lingüísticas são "vazias", adquirindo significado na relação com um modelo.

Exemplificando-se como se dá a representação semântica nos modelos de RC de base lógica, considere a unidade lexical *bicicleta*. Sob o ponto de vista extensional, nomes comuns (concretos e contáveis) como *bicicleta* denotam conjunto de indivíduos. Dessa forma, pode-se dizer que o significado da expressão *bicicleta* é "o conjunto de todas as coisas do mundo ao

qual ela se aplica". Suponha-se que o modelo de mundo possa ser reduzido ao conjunto U descrito na Figura 14.

Usando-se noções da teoria dos conjuntos, pode-se identificar o significado de *bicicleta* pelo subconjunto B de U, que agrupa todos os indivíduos que pertencem à classe das bicicletas. Logo, B fica formalmente como: $B = \{x \in U \mid B(x)\}$ (leia-se: B é o conjunto dos x que pertencem a U e que são bicicletas). Em outras palavras, diz-se que o significado de *bicicleta* é a sua extensão no mundo.

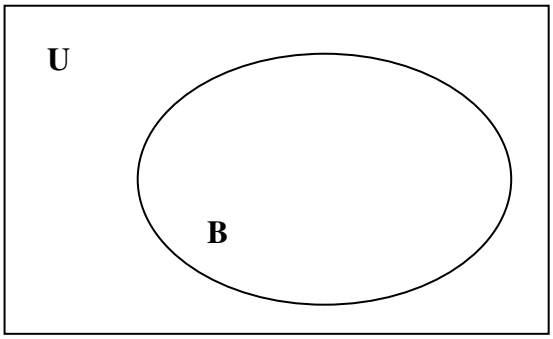


Figura 14: Representação da denotação de *bicicleta* em termos de conjuntos.

Alternativamente, pode-se dizer, sob o ponto de vista intensional, que o significado de *bicicleta* é o conjunto ou a classe dos indivíduos que têm a mesma propriedade, no caso, "ser bicicleta". Nesse caso, o significado de *bicicleta* é a sua intensão, ou seja, o conjunto das propriedades que são comuns a todos os objetos dessa classe (extensão).

Suponha-se que a essência de "ser bicicleta" em U possa ser caracterizada informalmente como <veículo que contém duas rodas, um quadro, um selim e é movido por pedais>. Para representar a semântica de expressões lingüísticas como *bicicleta,* lança-se mão do recurso dos **postulados de significado**, introduzido por Carnap (1958). Segundo Allan (2001) e Cruse (2006), um postulado de significado é o vocabulário não-lógico (no caso deste trabalho, expressões do PB) usado na especificação semântica das expressões.

Assim, os componentes [veículo], [duas rodas], [quadro], [selim] e [movido por pedais], por exemplo, são incluídos na representação semântica de *bicicleta,* como ilustrado na Figura 15. Nessa representação, o símbolo $\wedge$ representa "e". Dessa forma, leia-se: "B é o conjunto dos x que pertencem a U e que são veículos de duas rodas, com quadro, selim e movidos por pedais".

$$B = \{ x \in U \mid \text{veículo}(x) \wedge \text{duas\_rodas}(x) \wedge \text{quadro}(x) \wedge \text{selim}(x) \wedge \text{movido\_por\_pedais}(x) \}.$$

Figura 15: Descrição formal da intensão de *bicicleta*.

Utilizando-se o quantificador universal, (∀), pode-se dizer, considerando U, que:

$$\forall x \mid x \in U \; [\text{bicicleta}(x) \leftrightarrow \text{veículo}(x) \wedge \text{tem\_duas\_rodas}(x) \wedge \text{tem\_quadro}(x) \wedge \text{tem\_selim}(x) \wedge \text{movido\_por\_pedais}(x)]$$

(leia-se: para todo x, tal que x pertence a U, x é bicicleta se e somente se, x é um veículo que tem duas rodas, quadro, selim e é movido por pedais).

Exemplificando-se agora não a representação semântica da expressão em isolado, mas em contexto, considere a seguinte sentença: *A bicicleta quebrou*. Tal sentença pode ser traduzida, segundo o cálculo de predicados[32], na seguinte fórmula: $\exists x \, [B(x) \wedge Q(x)]$ (leia-se: existe um x tal que x é bicicleta e x quebrou).

Para que a interpretação dessa sentença seja possível, pode-se propor, mesmo verbalmente, um modelo ou estado-de-coisas em que haja pelo menos um indivíduo na intersecção entre os conjuntos das coisas que são bicicleta e o conjunto das coisas que quebram. De acordo com o modelo de mundo U, o significado da referida sentença pode ser representado pela Figura 16, sendo que o valor de verdade da proposição expressa por ela, $\exists x \, (B(x) \wedge Q(x))$, é verdadeiro.

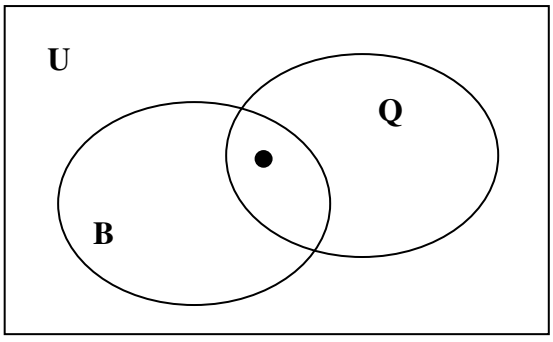


Figura 16: Representação da denotação de *A bicicleta quebrou* em termos de conjuntos.

[32] A lógica de predicados (LPO), conhecida também como cálculo de predicados de primeira ordem, estende o sistema da lógica proposicional, em as proposições são descritas como entidades atômicas. Na LPO, as proposições são analisadas em termos de predicados e argumentos, tendo o formato P (t1,…, tn) (um predicado com um ou mais "argumentos"). A LPO também utiliza vários operadores que designam operações lógicas e relações e permitem a construção de expressões mais complexas. Os operadores mais importantes são os operadores de quantificação universal (∀) (leia-se: "para todo x") e existencial (∃) (leia-se: "para algum x") (ALLWOOD et al., 1977, CRUSE, 2006).

Dessa forma, vê-se que as expressões recebem interpretações somente em um dado modelo ou universo de discurso. Para as sentenças em língua natural, a escolha de um modelo implica necessariamente na escolha de um contexto ou cenário. Assim, saber a denotação da sentença *A bicicleta quebrou* depende das condições fornecidas por U.

3.3.2. As RCs baseadas em *frames* e em redes semânticas

3.3.2.1. Abordagem teórica do significado: cognitiva

Os modelos de RC baseados em *frames* e em redes semânticas tratam o significado segundo a abordagem cognitiva, também denominada mentalista ou representacional.

Essa abordagem enquadra-se na subárea da Lingüística Cognitiva, em que Lakoff (1987) e Langacker (1987, 1991) são figuras de destaque. A Lingüística Cognitiva segue três princípios básicos: (i) a língua não é uma faculdade cognitiva autônoma; (ii) a gramática é conceitualização; e (iii) o conhecimento da língua emerge do uso da língua (TAYLOR, 1985; CROFT, CRUSE, 2004). Assim, a Lingüística Cognitiva se preocupa em determinar como a mente humana funciona, ou seja, como a mente recebe as informações advindas de fontes diversas (visual, motora, etc.) e as processa.

Na abordagem da semântica cognitiva, o significado subjacente às expressões lingüísticas é um **conceito**, o qual pode ser entendido como uma "descrição mental", uma idéia (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; ROSCH, MERVIS, 1975; TAYLOR, 1985; LÔBNER, 2002; CROFT, CRUSE, 2004; CRUSE, 2004).

A Figura 17 ilustra a reelaboração do triângulo semiótico de Ogden & Richards (1946) segundo as noções da abordagem cognitiva do significado, enfatizando a relação entre expressão lingüística, conceito e categoria.

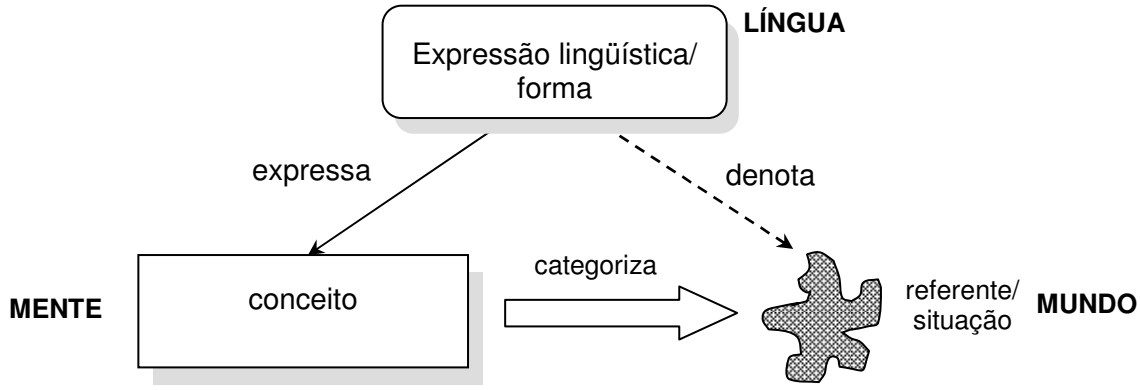


Figura 17: Relação entre língua, mente e mundo.

Mais especificamente, categorizar algo é percebê-lo como sendo de certo tipo. Ao olhar uma bicicleta, por exemplo, reconhece-se esse objeto como sendo desse tipo. Assim, a categoria BICICLETA consiste de todas as coisas que se categorizam como sendo desse tipo. Aos elementos que pertencem a uma categoria, destina-se a denominação membros ou exemplares.

A categorização só é possível, então, se a respectiva categoria estiver disponível no sistema cognitivo, isto é, na mente. Assim sendo, a categorização requer uma representação mental de uma categoria, um conceito.

Quando o falante se depara com um objeto novo, o sistema cognitivo gera uma descrição preliminar com base nas características perceptíveis do objeto, por exemplo: cor, tamanho, etc. Se a descrição for compatível com o conceito, o objeto será categorizado como tal. Esse conceito, segundo Rosch (1973) e Rosch e Mervis (1975), baseia-se no **protótipo** da categoria, ou seja, no exemplar prototípico da categoria. Em outros termos, um membro prototípico é aquele que apresenta o maior número de atributos comumente julgados como mais característicos da categoria.

Como salienta Allan (2001), intensão é, segundo a teoria dos protótipos, o conjunto de propriedades características de um membro prototípico de uma categoria.

Se uma categoria é definida por um protótipo, as condições que definem o protótipo não precisam ser condições necessárias para o resto da categoria. Para a categoria AVE, por exemplo, ter-se-ia uma representação mental (conceito) de uma entidade prototípica com as características: [animal], [pequeno] e [capaz de voar]. No entanto, há membros dessa categoria que não são capazes de voar, como dito anteriormente. De acordo com a teoria, isso não exclui pingüim da categoria AVE. Nesse caso, diz-se que os pingüins têm um grau de saliência menor ou são periféricos em relação ao protótipo da categoria AVE. Segundo Wittgenstein (1979), o agrupamento dos membros em uma categoria é feito por meio de similaridades parciais ou **semelhanças de família** (do inglês, *family resemblance*).

Com relação à definição de protótipo, ou seja, membro mais típico ou representativo de uma categoria, Lôbner (2002) salienta que seria mais adequado entendê-lo como um caso abstrato definido por um conceito que fixa certos traços. Assim, o conceito da bicicleta prototípica, por exemplo, conteria as propriedades mais salientes dos referentes. Neste ponto, é inevitável que se pergunte: quais são os traços que compõem o conceito? Segundo Lôbner (2002), os traços devem ser aplicáveis a uma grande parcela dos membros da categoria e a uma pequena parcela de não-membros. Por exemplo, [ter asas] e [botar ovos] são traços

compartilhados por vários tipos de animais e, por isso, não são "bons" traços para caracterizar o conceito <ave>. Já os traços [ter asas] e [voar] são mais distintivos.

A noção de protótipo, aliás, fica bastante evidente quando se trata das **categorias de nível básico** (do inglês, *basic level categories*) (ROSCH, 1973). As categorias podem, então, ser vistas em níveis, por exemplo: VEÍCULO – BICICLETA – VELOCÍPEDE. As categorias, como VEÍCULO, são chamadas **superordenadas** e as categorias como VELOCÍPEDE, **subordinadas** (CRUSE, 2004) (Figura 18).

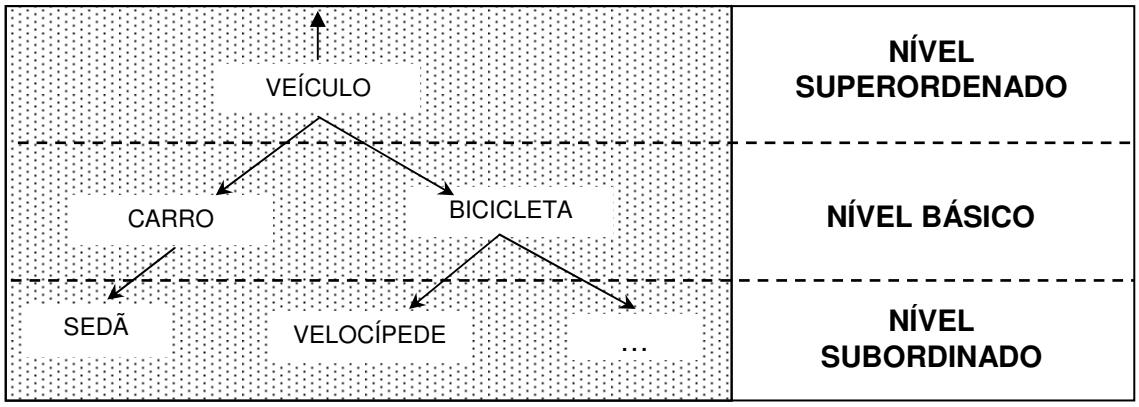


Figura 18: Níveis de categorização.

As categorias de nível básico, como BICICLETA, ocupam lugar de destaque no sistema cognitivo por várias razões: (a) os referentes são rapidamente reconhecidos; (b) os membros dessas categorias têm formato similar; (c) as categorias básicas visíveis permitem que se tenha uma figura ou imagem mental (ou seja, uma *gestalt*[33]) de seu conceito; (d) as categorias de nível básico são definidas pelo modo como se dá a interação com seus membros; para a categoria artefato, por exemplo, as partes características estão intimamente ligadas ao uso que se faz esses objetos (LAKOFF, 1987. CRUSE, 2004; CROFT, CRUSE, 2004).

O formato similar e o modo uniforme em que se dá a interação com os referentes/ membros das categorias de nível básico visíveis facilitam a determinação do protótipo e a elaboração do conceito (TAYLOR, 1985; LAKOFF, 1987; ALLAN, 2001; LÔBNER, 2002, CRUSE, 2004; CROFT, CRUSE, 2004). Aliás, o papel de destaque das categorias básicas no sistema cognitivo tem reflexo no nível lingüístico. As expressões lingüísticas dessas categorias tendem a ser mais curtas e simples.

Vale ressaltar, por fim, que, similar à noção de protótipo como *gestalt*, encontra-se a noção de **estereótipo**, proposta por Putnam (1975). Para esse autor, o conceito subjacente a uma

[33] O termo "gestalt" é emprestado do alemão e significa grosso modo "imagem mental única" (ALLAN, 2001).

expressão lingüística (tipicamente lexicalizada) não é composto por um conjunto bem-definido de propriedades necessariamente encontradas em todo referente, mas sim por um "conjunto mínimo de fatos estereotipados" sobre um referente típico. Como salienta Allan (2001), enquanto a noção de protótipo implica na escolha do "melhor exemplar" da categoria, a noção de estereótipo implica na escolha de um "exemplar típico".

3.3.2.2. A organização dos conceitos

Na abordagem representacional, defende-se amplamente a organização global dos conceitos. Em outras palavras, isso quer dizer que os conceitos não estão isolados na mente ou organizados aleatoriamente. Ao contrário, há relações entre os conceitos. De um modo geral, pode-se identificar dois tipos de relação que se estabelecem entre os conceitos.

Segundo Cruse (1986, 2004), por exemplo, os conceitos estariam organizados principalmente pelas relações **paradigmáticas**. Tais relações refletem as escolhas semânticas disponíveis em um ponto particular da sentença, por exemplo, (1) *Eu quero um copo de* ____ (vinho, cerveja, leite, etc.). Dentre elas, citam-se as relações que expressam inclusão.

A relação de hiponímia é normalmente vista como sendo de inclusão e como a relação estrutural mais importante. Essa relação é aquela que se estabelece entre, por exemplo, os conceitos <veículo> e <bicicleta>. Nesse caso, diz-se que <bicicleta> é hipônimo de <veículo> e <veículo> é hiperônimo de <bicicleta>.

Do ponto de vista extensional, diz-se que a classe denotada por hiperônimo inclui a classe denotada pelo hipônimo. Dessa forma, a classe dos veículos inclui a classe das bicicletas. Do ponto de vista intensional, diz-se que o conceito <bicicleta> é mais rico que o conceito <veículo>. Vale ressaltar que, do ponto de vista lógico, a hiponímia é uma relação transitiva, ou seja, se A é um hipônimo de B e B é um hipônimo de C, então A é necessariamente hipônimo de C (A=velocípede, B=bicicleta, C=veículo) (CRUSE, 2004).

Muitas relações de hiponímia também são relações taxonômicas. Esse é o tipo, por exemplo, da relação entre <bicicleta> e <veículo>, a qual pode ser caracterizada como "é um tipo de" (isto é, bicicleta é um tipo de veículo). Isso não ocorre, por exemplo, entre <garanhão> e <cavalo>, já que a relação entre esses conceitos não pode ser caracterizada como (<garanhão>) "é um tipo de" (<cavalo>). Cruse (2004) salienta que as relações taxonômicas são um subtipo de hiponímia.

Outra relação de inclusão é a "parte de", cuja relação refletida na língua é denominada meronímia. Essa relação é a que se estabelece entre os conceitos <mão> e <dedo>,

<telescópio> e <lentes>, por exemplo. Nesse caso, diz-se que <dedo> é merônimo de <mão> e <mão> é holônimo de <dedo>.

As relações de hiponímia e meronímia e outras são a base para a estruturação dos conceitos nos modelos de representação do conhecimento baseados em redes semânticas. Tais relações estruturam, por exemplo, as redes lexicais em formato *wordnet*, que seguem o desenvolvimento da base da WN.Pr (FELLBAUM, 1998). Nessa base, o conceito codificado pelo *synset* {car; auto; automobile; machine; motorcar}, por exemplo, está relacionado ao conceito mais geral codificado pelo *synset* {motor vehicle; automotive vehicle}.

Segundo autores como Croft e Cruse (2004), por sua vez, os conceitos não estão organizados em função de relações como hiponímia e meronímia. Sob esse ponto de vista, aliás, Fillmore propôs a Semântica de Frames (FILLMORE, 1976). Segundo essa teoria, os conceitos organizam-se em função de **frames semânticos**, ou seja, esquemas de situações típicas, em que participantes, propriedades, papéis e outros elementos conceituais são estabelecidos (FILLMORE, WOOTERS, BAKER, 2001). Os frames semânticos também podem ser entendidos como padrões globais de conhecimento de senso comum sobre conceitos centrais, de tal forma que um item lexical denotando o conceito evoca todo o frame semântico (BEAUGRANDE, DRESSLER, *apud* TAYLOR, 1985). No conhecido exemplo de Schank e Abelson (1977), o conceito <restaurante> está associado a outros conceitos, os quais são denominados elementos-frame, como <freguês>, <pedido>, <conta>, <garçom>, etc.

Os frames podem representar também seqüências de eventos. Nesse caso, os *frames* são nomeados **frames dinâmicos** ou **scripts**. Um *script* define uma seqüência de eventos em função de protótipos para as pessoas, propriedades e objetos envolvidos na seqüência de eventos (HELBIG, 2006). Em outras palavras, um *script* é a composição de uma seqüência de ações que se repetem regularmente em situações similares e são representadas como um esquema padronizado (*frame*). Por exemplo, o significado de *comer em um restaurante* pode envolver os (i) eventos de *entrar, fazer o pedido, comer, pagar* e *sair* e os (ii) papéis de *freguês, garçom* e *caixa* (ALLAN, 2001).

Os frames semânticos são responsáveis pela organização dos conceitos nos modelos de representação do conhecimento baseados em *frames.* No PLN, os frames semânticos constituem a base de organização de um recurso lingüístico-computacional do AmE bastante reconhecido, a base FrameNet (BAKER et al, 1998). Nessa base, o conceito <frasco> (do inglês, *bottle*) está associado a elementos-frame como [conteúdo], [material], [possuidor], [parte], entre outros.

3.3.2.3. Outras abordagens do significado: estruturalista e pragmática

Antes de se enfatizar as tecnologias de representação utilizadas pelos modelos de RC baseados em *frames* e em redes semânticas, salienta-se que, além das abordagens do significado referencial e cognitiva, há outras duas: a estruturalista e a pragmática (TAYLOR, 1985; CHIERCHIA, McCONNELL-GINET, 1990; CHIERCHIA, 2003; CRUSE, 2004). Essas duas abordagens são descritas a seguir para contextualizar a abordagem referencial e cognitiva do cenário mais amplo dos estudos semânticos.

A abordagem estrutural do significado, desenvolvida no âmbito do Estruturalismo de Saussure (1999)[34], engloba o pressuposto de que a língua é um sistema abstrato e complexo de relações e regras que subjazem às regularidades a serem observadas no uso real da língua. Tal sistema é formado pelos signos, que se relacionam de várias maneiras. Um signo, no caso, uma palavra, consiste de duas partes. A primeira é a forma (ou significante) e a segunda é o significado. A associação entre significante e significado é arbitrária. Não há nenhuma razão, além da convenção, que una, por exemplo, a forma fonética [kaza] ao significado específico <casa>; qualquer forma aceita pelos falantes poderia ser associada a esse significado.

Segundo Culler (1976), há outra leitura que se pode fazer a respeito da arbitrariedade dos signos. Segundo esse autor, Saussure concebe que o significado propriamente dito é arbitrário. Para Saussure, não há significados pré-existentes, independentes da língua. Nesse sentido, não há nenhuma razão para que determinada porção do espaço das cores, por exemplo, seja lexicalizada (TAYLOR, 1985).

A arbitrariedade do signo lingüístico está intimamente ligada a outro princípio saussureano: o de que a língua é um sistema autônomo. Assim, o significado não é uma propriedade do signo, mas uma função do valor do signo dentro do sistema lingüístico ao qual pertence. Em outras palavras, o significado é definido "contrastivamente", isto é, pela comparação com elementos da mesma ordem, signos. Dessa forma, apesar de a palavra *vermelho* ser devidamente usada pelo falante da língua para se referir às propriedades do mundo e evocar provavelmente uma imagem mental das mesmas, o significado da palavra não é dado por nenhuma propriedade do mundo e também não reflete qualquer ato cognitivo não-lingüístico por parte do falante (LÔBNER, 2002). O significado de *vermelho* resulta do valor da palavra dentro do sistema semântico das cores nomeadas no PB e o que é externo a esse sistema não representa nada para a determinação do significado de vermelho. O fato de

[34] Tradução do texto original em francês *'Cours de linguistique général'*, publicado em 1916.

essa língua ter "palavras" como *laranja* (<laranja>)*, rosa* (<rosa>) e *violeta* (<violeta>) limita o campo denotacional de *vermelho* (<vermelhor>).

Resumidamente, de acordo com a semântica estrutural, o significado de um signo é o resumo das diferenças e relações deste com o significado dos demais signos. Nessa abordagem, a denotação seria apenas algo secundário em relação à combinatória dos signos. Assim, o significado não é uma questão de sentido e denotação; o interesse está no signo, ou seja, no par forma/ significado, e no modo como ele se relaciona com os demais signos.

Nessa abordagem, há três relações distintas que devem ser consideradas. A primeira delas é o relacionamento de similaridade semântica que há na base dos campos lexicais; a análise desse tipo de similaridade foi inaugurada por Trier e desenvolvida posteriormente pelos antropólogos americanos Goodenough (1956), Lounsbury (1956) e pelo europeu Pottier (1974) (BORBA, 1991). A segunda são as relações de sinonímia, antonímia e hiponímia, que foram sistematicamente selecionadas como a base da semântica estrutural por Lyons (1963). A terceira são as relações sintagmáticas identificadas por Porzig (PORZIG, 1934, apud GEERAERTS, 2006). Essas relações reapareceram posteriormente como as restrições selecionais na semântica "neo-estrutural" e na gramática gerativa de Katz e Fodor (1963). Segundo Geeraerts (2006), o termo "neo-estrutural" recobre os trabalhos contemporâneos descendentes do estruturalismo. No que diz respeito à semântica, citam-se: a Teoria dos Campos Semânticos desenvolvida por Coseriu (COSERIU, GECKLER, 1981); o trabalho de Cruse (CRUSE, 1986) sobre as relações de sentido[35]; os trabalhos de Halliday e Hasan (HALLIDAY, HASSAN, 1976) e de Sinclair (SINCLAIR, 1987) a respeito da concepção sintagmática das relações lexicais.

Dentro da perspectiva estrutural, um dos enfoques a respeito do significado e da análise semântica que se encaixa nos pressupostos da teoria "clássica" do significado é a Análise Componencial (LYONS, 1979). De acordo com o enfoque componencial, a estrutura do significado de uma categoria (p. ex. VACA) se organiza em termos de traços ou componentes necessários e suficientes (essenciais), compartilhados por todos os seus membros. (p. ex. [fêmea], [adulto], [bovino] etc.). Esses traços distinguem uma categoria das outras, dentro de um mesmo campo semântico (p. ex. *touro*, *novilho*, *bezerro*) (LEHRER, 1974).

No âmbito dos estudos gerativistas, Katz e Fodor (1963) e Katz e Postal (1964) adotaram a análise componencial também como um aparato formal (GEERAERTS, 2006).

[35] No trabalho mais recente de Cruse (2004), entretanto, as origens estruturalistas do autor foram atenuadas pela adoção do ponto de vista da Semântica Cognitiva.

Na abordagem do significado denominada **pragmático-social** ou **sócio-pragmática**, renuncia-se à visão de significado como entidade, em benefício de uma abordagem totalmente pragmática (cf. AUSTIN, 1965; GRICE, 1975; SEARLE, 1979; LEECH, 1983).

Essa abordagem qualifica o significado como uma práxis social, assimilando-o à maneira como as expressões são usadas. Aliás, Wittgenstein, em suas *Investigações Filosóficas* (WITTGENSTEIN, 1979) já falava das regras e dos acordos, ou **jogos de linguagem,** que se deve seguir ao se engajar nas práticas sociais. Tais regras não são individuais, mas sim públicas e se estabelecem pelo contexto de uso entre os sujeitos da interação, do jogo. Essas regras são acordadas socialmente não existindo um significado *a priori* do uso. O uso das palavras e as regras dos jogos de linguagem determinam o significado. De fato, há tantas significações possíveis quantos contextos possíveis (FAUSTINO, 1995; GLOCK, 1998).

Na esteira da reflexão de Wittgenstein sobre os jogos de linguagem, reflexão essa que tende a sublinhar a diversidade de ações realizadas pelos falantes no e pelo uso quotidiano da linguagem, a abordagem pragmático-social tende a privilegiar a dimensão comunicativa da linguagem em detrimento da sua função de representação do mundo. Assim, confere-se particular relevo ao uso da linguagem num processo de interação: falar é agir intencionalmente em função de certos objetivos e de acordo com regras de natureza contratual e institucional.

É justamente nesta linha de investigação do fenômeno lingüístico que se inserem os trabalhos de Austin (1965) sobre os atos de fala, bem como o aprofundamento posterior desenvolvido por Searle (1979), com a sua tipologia das modalidades das ações verbais.

Destaca-se também a figura de Grice (1975), segundo a qual a linguagem é um instrumento utilizado pelo locutor para comunicar ao seu destinatário suas intenções, nas quais está embutido o significado e a comunicação verbal organiza-se pelo **Princípio da Cooperação**, cuja pressuposição é a de que os indivíduos que se comunicam, então constroem enunciados, obedecendo a quatro máximas conversacionais: **verdade**, **quantidade, relevância** e **modo**. Pela primeira máxima, pressupõe-se que tudo que o interlocutor diz é verdadeiro; pela segunda, que ele só diz o necessário; pela terceira, que só diz o que é pertinente para aquela situação de comunicação e, por fim, o faz do melhor modo possível. O Princípio da Cooperação de Grice assenta-se, principalmente, na suposição de que há uma lógica que orienta o fluxo conversacional. Essa hipótese, aliás, está expressa no título de seu mais famoso texto Lógica e Conversação (do inglês, *Logic and Conversation*) (GRICE, 1975), em que admite a idéia de conhecimento e de significados compartilhados e de comunicação harmônica entre os sujeitos. Quando se quebram as regras desse "jogo lógico",

ocorre o que o autor denomina **implicatura conversacional**, ou seja, o uso de expressões que rompem a máxima da verdade, quantidade, relevância e modo.

Resumidamente, a abordagem pragmático-social dá destaque ao aspecto da língua enquanto uma forma de agir sobre o mundo, focalizando o processo de comunicação verbal.

3.3.2.4. A metalinguagem formal dos modelos baseados em frames

Do ponto de vista formal e interno, um *frame* é uma estrutura **atributo-valor**, composta por atributos (do inglês, *slots*) que são preenchidos com valores (do inglês, *fillers*) apropriados (MINSKY, 1975; HANDKE, 1995), assim como ilustrado na Figura 19.

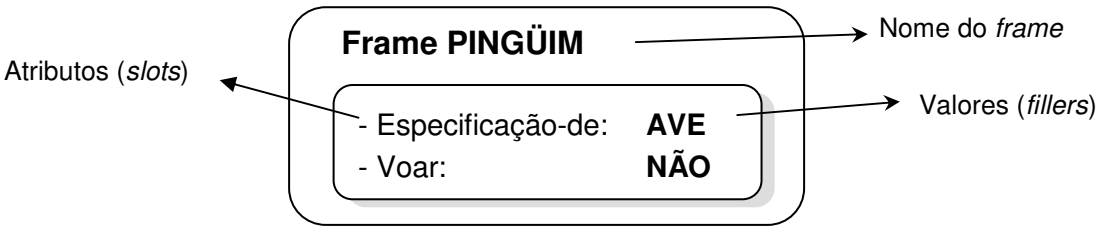


Figura 19: Elementos da estrutura interna de um *frame*.

Do ponto de vista da organização global, um *frame* está relacionado a outros (subjacente, por exemplo, a um objeto, um fato ou um evento) em uma hierarquia de *frames.* A Figura 20, baseada em Handke (1995, p. 100), ilustra o *frame* subjacente à expressão *pingüim* e sua relação com outro *frame*.

O *frame* ilustrado na Figura 19 possui dois tipos de atributos. O atributo Especificação-de é usado para estabelecer a relação de hiponímia entre o *frame* em questão e seu *frame* hiperônimo, no caso, AVE. Esse atributo é necessário para a classificação de um *frame* e essencial para estabelecer a hierarquia entre os *frames*. Além desse atributo central, mais atributos podem ser especificados. No caso, o *frame* em questão descreve o atributo Voar, que apresenta um *valor padrão* (do inglês, *default value*) [não].

Os valores-padrão de um *frame* são especificados com base na noção de protótipos. Por essa razão, diz-se que os *frames* podem ser considerados como uma formalização da teoria dos protótipos (HANDKE, 1995).

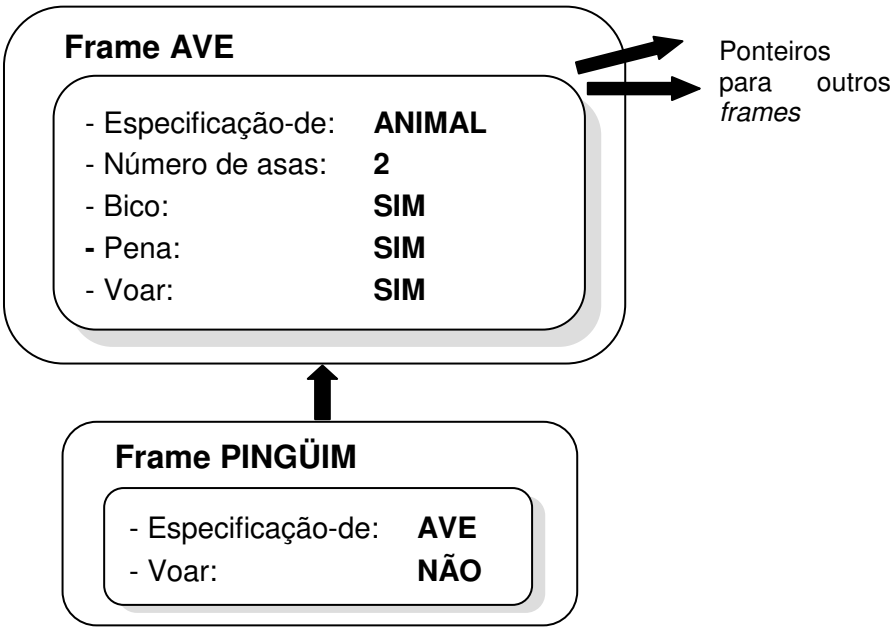


Figura 20: Conexões entre *frames*.

Com base na noção de protótipo, pode-se especificar, por exemplo, que o *frame* AVE, ilustrado na Figura 20, tem os seguintes atributos e valores-padrão: [Número de asas: 2; Bico: SIM; Pena: SIM; Voar: SIM]. O *frame* PINGÜIM, enquanto hipônimo do *frame* AVE, herda os atributos e valores deste. O mecanismo de herança empregado nas representações baseadas em *frames* é dito não-monotônico, ou seja, a informação de um *frame* genérico é herdada até que não haja outra informação disponível nos *frames* mais específicos. O *frame* PINGÜIM, por exemplo, especifica o seguinte par atributo-valor, [Voar: NÃO], cujo valor é conflitante com o valor do *frame* hiperônimo [Voar: SIM]. Quando ocorre esse tipo de conflito, considera-se a informação mais específica. No caso, sobrescreve-se o valor *padrão* do *frame* AVE pelo valor do *frame* PINGÜIM.

Do ponto de vista da teoria dos protótipos, esse conflito é gerado pelo fato de os pingüins não serem membros prototípicos da categoria das aves e, por isso, não apresentarem todas as propriedades características das aves.

A metalinguagem dos *frames* foi utilizada no projeto Cyc (do inglês, *en**cyc**lopedia*), que teve início em 1984 e ainda está em andamento. Nesse projeto, propõe-se a construção de uma base de conhecimentos, para a língua inglesa, que contenha uma grande quantidade de entradas provenientes do conhecimento enciclopédico (LENAT, GUHA, 1990; WILKS et al., 1996). Essa base de conhecimento tem aplicação direta na computação; segundo Copeland (2000) e Helbig (2006), trata-se do maior projeto existente em Inteligência Artificial para a aquisição de uma base de conhecimento.

Sua metalinguagem envolve: (i) uma base de conhecimentos (KB); (ii) um motor de inferências baseado na dedução lógica; (iii) uma linguagem de representação que serve como *interlíngua*, a CycLanguage; e (iv) um Subsistema de Processamento de Língua Natural. A base de conhecimento do projeto Cyc organiza atualmente uma vasta quantidade de entidades e relações. Segundo informações disponíveis no site do projeto[36], essa base registra aproximadamente duzentos mil termos e várias dezenas de asserções inseridas manualmente para cada termo, além de outros milhões de relações estabelecidas automaticamente a partir de processos de inferências. As entidades do Cyc foram extraídas de diversos textos on-line (histórias, artigos de enciclopédias, base de dados sobre cinema, etc.) e convertidas para a interlíngua Cyc e, em seguida, incluídas na base.

Ao menos na descrição fornecida por Lenat e Guha (1990), a base do Cyc é um sistema de *frames*. Segundo Helbig (2006), um dos maiores problemas do Cyc diz respeito à relativa distância em relação à língua natural e à ausência de adequação cognitiva na especificação das descrições dos *slots* (ou descrição dos predicados). Por exemplo, os rótulos computersFamiliarWith ou objectFoundInLocation (LENAT, GUHA, 1990) apresentam algum significado para os humanos, mas o encaixe de tais informações na estrutura dos conceitos gerais não é clara. Esse problema, aliás, dificulta consideravelmente o estabelecimento de uma interface dos conceitos com as línguas naturais (principalmente no processo de construção de léxicos lingüístico-computacionais) e, por conseguinte, o emprego dessa base em sistemas de processamento automático de línguas naturais (MAHESH et al., 1996).

3.3.2.5. A metalinguagem formal dos modelos baseados em redes semânticas

As redes semânticas (RSs) constituem todo um paradigma de representação do conhecimento. Nesse paradigma, a representação dos conceitos é feita por **nós** e a das relações por **arcos** rotulados entre os nós (HANDKE, 1995). A Figura 21 ilustra os construtos básicos de uma rede semântica, ou seja, os nós (ou conceitos, C) e os arcos (relações, REL), que serão amplamente discutidos na Seção IV.

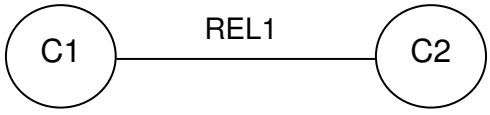


Figura 21: Ilustração dos construtos básicos de uma rede semântica.

[36] http://www.cyc.com/

As redes semânticas têm uma longa tradição na área de representação do conhecimento. Elas se tornaram conhecidas por meio do trabalho do botânico Ritchens, que em 1956 propôs o formalismo (HUTCHINS, 1986).

As justificativas psicológicas para esse tipo de representação foram configuradas na década de 1960, por meio do trabalho colaborativo do cientista da computação A. M. Collins e do psicólogo M. R. Quillian (HANDKE, 1995).

O primeiro a considerar que esse tipo de representação reflete a organização da memória humana foi Quillian (1967, 1968), que propôs um modelo computacional da memória humana chamado **memória semântica**. Segundo o autor, esse modelo, em que conceitos são representados por nós, e relações entre conceitos, por arcos, explicaria diversos resultados experimentais sobre o comportamento da memória humana. Explicaria, por exemplo, o fato de que o reconhecimento de objetos que pertencem a classes mais numerosas requer mais tempo do que o reconhecimento dos pertencentes a classes menos numerosas (COLLINS, QUILLIAN, 1969, WILKINS, 1971; SMITH, 1978; MILLER, CHARLES, 1991).

Nos testes feitos por Quillian e reportados por Collins e Quillian (1969), assumiu-se o tempo de reação para indicar o número de níveis hierárquicos que separavam dois conceitos. Eles observaram, por exemplo, que os sujeitos dos experimentos levavam mais tempo para responder se é verdade que “um canário pode cantar” que para responder se “um canário pode voar”, e ainda mais tempo para responder se “um canário tem pele”. Nesse exemplo, assume-se que [cantar] é armazenado na memória como um traço de canário, [voar] é armazenado como traço de pássaro e [ter pele], como traço de animal. Segundo os autores, se esses três traços estivessem diretamente associados a canário, o tempo de resposta às perguntas seria o mesmo.

A Figura 22 ilustra um fragmento de uma RS em que, por questão de simplicidade, as relações entre os nós são especificadas por 2 tipos de arcos, rotulados por “é um (tipo de)”, que expressa a relação taxonômica, e por “tem”, que expressa a relação de meronímia.

Em uma RS, como a ilustrada na Figura 22, os conceitos estão organizados hierarquicamente; nessa hierarquia, há um nó superior ao qual estão ligados os nós filhos, os quais, por sua vez, também têm outros conceitos como filhos e assim sucessivamente (BRACHMAN, 1979). Como salienta Dias-da-Silva (1996), cada nó pode representar um tipo ou subtipo semântico (ou conceitos) e cada arco direcionado pode não só representar relações “é-um (tipo de)” ou “é parte de”, mas também relações temáticas (agente, paciente, etc.) que se estabelecem entre predicados e seus argumentos (cf. Figura 23).

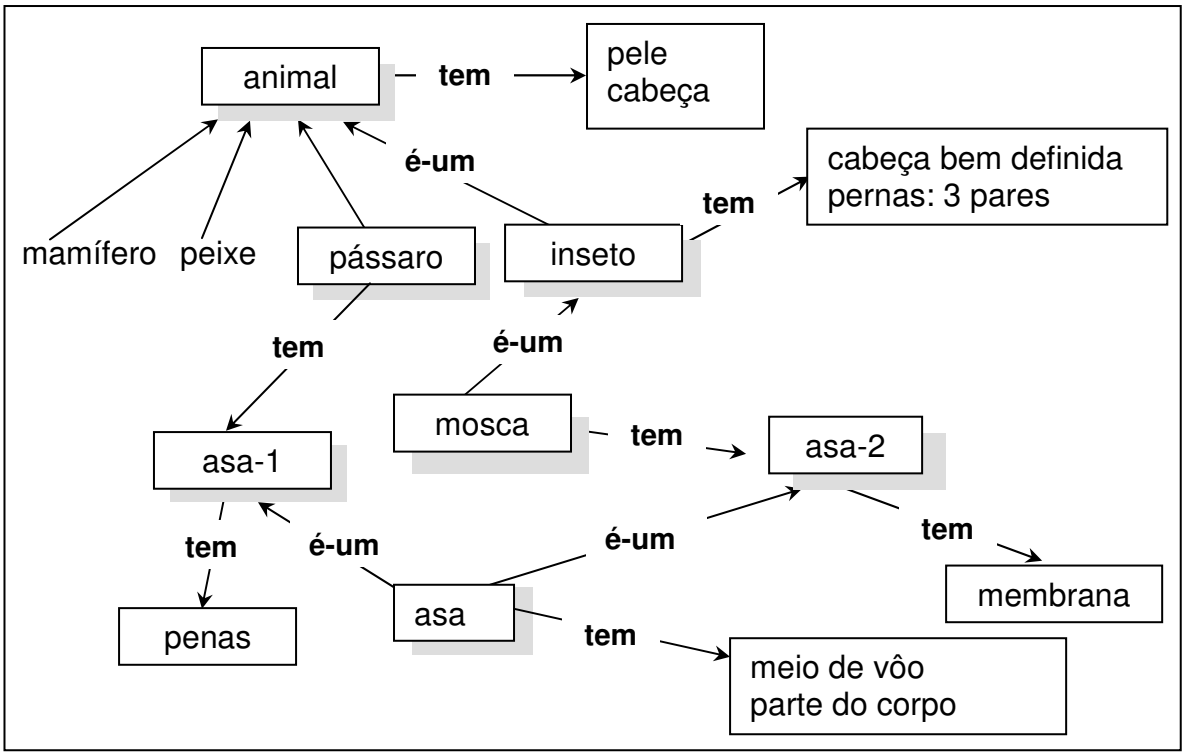


Figura 22: Exemplo de um fragmento de uma rede semântica.

Outro aspecto importante desse formalismo é que, além de permitir a representação de tipos de estruturas, as redes semânticas permitem representar estruturas realizadas. Antes, porém, de se apresentar essa possibilidade, destacam-se os níveis de representação (ou domínios) propostos por Jackendoff (1983). O autor assume que a estrutura do pensamento humano organiza-se em função de quatro domínios – **mundo real, mundo projetado** (ou mundo da experiência)**, estrutura conceitual** (ou conceitos) e **expressões lingüísticas** (ou formas), a partir dos quais o autor discute as relações entre as expressões lingüísticas e a realidade.

Ao se considerar que o "micromundo" a ser construído em um sistema de PLN corresponde a uma espécie de mundo projetado, é possível estabelecer que a expressão lingüística *bicicleta* (forma) é um tipo de <bicicleta> (conceito) e que uma realização concreta desse conceito como em, por exemplo, *Eu quero sua bicicleta,* é #bicicleta77# (referente) (DIAS-DA-SILVA, 1996). Dessa forma, as redes semânticas têm a possibilidade de suportar a especificação do significado intensional e extensional. Em outras palavras, é possível estabelecer a relação entre a expressão lingüística, o conceito (significado potencial) e um possível referente, assim como ilustrado na Figura 23, elaborada com base em Dias-da-Silva (1996).

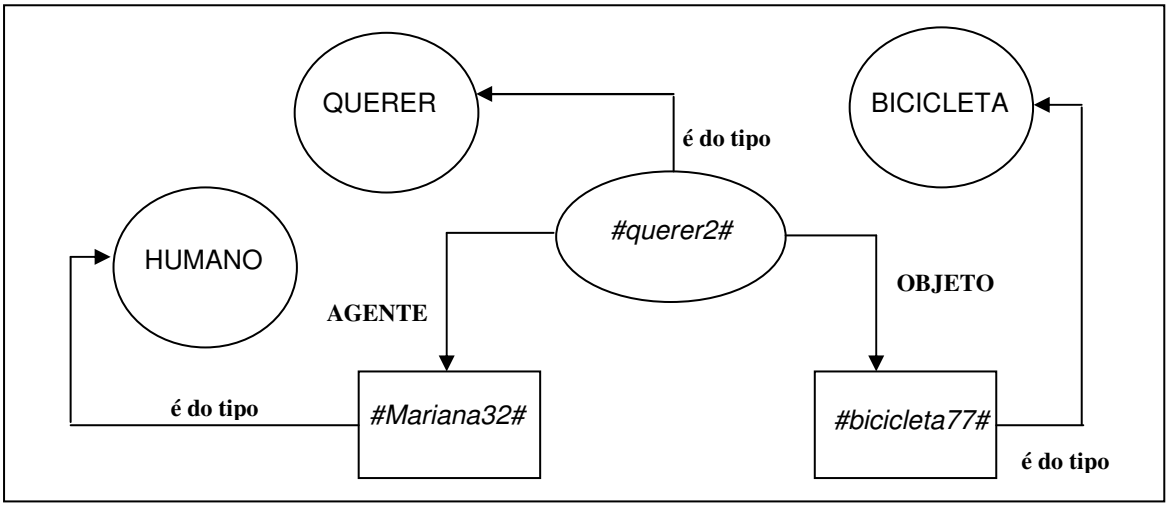


Figura 23: Rede semântica que representa parte do significado expresso pela frase *Mariana quer a bicicleta*.

Segundo Handke (1995), dois fatores foram fundamentais para a popularidade das redes semânticas como modelo de representação do conhecimento. O primeiro deles é a elegância com que deduções podem ser feitas; por exemplo, para se deduzir que uma mosca tem 3 pares de pernas, é preciso apenas seguir a hierarquia "é um (tipo de)", assumindo-se que os atributos associados ao nó superior, no caso, associados a <inseto>, são válidos para nós inferiores, no caso, o conceito <mosca>. O segundo fator de popularidade é a simplicidade e elegância com que as RSs podem ser implementadas.

No final da década de 70, duas linhas de desenvolvimento podem ser observadas dentro do paradigma das RSs.

Em uma dessas linhas, enquadram-se, por exemplo, as *Redes de hierarquia estruturada* (do inglês, *Structured Inheritance Networks* - SINs) (BRACHMAN, SCHMOLZE, 1985). As SINs, cujo representante mais típico é a KL-ONE (do inglês, *Knowledge Language One*), baseiam-se essencialmente nos trabalhos de Brachman e Schmolze (1985). As SINs utilizam dois construtos básicos: os conceitos (genéricos ou individuais) e os papéis, sendo que a definição dos papéis é o ponto de partida para a organização global dos conceitos. Nas SINs, as relações hierárquicas entre os conceitos desempenham papel fundamental, daí a origem de sua denominação. As SINs utilizam as seguintes categorias para estruturar o conhecimento: papéis ou atributos definitórios, restrições sobre os papéis e diferenciação, subsunção e classificação dos papéis.

A Figura 24, elaborada com base em Helbig (2006) e Brachman e Shmolze (1985), ilustra a organização hierarquicamente estruturada segundo o formalismo KL-ONE. A Figura 24 tem

4 elipses, que representam os conceitos, e cinco arcos com um círculo no meio, que indicam papéis. A notação *v/r* indica *restrições* (do inglês, *value restrictions*) que os conceitos aplicam aos papéis.

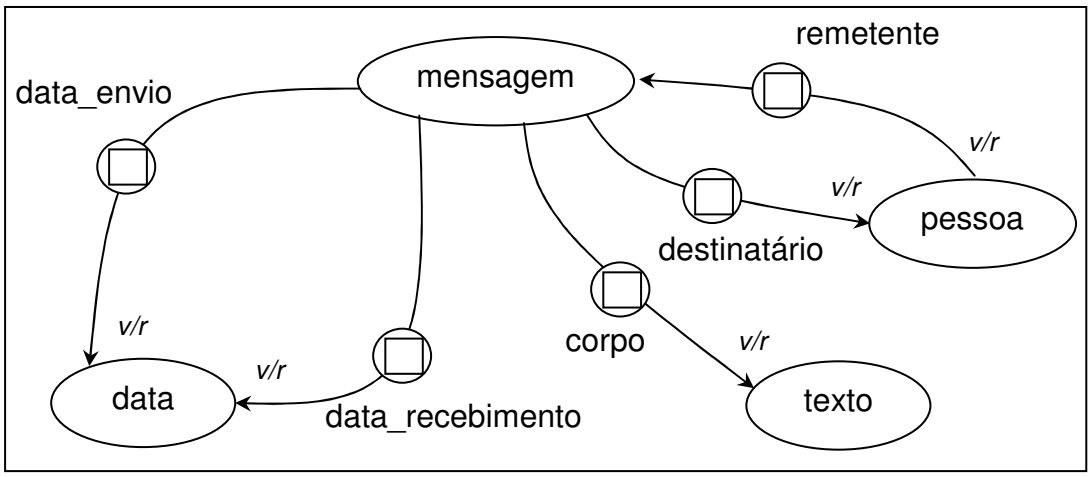


Figura 24: Fragmento da rede de hierarquia estruturada KL-ONE.

Nessa Figura, nota-se, por um lado, que remetente e destinatário são representados como relações e, portanto, como arcos, e pessoa, por outro lado, é representado como um conceito, ou seja, um nó. Dessa forma, os conceitos podem ser representados tanto por nós (conceitos) como por arcos (papéis). Não há, no entanto, um limite claro entre os conceitos que devem ser representados por nós e aqueles que devem ser representados por arcos. Aliás, a especificação dos tipos de arcos é discutida por Woods (1975), que salienta sua relevância na elaboração de uma rede semântica. Helbig (2006) classifica essa arbitrariedade na especificação dos nós e arcos como uma espécie de inadequação cognitiva.

Na outra linha de desenvolvimento do paradigma baseado em redes, enquadra-se, por exemplo, o paradigma das *Redes semânticas estendidas estratificadas* (do inglês, *Multilayered Extended Semantic Networks* – MultiNet) (HELBIG, 2006). Essa linha foi influenciada pelas pesquisas advindas da Ciência Cognitiva, em especial, de Miller e Johnson-Laird (1976). Os formalismos desenvolvidos nessa segunda linha de pesquisa são denominados “redes semânticas cognitivamente motivadas” (do inglês, *cognitively-oriented semantic networks* - CSNs).

3.4. Síntese da Seção III

Ao assumir que um sistema de PLN é um tipo de sistema especialista, assume-se, conseqüentemente, que as pesquisas nessa área envolvem uma espécie de "engenharia do conhecimento lingüístico". Como tal, as pesquisas em PLN podem usufruir dos modelos de RC da área da Engenharia do Conhecimento (DIAS-DA-SILVA, 2006).

Os modelos de RC fornecem, ao mesmo tempo, o arcabouço teórico-metodológico segundo o qual o objeto é investigado e a metalinguagem formal para a descrição do conhecimento.

Quanto às abordagens teóricas do significado subjacentes aos paradigmas de RC descritos nesta seção, salienta-se que a variedade de abordagem mostra que o estudo do significado pode ser feito de vários ângulos. Os modelos baseados na lógica clássica, por exemplo, focalizam a relação entre expressões lingüísticas e mundo (abordagem extensional), enquanto os modelos baseados em redes semânticas e *frames* focalizam a relação entre expressões lingüísticas e representações mentais (abordagem cognitiva). Provavelmente, essas perspectivas não são totalmente incompatíveis, pois o significado possui realmente várias facetas.

Ao se adotar determinado modelo ou paradigma de RC, adota-se conseqüentemente um conjunto específico de princípios fundamentais que geram o estudo do conhecimento, no caso, semântico, e uma tecnologia ou metalinguagem formal para a descrição desse conhecimento.

Na próxima Seção, apresenta-se o modelo de RC denominado MultiNet, o qual inaugura um novo paradigma de representação de conhecimento. Apresentam-se, especificamente, os critérios de elaboração desse paradigma e suas principais características, as quais, aliás, fundamentam a escolha desse modelo enquanto norte teórico-metodológico e metalinguagem formal para a descrição dos conceitos lexicalizados.

Seção IV

O paradigma MultiNet: uma linguagem de representação do conhecimento lingüisticamente codificado

Nesta Seção, apresenta-se o paradigma de representação do conhecimento codificado em língua natural denominado **MultiNet** (do inglês, *"**Multi**layered Extended Semantic **Net**works"*) (HELBIG, GNÖRLICH, 2002; HELBIG, 2006), destacando os critérios segundo os quais tal modelo de representação foi proposto, seus principais recursos expressivos e os recursos específicos para a ligação entre a representação de conhecimento e o léxico de uma língua natural.

4.1. Introdução

O **MultiNet** foi proposto por Hermann Helbig (1997, 2006) após anos de trabalho nas áreas de Representação do Conhecimento e Processamento Automático das Línguas Naturais. Inicialmente, o acrônimo para ***Multi**layered Extended Semantic **Net**works* era **MESNET** (HELBIG, 1995, 1997), mas, em seguida, passou a ser MultiNet (HELBIG, 2002), o qual permaneceu. Esse modelo liga-se ao paradigma das representações baseadas em redes semânticas, nos moldes propostos por Quillian (1967, 1968). Na verdade, o MultiNet é um tipo específico de **rede semântica**, ou seja, um modelo matemático de representação de estruturas conceituais que descreve conjuntos de conceitos e conjuntos de relações entre eles. Na forma de grafo, esse tipo de rede semântica codifica os conceitos nos nós e as relações entre os conceitos nos arcos, assim como exemplifica a Figura 25. Essa Figura ilustra, de modo simplificado, a rede semântica subjacente à "Napoleão perdeu a batalha de Waterloo".

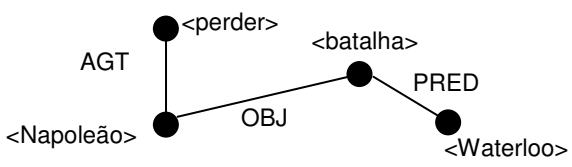


Figura 25: Fragmento de uma rede semântica nos moldes do MultiNet.

Os **conceitos,** representações mentais do mundo, representam o significado subjacente às expressões lingüísticas. Dessa forma, os conceitos e as relações entre eles são elementos essenciais do aparato cognitivo. Um conceito caracteriza-se basicamente por uma expressão lingüística que o denomina, por um conjunto de relações estabelecidas com outros conceitos e por um padrão complexo de origem visual (ou imagem) (HELBIG, 2006).

Para elucidar o estatuto epistemológico do MultiNet, utiliza-se a Figura 26, elaborada com base em Helbig (2006).

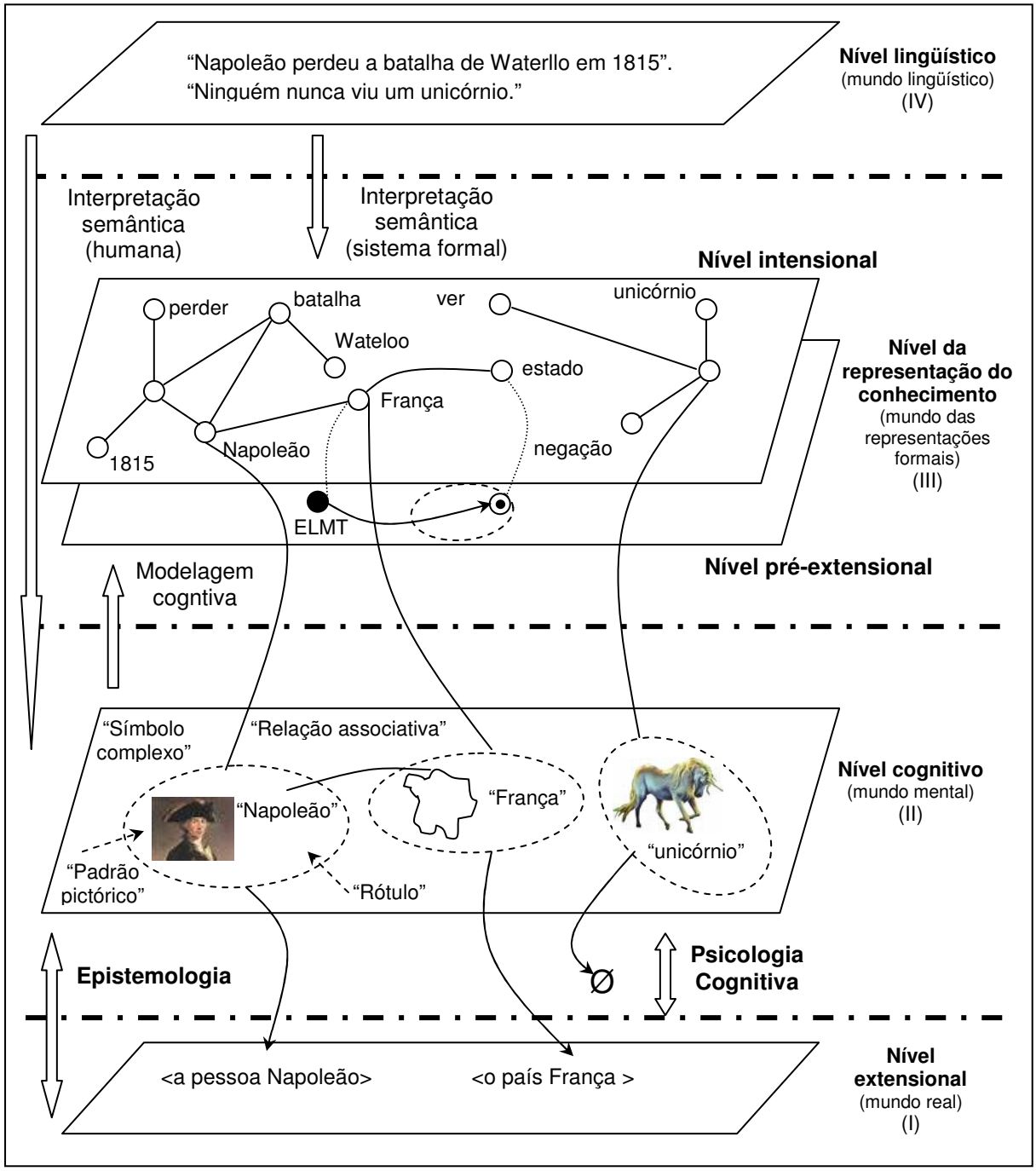


Figura 26: O lugar da representação do conhecimento nos diferentes "mundos".

Por meio dessa Figura, observa-se que o paradigma assume quatro níveis de "realidade". Os dois níveis inferiores, os Níveis I e II, são referentes, respectivamente, ao "mundo real" (fora do aparato mental) e ao "mundo mental". Assim, o Nível II, em especial, representa a memória humana e o seu conteúdo. A relação entre os Níveis I e II, tópico de pesquisa da Filosofia e da Psicologia Cognitiva, não é objeto de investigação do paradigma MultiNet. O MultiNet, como um tipo de rede semântica, busca modelar o Nível II e sua interrelação com o Nível IV ("mundo lingüístico"), já que ele foi criado para representar o conhecimento semanticamente registrado por meio de línguas naturais.

O MultiNet, que se enquadra no Nível III, distingue dois planos de representação básicos: o "plano intensional" e o "plano pré-extensional". O primeiro modela as relações entre os conceitos (relações intensionais) e o segundo modela os conceitos e suas extensões (relações extensionais). No plano pré-extensional, são modeladas, por exemplo, a cardinalidade das extensões e as relações entre essas extensões, modeladas pela teoria dos conjuntos. Esse plano é importante para a interpretação de conceitos quantitativos como <todos os x exceto y> ou <três deles>.

Diz-se "pré-extensional" porque os aspectos puramente extensionais do mundo real são modelados no Nível II (cf. Figura 26). Assim, é nos planos intensional e pré-extensional que os aparatos da representação do conhecimento do MultiNet operam, sendo que no plano pré-extensional são representadas, de modo indireto, as entidades do mundo real. Além disso, no MultiNet, é a combinação das representações dos planos pré-extensional e intensional que fornece a representação completa dos conceitos. A conexão entre as representações do MultiNet e as expressões lingüísticas é feita em termos da especificação da relação semântica ("interpretação semântica" da Fig. 26) entre os construtos do MultiNet e as expressões lingüísticas.

Cabe destacar que a elaboração do paradigma MultiNet foi fortemente influenciada: (i) por pesquisas nas áreas da Psicologia Cognitiva, (ii) pela Gramática de Casos de Fillmore (1968) e (iii) pelos trabalhos de Woods (1985) e Brachman (1979) que discutem o papel das redes semânticas na representação do conhecimento. Além disso, o MultiNet foi proposto com base em uma série de requisitos considerados essenciais para que um modelo de representação do conhecimento fosse adequado ao tratamento da semântica das línguas naturais, requisitos que podem ser divididos em: requisitos globais e internos.

A seguir, apresenta-se cada um desses requisitos (HELBIG, GNÖRLICH, 2002; HELBIG, 2006).

1. Requisitos globais do modelo

a) **Universalidade**: os meios de representação semântica devem ser definidos independentemente de uma língua natural específica ou do domínio da aplicação computacional e não deve ser uma "modelagem" *ad hoc* para um campo especial do discurso;

b) **Adequação cognitiva**: os meios de representação devem permitir uma modelagem cognitivamente motivada das estruturas conceituais e da manifestação dessas estruturas na semântica das línguas naturais. Além disso, o modelo deve ser centrado em conceitos (do inglês, *"concept-centering"*), ou seja, todo conceito deve ter uma representação única que torna acessível toda a informação a ele associada[37];

c) **Interoperabilidade**: os meios de representação devem ser aplicáveis tanto a investigações lingüísticas como lingüístico-computacionais. Em outras palavras, os recursos expressivos devem ser adequados e apropriados tanto para a construção de bases de conhecimento como para a representação das estruturas semântico-gramaticais das línguas naturais;

d) **Homogeneidade**: os mesmos meios de representação devem ser capazes de expressar o significado subjacente aos itens lexicais, às sentenças e aos textos;

e) **Comunicabilidade**: os construtos e suas definições devem ser claros para serem compartilhados por diferentes equipes de pesquisadores;

f) **Praticidade**: toda representação do conhecimento deve ser tecnicamente tratável e implementável;

g) **Automaticidade**: o repertório pré-definido dos meios de expressão deve permitir o processamento automático do conhecimento.

2. Requisitos internos do modelo

a) **Completude**: todo conceito expresso por meio de línguas naturais deve ser representado pelos meios previstos no modelo;

b) **Granularidade ótima**: significados distintos devem ser mapeados a estruturas conceituais também distintas, resguardadas as decisões sobre o grau de granularidade que deve ser considerado pelo modelo;

c) **Consistência:** partes de informação que sejam logicamente contraditórias não devem ser derivadas umas das outras;

---

[37] No âmbito da Ciência da Computação, essa caracterização é denominada "orientada para objetos" (do inglês, *objected-oriented*).

d) **Multidimensionalidade**: a distinção qualitativa dos diferentes aspectos do conhecimento deve ser espelhada na designação dos conceitos em termos de diferentes níveis de representação;
e) **Interpretabilidade local**: os construtos do modelo devem ser logicamente interpretáveis por si, independentemente de sua inclusão em uma base de conhecimento particular.

Esses requisitos permitem observar que o MultiNet distingue-se das representações baseadas em redes semânticas como o KL-ONE, assim como das representações semânticas dos seus sucessores (cf. ALLGAYER, REDDIG-SIEKMANN, 1990), e das representações baseadas na lógica devido principalmente aos critérios de adequação cognitiva (1.b) e homogeneidade (1.d). Essas representações do conhecimento baseiam-se em um modelo extensional, que não é capaz de descrever conceitos de natureza intensional como <intensão>, <charme>, etc. Além disso, o MultiNet diferencia-se desses modelos devido à sua "estrutura de multicamadas" e ao "encapsulamento de conceitos"; essas duas características serão descritas ao longo desta Seção. Aliás, é a representação do conhecimento em multicamadas (ou estratificado) que dá ao modelo sua denominação e justifica a abertura desse novo paradigma de representação do conhecimento.

Ressalta-se também que, segundo o critério de universalidade (1.a), o MultiNet pode ser empregado como uma **interlíngua**, já que é independente de língua. Conseqüentemente, os rótulos dos nós são apenas recursos mnemônicos; os rótulos para os conceitos <França> e <batalha>, por exemplo (cf. Figura 26, pág. 64), poderiam ser meros códigos como C(onceito)1 e C2, respectivamente. Destaca-se que uma das mais importantes aplicações do MultiNet é como interlíngua semântica para recuperação de informação na *Web* por meio de interfaces em língua natural (LEVELING, HELBIG, 2002; LEVELING, 2003, 2004).

4.2. Os meios de representação dos conceitos no MultiNet

O repertório dos meios de representação semântica do MultiNet, descritos na seqüência desta Subseção, está sistematizado na Figura 27, elaborada com base em Helbig (2006).

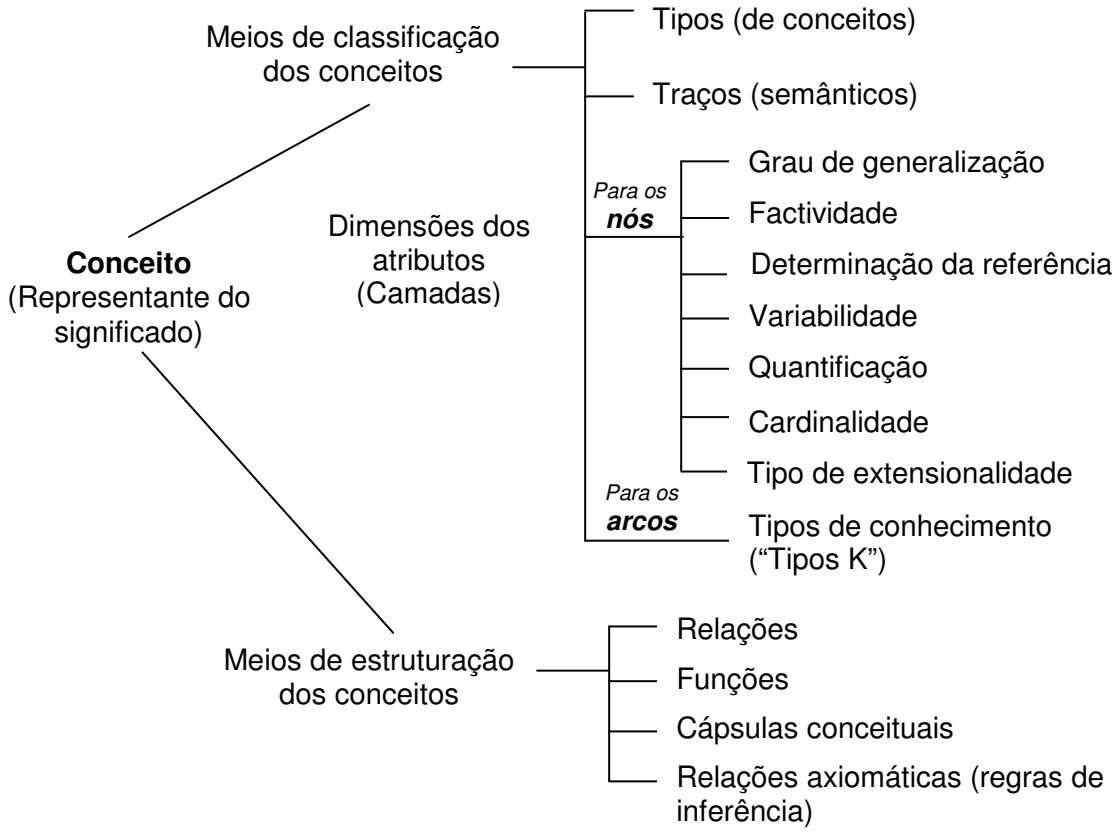


Figura 27: Os meios de representação semântica do MultiNet.

4.2.1. A classificação dos conceitos

4.2.1.1. Tipos conceituais e traços semânticos

O MultiNet baseia-se em uma **ontologia** a ser representada, ou seja, "em uma classificação dos conceitos do ponto de vista epistemológico" (HELBIG, 2006). As classes de conceitos definidas por essa ontologia são os **tipos** (do inglês, *sorts*). Os tipos desempenham um papel importante no aparato formal de representação do conhecimento porque são condições indispensáveis para a definição do domínio (do inglês, *domain*) e contradomínio (do inglês, *range*) das relações e funções[38]. Por exemplo, é essencial que a relação CAUS (causa) seja usada entre dois fatos (e não entre dois objetos concretos). Isso quer dizer que o domínio e o contradomínio da relação CAUS contêm elementos do tipo fato. Os tipos também desempenham papel importante na análise da língua e na definição do conhecimento lexical. Nesse sentido, é possível reconhecer, a partir do tipo [SORT=*d*] (isto é, o tipo que designa

[38] **Função** (matemática) f: $D \rightarrow Y$ é uma lei que associa elementos de um conjunto D, chamado o "domínio da função", a elementos de um outro conjunto Y, chamado o "contradomínio da função". Em outras palavras, o "domínio" é o conjunto que contém todos os elementos x para os quais a função deve ser definida. O "contradomínio", por sua vez, é o conjunto que contém os elementos que podem ser relacionados a elementos do domínio por meio da aplicação de função a elementos do domínio.

objetos discretos) da denotação do objeto gramatical da sentença *Os estudantes conjugaram as bicicletas*, que essa sentença não é aceitável. O verbo *conjugar* exige que a restrição selecional do seu objeto seja do tipo [SORT=*io*] (isto é, o tipo que designa objetos idealizados).

Do ponto de vista das ontologias modernas (CHANDRASEKARAN et al., 1999), a ontologia de tipos do MultiNet contempla a parte superior e mais geral de toda hierarquia conceitual (do inglês, *upper ontology, foundation ontology* ou *top-level ontology*). Os demais níveis da ontologia são essencialmente estabelecidos por relações hierárquicas, como SUB (ou seja, a subordinação entre conceitos do tipo entidade). Todo projeto que envolve o desenvolvimento de ontologias, como o projeto CYC (LENAT, GUHA, 1990), SUMO (do inglês, ***S**uggested **U**pper **M**erged **O**ntology*) (NILES, PEASE, 2001), WordNet (FELLBAUM, 1998), EuroWordNet (VOSSEN, 1998) e Mikrokosmos[39], pressupõe o estabelecimento desses tipos superiores ou gerais. Quanto à lexicalização, ressalta-se que o léxico das línguas tende a ser particularmente esparso nesse nível conceitual (HIRST, 2004). Certamente, todas as línguas incluem itens lexicais similares aos itens do português *objeto, substância* e *ações,* etc.; no entanto, esses itens tendem a ser vagos, variando conceitualmente de uma língua para outra. Daí a dificuldade de se delinear os conceitos mais gerais em uma ontologia. Essa dificuldade é reconhecida por Gangemi et al. (2001) quando da análise dos conceitos de nível superior da WordNet.

Os tipos previstos pelo MultiNet apresentam, de certa forma, pontos de contato com a proposta de Lyons (1977). Do ponto de vista lingüístico, destaca-se a proposta de organização das classes de entidades de Lyons (1977), a qual, como já se observou, fundamenta a ontologia do projeto EuroWordNet.

Na seqüência, esses tipos são definidos e exemplificados (HELBIG, 2006, p. 411-417).

- **Entidades** (do inglês, *entities*) **[*ent*]** - esse é o tipo conceitual mais geral; engloba todos os conceitos representados nos nós de uma rede semântica. Há 7 tipos gerais de entidades: Objetos, Situações Descritores de Situações, Qualidades, Quantidades, Graduadores e Entidades Formais. Esses tipos, por sua vez, subdividem-se em subtipos mais específicos, conforme apresentação a seguir.

[39] http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html

1. **Objetos** (do inglês, *objects*) **[*o*]** – os objetos desse tipo subdividem-se em objetos concretos, que intuitivamente categorizam coisas que podem ser percebidas pelos sentidos, e objetos abstratos, aos quais essa propriedade não se aplica.

   - **Objetos concretos** (do inglês, *concrete objects*) **[*co*]**: incluem-se nesse tipo as substâncias e os objetos discretos:
     - **Substâncias** (do inglês, *substances*) **[*s*]**: definem-se por possuírem uma existência (quase) contínua, são divisíveis, mas não contáveis, p.ex.: <leite>, <ferro>, <30g de urânio>, etc.
     - **Objetos discretos** (do inglês, *discrete objects*) **[*d*]**: definem-se por serem entidades contáveis e não divisíveis; p.ex.: <casa>, <bicicleta>, etc.
   - **Objetos abstratos** (do inglês, *abstract objects*) **[*ab*]**: incluem-se nesse tipo produtos do raciocínio humano. Distinguem-se os seguintes tipos:
     - **Objetos situacionais** (do inglês, *situational objects*) **[*abs*]**: definem-se por representarem situações em termos de objetos. São subdivididos em:
       - **Abstrações a partir de situações dinâmicas** (do inglês, *abstractions from dynamic situations*) **[*ad*]**: <corrida>, <assalto>, etc.
       - **Abstrações a partir de situações estáticas** (do inglês, *abstractions from static situations*) **[*as*]**: <equilíbrio>, <sono>, etc.
     - **Atributos** (do inglês, *attributes*) **[*at*]**: subdividem-se em:
       - **Atributos mensuráveis [*oa*]**, p.ex.: <altura>, <peso>;
       - **Atributos não-mensuráveis [*na*]**, p.ex.: <flexibilidade>, etc.
     - **Relações** (do inglês, *relationships*) **[*re*]**: definem-se por relações do tipo <causalidade>, <similaridade>, <diferença>, etc.
     - **Objetos idealizados** (do inglês, *ideal objects*) **[*io*]**: definem-se por abstrações como <religião>, <justiça>, <categoria>, etc.
     - **Objetos abstratos temporais** (do inglês, *abstract temporal objects*) **[*ta*]**: definem-se por eventos temporariamente delimitados como <Renascença>, <feriado>, <Páscoa>, etc.
     - **Modalidades** (do inglês, *modalities*) **[*mo*]**: definem-se por classificarem as modalidades, p.ex.: <necessidade>, <probabilidade>, etc.

2. **Situações (estados-de-coisas)** (do inglês, *situations/ states-of-affair*) **[*si*]** – conjuntos de objetos, seus modos de ser ou mudanças a que se submetem. Incluem-se nesse tipo situações estáticas e situações dinâmicas:
   - **Situações estáticas (estados)** (do inglês, *static situations/ states*) **[*st*]**: incluem os estados físicos e psíquicos, p.ex.: <descanso>, <estar doente>, etc.
   - **Situações dinâmicas (eventos)** (do inglês, *dynamic situations/ events*) **[*dy*]**: distinguem-se dois tipos:
     - **Ações** (do inglês, *actions*) **[*da*]**: definem-se por denotarem situações dinâmicas que são desempenhadas por um agente, p.ex.: <trabalhar>, <escrever>, etc.
     - **Acontecimentos** (do inglês, *happenings*) **[*dn*]**: definem-se por serem eventos com causa em que a causa não está associada a um agente, p.ex.: <chover>, <explodir>, etc.
3. **Descritores situacionais** (do inglês, *situational descriptors*) **[*sd*]** – especificam a ancoragem espaço-temporal e modal das situações. Distinguem-se:.
   - **Descritores do tempo da situação** (do inglês, *temporal situational descriptors*) **[*t*]**: esse tipo inclui as especificações de tempo na forma de momentos ou intervalos de tempo, p.ex.: <às 19 horas>, <na segunda-feira>, etc.
   - **Descritores do local da situação** (do inglês, *local situational descriptors*) **[*l*]**: esse tipo especifica as localidades como caracterização espacial das situações, p.ex.: <no telhado>, <embaixo da mesa>, etc.
   - **Modalidades** (do inglês, *modalities*) **[*md*]**: esse tipo especifica a atitude do falante sobre o que é dito e sobre a validade das situações, p.ex.: <provavelmente>, <impossível>, etc.
4. **Qualidades** (do inglês, *qualities*) **[*ql*]** – especificam as qualidades ou as especificações das propriedades. Distinguem:
   - **Propriedades no sentido amplo** (do inglês, *properties in the broadest sense*) **[*p*]**: incluem as qualidades semanticamente completas [*tq*] e as qualidades graduáveis [*gq*]; ds últimas podem ser divididas em:
     - **Qualidades mensuráveis [*mq*]**: como <peso>;
     - **Qualidades não-mensuráveis [*nq*]**: como <cruel>;
   - **Qualidades relacionais** (do inglês, *relational qualitites*) **[*rq*]**: esse tipo especifica relações entre entidades; essas qualidades podem ser atribuídas somente a pluralidades com ao menos dois elementos, p.ex.: <equivalente>, <inverso>, etc.

- **Qualidades funcionais** (do inglês, *functional qualities*) **[*fq*]**: esse tipo especifica qualidades que obtêm significado pleno em conexão a outras entidades. Combinadas com as últimas, elas formam uma unidade conceitual nova. Distinguem-se:
    - **Associativas [*aq*]**: como <químico>;
    - **Operacionais [*oq*]**: como <último> e <terceiro>;

5. **Quantidades** (do inglês, *quantities*) **[*qn*]** – esse tipo expressa o aspecto quantitativo dos conceitos. Distinguem-se:
    - **Quantificadores** (do inglês, *quantificators*) **[*qf*]**: subdividem-se em:
        - **Numéricos [*nu*]**: como <dois>;
        - **Não-numéricos** [*nm*]: como <mais que a metade>;
    - **Unidades de medida** (do inglês, *units of measurements*) **[*me*]**: esse tipo especifica unidades de medida (p.ex.: kg, C^o), p.ex.: <3 kg>, <muitas horas>, etc.
6. **Graduadores** (do inglês, *graduators*) **[*gr*]**: esse tipo especifica qualidades e quantidades de modo refinado. Distinguem-se:
    - **Graduadores qualitativos** (do inglês, *qualitative graduators*) **[*lg*]**: refinam a especificação das propriedades, p.ex.: <especificamente>, etc.
    - **Graduadores quantitativos** (do inglês, *quantitative graduators*) **[*ng*]**: são usualmente empregados com propriedades vagas, p.ex.: <aproximadamente>, <menos que>, etc.
7. **Entidades formais** (do inglês, *formal entities*) **[*fe*]**: representam objetos extralingüísticos como figuras, quadros, fórmulas, etc.

Além dos tipos de conceitos, o MultiNet possibilita também a especificação de atributos de conceitos, os chamados **traços** (do inglês, *features*), que desempenham papel fundamental na classificação dos conceitos e na análise léxico-semântica. Mais especificamente, os traços facilitam a formulação de restrições de seleção e da subcategorização dos itens lexicais, possibilitando a especificação da valência de verbos, adjetivos e nomes.

A Figura 28 ilustra os principais subtipos do tipo [SORT=*co*] (objetos discretos), definidos em função de uma hierarquia de estruturas atributo-valor. Os atributos, como ANIMAL e ARTIF, estão sistematizados no Quadro 2 (pág. 90) e descritos na seqüência desse Quadro.

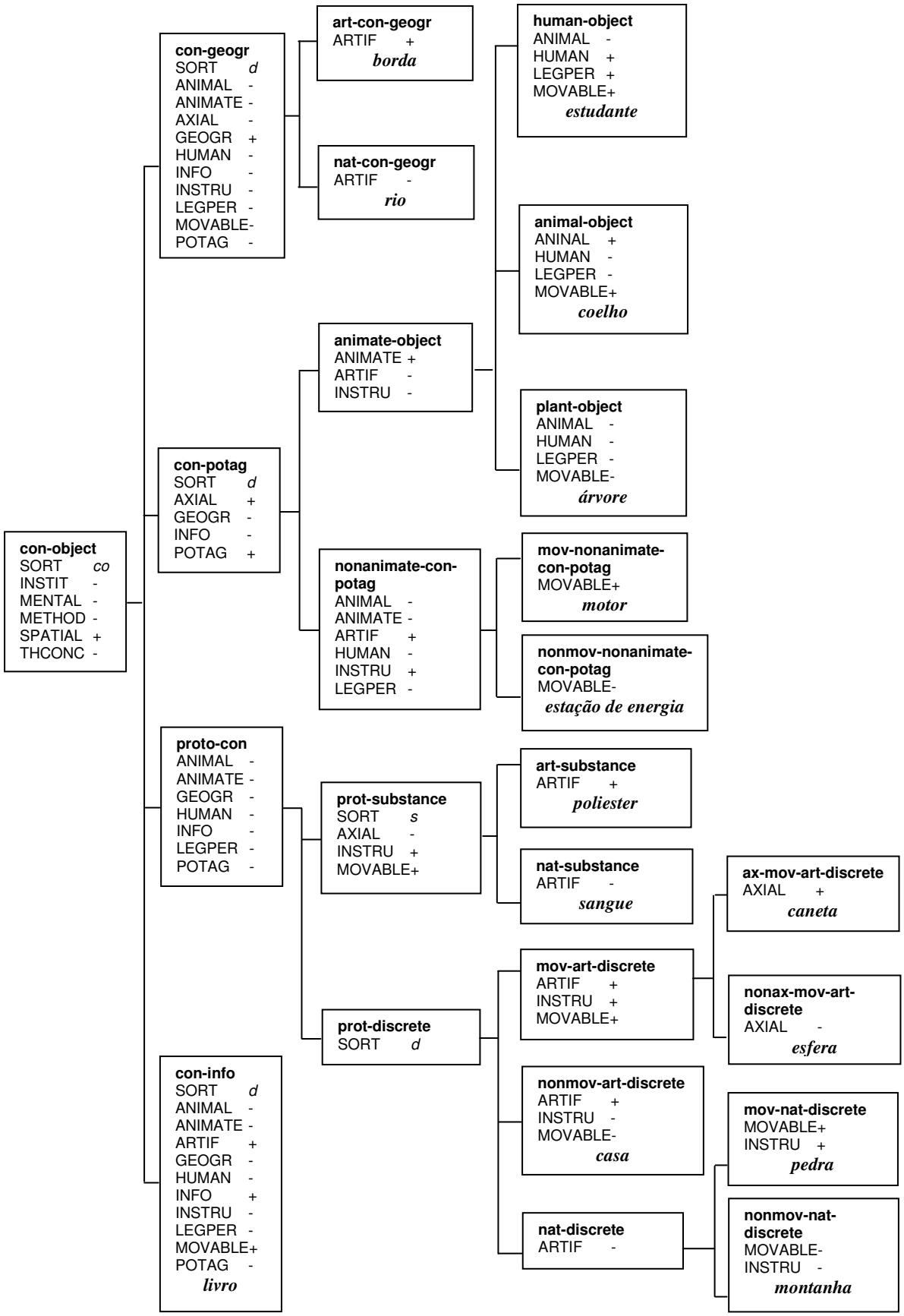


Figura 28: Exemplo das combinações do tipo [SORT=*co*] e dos traços dos objetos concretos (*con-object*).

Os subtipos do tipo [SORT=*co*], contidos na Figura 28, podem ser assim descritos:

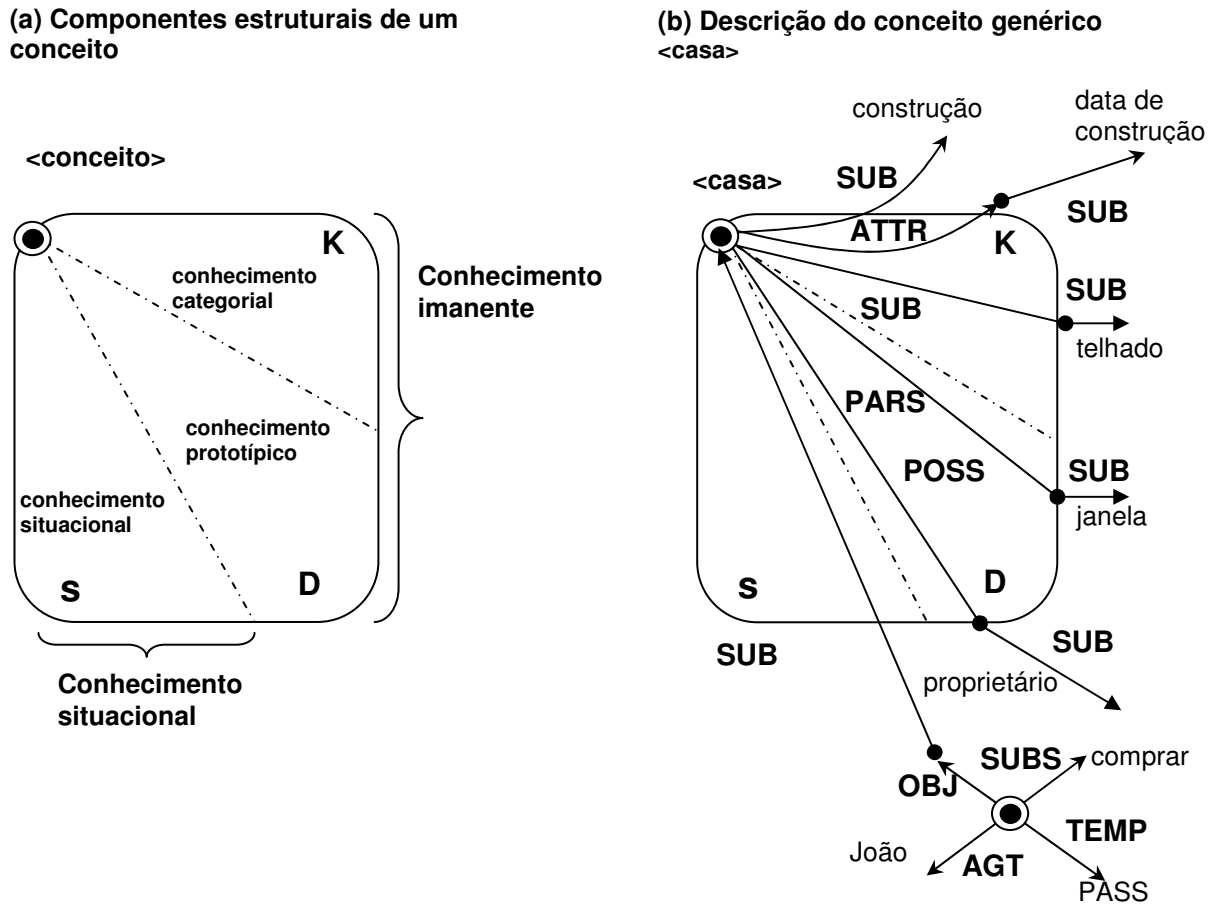
- ***con-object*****:** objeto **con**creto
    - ***con-geogr*****:** geográfico
        - ***art-con-geogr*****:** artefato
        - ***nat-con-geogr*****:** natural
    - ***con-potag*****:** agente potencial
        - ***animate-object*****:** animado
            - ***human-object*****:** humano
            - ***animal-object*****:** animal
            - ***plant-object*****:** vegetal
        - ***nonanimate-con-potag*****:** não-animado
            - ***mov-nonanimate-con-potag*****:** móvel
            - ***nonmov-nonanimate-con-potag*****:** imóvel
    - ***proto-con*****:** prototípico;
        - ***proto-substance*****:** substância
            - ***art-substance*****:** artificial
            - ***nat-substance*****:** natural
        - ***proto-discrete*****:** discreto
            - ***mov-art-discrete*****:** móvel
                - ***ax-mov-art-discrete*****:** com eixos caracteríticos
                - ***nonax-mov-art-discrete*****:** sem eixos caracteríticos
            - ***nonmov-art-discrete*****:** artefato imóvel
            - ***nat-discrete*****:** natural
                - ***mov-nat-discrete*****:** móvel
                - ***nonmov-nat-discrete*****:** imóvel
    - ***con-info*****:** objeto de informação

Os atributos e seus respectivos valores são herdados por uma operação de unificação das estruturas atributo-valor. A consistência é garantida no processo de herança pelo fato de que os subtipos podem ser unificados e formar os supertipos. Na Subseção em que se destaca a relação entre representação do conhecimento e léxico, mostra-se como tal hierarquia pode ser gerada automaticamente a partir de um dado conjunto de dependências entre traços e valores.

A seguir, são descritos os atributos dos nós, que se distribuem em sete dimensões.

4.2.1.2. Tipos de conhecimento classificados no MultiNet

No MultiNet, uma distinção importante é feita entre dois tipos de conhecimento (atributo K-TYPE): **conhecimento imanente** e **conhecimento situacional**. O conhecimento imanente engloba todos os componentes de significado de um conceito que são independentes da sua manifestação em uma situação ou contexto específico. De certa forma, esse tipo de conhecimento relaciona-se ao nível do significado da expressão. Como exemplificação, apresenta-se o conceito <casa> na Figura 29, com base em Helbig (2006, p. 435).



K	conhecimento imanente categorial: Uma casa é um edifício; necessariamente tem telhado e uma data de construção.
D	conhecimento imanente prototípico: Uma casa (normalmente) possui janelas e tem um proprietário.
S	conhecimento situacional: João comprou uma casa.

Figura 29: Os três tipos de conhecimento no MultiNet.

O fato de uma casa, por exemplo, ter partes como <janela>, <telhado>, etc., são informações que pertencem ao tipo de conhecimento imanente, que é essencial para a caracterização do conceito <casa>. Já o fato de João ter comprado uma casa não é essencial para a caracterização desse conceito; esse fato não caracteriza propriamente o conceito e se relaciona à ocorrência do conceito na descrição de um estado-de-coisas específico; assim, enquadra-se no conhecimento situacional.

O conhecimento imanente é raramente explícito ou descrito no discurso, pois, em geral, é pressuposto pelos falantes. Para demonstrar a importância desse conhecimento e o seu uso no processo de desambiguação de uma sentença gramaticalmente ambígua. Observem-se as sentenças em (4):

(4) O preço do livro é 25 reais. Seu autor vive no Rio de Janeiro.

Por meio de recursos puramente gramaticais não é possível determinar se o pronome *seu*, refere-se a *preço* ou a *livro*. Isso pode ser feito apenas por meio de conhecimento de mundo: como livros têm preços e autores, mas o preço (como um atributo) não é caracterizado por ter um autor, o conceito <preço> não pode ser o antecedente para o sintagma *seu autor*. Assim, o antecedente pretendido é <livro>.

O conhecimento imanente subdivide-se em **conhecimento categorial** (rótulo K, do alemão, *Kategorisch*) e **conhecimento prototípico** (rótulo D, do inglês, *Default knowledge*) (cf. Figura 29). 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. Dessa forma, é possível a apresentação tanto uma casa com janelas e portas e tendo um proprietário como de uma casa que não possua janelas ou portas. Conseqüentemente, é importante salientar que o conhecimento categorial está conectado ao mecanismo de herança monotônica e o conhecimento prototípico, à herança não-monotônica (cf. Subseção 5.1.1, pág. 96).

Os conhecimentos imanente e situacional ainda são classificados como **conhecimentos descritivos** (do inglês, *descriptive knowledge*), porque são usados para "descrever" objetos e situações. Há partes do conhecimento, no entanto (como condições, restrições modais, especificações contextuais), que não descrevem um objeto ou situação, mas sim restringem sua condição de existência, por exemplo, no caso dos objetos. Essas partes são classificadas como **conhecimento restritivo** (do inglês, *restrictive knowledge*) (HELBIG, 2006). O

MultiNet provê meios especiais para atribuir a cada arco o tipo de conhecimento (K-TYPE) que descreve. Os tipos de conhecimentos estão sistematizados na Figura 30.

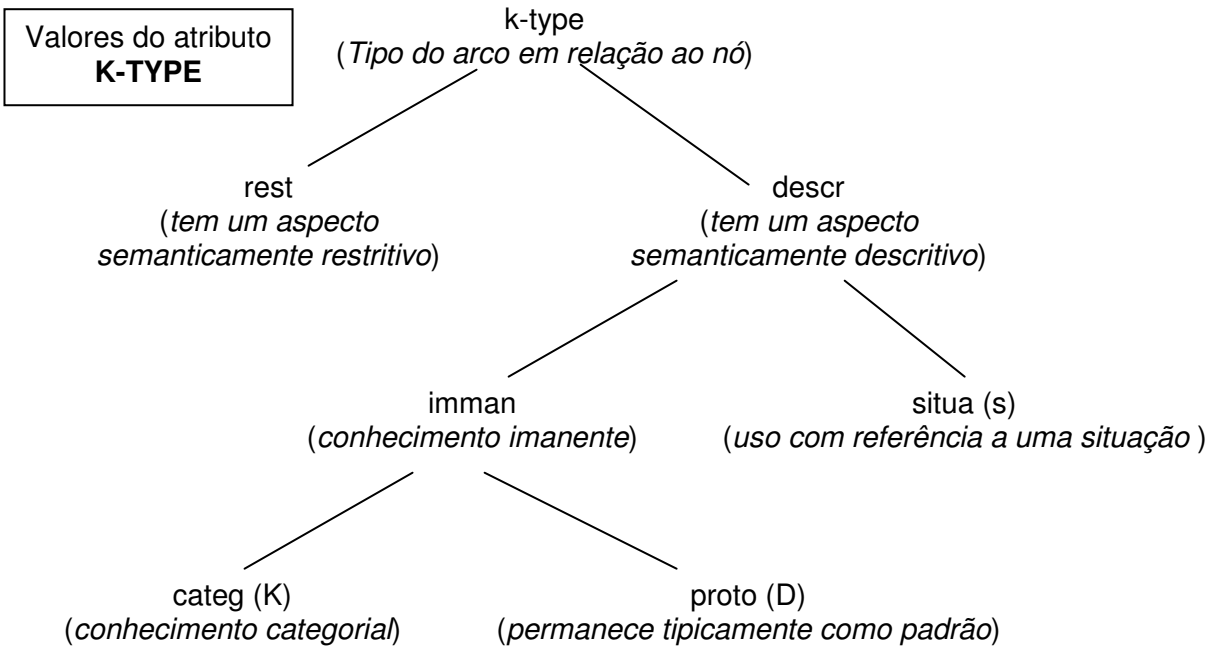


Figura 30: Os valores do atributo K-TYPE.

Vale ressaltar, assim, que todo arco (k1 REL k2) rotulado por uma relação REL e que liga o nó k1 ao nó k2 é caracterizado por diferentes valores do atributo K-TYPE. Por exemplo, na relação SUB entre <esta mosca> e <inseto>, o arco SUB é do tipo *categ* (conhecimento categorial) em relação ao conceito <esta mosca>, já que é uma parte intrínseca das moscas serem insetos; e do tipo *situa* (conhecimento situacional) em relação ao conceito <inseto>, pois este não é imanentemente caracterizado pela inclusão de mosca em sua descrição.

4.2.2. Os atributos multidimensionais

A característica essencial do MultiNet é a organização dos atributos associados aos nós em um espaço multidimensional (HELBIG, GNÖRLICH, 2002; HELBIG, 2006). Essa característica também é responsável por distinguir o MultiNet dos demais formalismos baseados em redes semânticas, que organizam o significado em uma estruturada “plana” (do inglês, *flat*). No MultiNet, os nós estão inseridos em um espaço multidimensional caracterizado por esses atributos, que caracterizam camadas na rede. A especificação das camadas é dada pelo atributo complexo LAY, cujas sete dimensões e seus valores estão sistematizados na Figura 31.

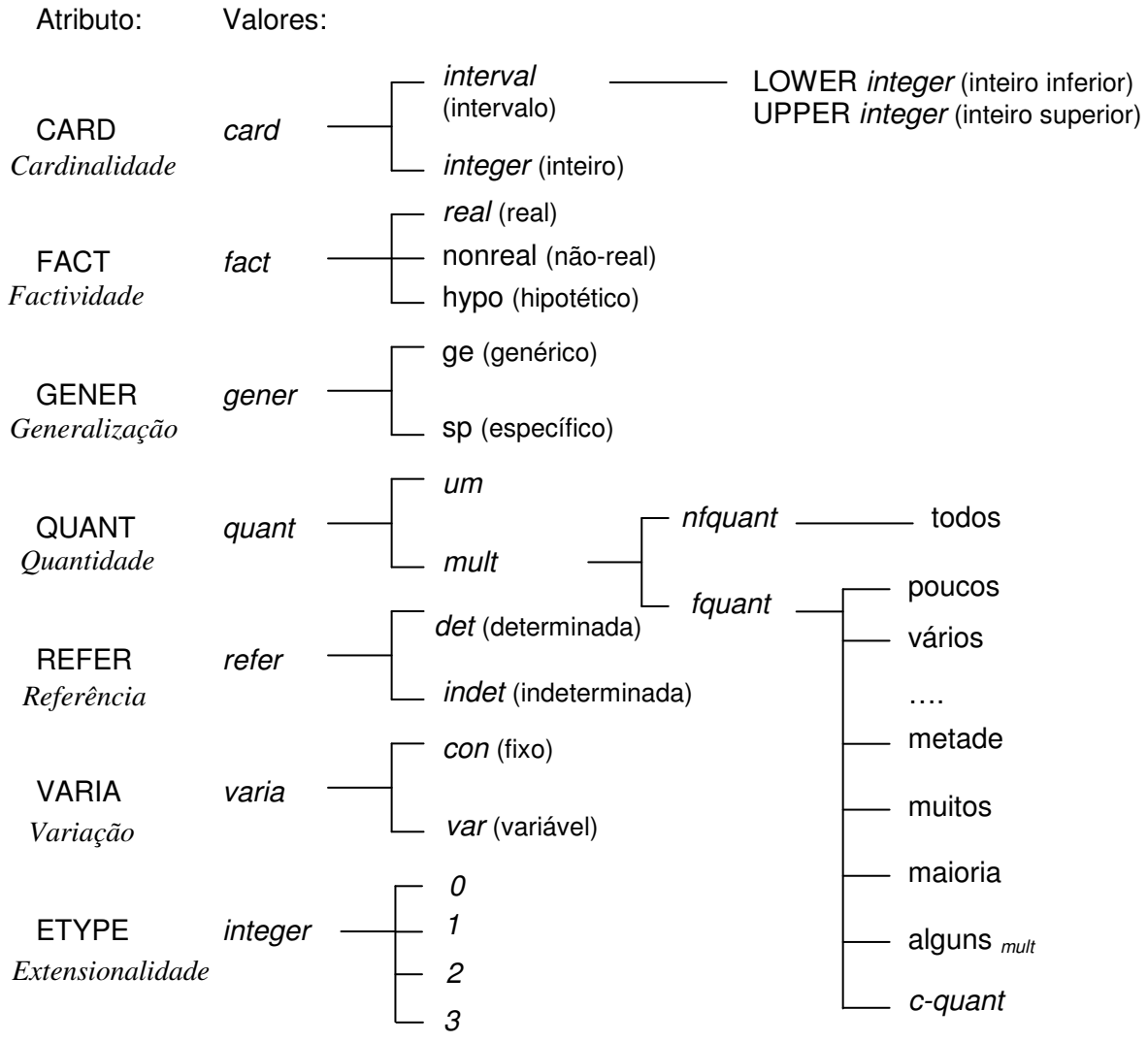


Figura 31: Os atributos multidimensionais do MultiNet e seus valores.

As setes dimensões do atributo LAY podem ser divididas em dois níveis: nível **intensional** e **pré-extensional**. Para a proposição desses níveis, o MultiNet baseou-se na distinção entre os aspectos intensional e extensional na interpretação semântica das línguas naturais, a qual foi amplamente discutida na Filosofia (CARNAP, 1958).

No nível intensional, os nós são caracterizados pelos atributos **grau de generalização** (do inglês, *degree of generality*) (GENER), **quantificação** (intensional) (do inglês, *quantification*) (QUANT) e **determinação da referência** (do inglês, *determination of reference*) (REFER). No nível pré-extensional, os nós são especificados pelos atributos **cardinalidade** (do inglês, *cardinality*) (CARD), **factividade** (do inglês, *facticity*) (FACT), **grau de variação** (do inglês, *variability*) (VARIA) e **tipo de extensionalidade** (do inglês, *type of extensionality*) (ETYPE). Esses sete atributos são descritos a seguir na mesma ordem em que aparecem na Figura 31.

a) **Cardinalidade pré-extensional:** descreve o conceito expresso lingüisticamente por um sintagma nominal quantificado ou determinado nos níveis intensional e pré-extensional. Mais especificamente, o atributo cardinalidade (CARD) descreve o conceito no nível pré-extensional (HELBIG, GNÖRLICH, 2002). A sentença, a seguir, ilustra esse atributo.

(5) Vários alunos da 1ª série [CARD≤40] fizeram o exercício.

A cardinalidade dos conceitos como <vários alunos da 1ª série> não é dada explicitamente; ela é deduzida do conhecimento de mundo dos falantes, sendo especificada por meio de valores numéricos, como 1, 2, etc., ou mesmo por intervalos numéricos (1...10, ≤2, ≥8, etc.). Assim, o conceito <vários alunos da 1ª série> presente na sentença (5), com base no conhecimento de mundo sobre o número médio de alunos em uma sala de aula, permite a especificação do atributo [CARD≤40].

b) **Factividade:** define a propriedade dos conceitos de representarem objetos reais como <Nova Iorque> e <Torre Eiffel>, objetos hipotéticos como <buraco negro>, e objetos imaginários (não-reais) como <unicórnio>. Nesses casos, os valores do atributo FACT são, respectivamente: [FACT=*real*], [FACT=*hypo*] e [FACT=*nonreal*], conforme exemplificam (6.a) e (6.b).

(6) a. João [FACT=*real*] achou que ela estivesse doente [FACT=*hypo*].
b. João [FACT=*real*] lembrou que ela estava doente [FACT=*real*].

c) **Grau de generalização:** indica se uma entidade conceitual é genérica ([GENER=*ge*]) ou específica ([GENER=*sp*]), como ilustram os exemplos (7), (8) e (9), adaptados de Helbig (2006):

(7) a. Paulo encontrou quatro homens. [GENER=*sp*]
b. Quatro homens são [GENER=*ge*] necessários para carregar uma geladeira.
(8) a. Max tem um carro. [GENER=*sp*]
b. Um carro [GENER=*ge*] garante máxima mobilidade.
(9) a. Estes cachorros [GENER=*sp*] são perigosos.
b. Cachorro [GENER=*ge*] que late não morde.

d) **Quantificação intensional:** descreve o conceito expresso lingüisticamente por um sintagma nominal quantificado ou determinado nos níveis intensional e pré-extensional. Mais especificamente, o atributo quantificação (QUANT) descreve o conceito no nível intensional (HELBIG, GNÖRLICH, 2002). Para ilustar esse atributo, utiliza-se novamente o exemplo (5).

(5) Vários alunos da 1ª série [QUANT=*mult*]fizeram o exercício.

O conceito <vários alunos da 1ª série> é intensional e caracterizado por certa vagueza, posto que é impossível especificar o número de estudantes a partir do conceito <vários alunos da 1ª série>. Dessa forma, o valor do atributo QUANT para o referido conceito pode ser definido como [QUANT=*mult*].

e) **Determinação da referência:** define o grau com que um conceito determina a entidade a que se aplica. O valor desse atributo não se aplica a conceitos genéricos (isto é, [REFER=*refer*]). Para os outros casos, distinguem-se dois valores: [REFER=*det*] e [REFER=*indet*]., conforme exemplifica a frase (10):

(10) a. O passageiro [REFER=*det*] de quem falei presenciou um acidente. [REFER=*indet*].

f) **Variação:** define o grau de variação do conceito, isto é, se o objeto é fixo ou variável. Os valores do atributo VARIA podem ser:

- [VARIA=*con*]: conceito é fixo;
- [VARIA=*var*]: conceito é variável;
- [VARIA=*varia*]: o conceito é genérico. Esse valor muda para [VARIA=*con*] ao se descrever um elemento prototípico da extensão de um conceito genérico. Exemplificam-se, a seguir, esse atributo e seus valores.

(11) a. O policial [VARIA=*con*] checou o passaporte [VARIA=*var*].
b. Há um livro [VARIA=*con*] que foi lido por todos os alunos [VARIA=*varia*].
c. Estudantes [VARIA=*varia*] lêem livros [VARIA=*varia*].

g) **Tipos de extensionalidade:** caracteriza a extensão dos conceitos. Da mesma forma que os tipos e traços são usados para especificar, no nível intensional, que relação e função podem

em princípio descrever certos relacionamentos entre conceitos, uma classificação dos nós é necessária para a realização do processo análogo no nível pré-extensional. Assim, a relação ELMT (x é um elemento de y) estabelece-se entre um elemento e um conjunto, entre um conjunto e uma coleção (isto é, uma família de conjuntos), etc. Os valores do atributo ETYPE são descritos na seqüência (HELBIG, 2002). Vale ressaltar, antes disso, que o valor ETYPE-*nil* caracteriza conceitos como <intensão>, <regularidade>, etc., que não têm extensão. E mais, o atributo CARD não é aplicável [ETYPE=Ø]. Os valores do atributo ETYPE são assim definidos:

0 – elemento de um conjunto, p.ex.: <Elizabeth I>, <a casa>, <esta escola>, etc.
1 – conjunto de elementos do tipo [ETYPE=0], p.ex.: <muitas casas>, <uma família>, etc.
2 – coleção de conjuntos do tipo [ETYPE=1], p.ex: <muitas famílias>, <três tripulações>, etc.
3 – aglomerado de coleção do tipo [ETYPE=2], p.ex.: <organização guarda-chuva>, etc.

A quantificação dos sintagmas nominais, como salienta Helbig (2006), é uma questão difícil. A seguir, no Quadro 1, apresenta-se a proposta de Helbig (2006) para a classificação dos sintagmas nominais em função dos seguintes atributos do atributo complexo LAY: GENER, REFER, VARIA, FACT, QUANT e ETYPE.

No	GENER	REFER	VARIA	FACT	QUANT	ETYPE
1	*Sp*	*Det*	*Con*	*Real*	*Um*	*0*
	Exemplo: *Este urso tem uma pele grossa.*					
2	*Sp*	*Det*	*Con*	*Non*	*Um*	*0*
	Exemplo: *Este unicórnio é inofensivo.*					
3	*Sp*	*Indet*	*Con*	*Hypo*	*Um*	*0*
	Exemplo: *Ele acreditava ter descoberto um novo planeta.*					
4	*Sp*	*Indet*	*Var*	*Real*	*Um*	*0*
	Exemplo: *Todo mundo já viu um avião.*					
5	*Sp*	*Det*	*Con*	*Real*	*Um*	*1*
	Exemplo: *Os Alpes são o habitat natural dos íbex.*					
6	*Sp*	*Det*	*Con*	*Real*	*Mult*	*1*
	Exemplo: *Os ursos no zôo de XX são especialmente agressivos.*					
7	*Ge*	*Refer*	*Varia*	*Real*	*Mult*	*1*
	Exemplo: *Ursos são animais agressivos.*					
8	*Ge*	*Refer*	*Con*	*Real*	*Um*	*0*
	Exemplo: *O/ Um urso é um animal agressivo.*					
9	*Gener*	*Det*	*Con*	*Real*	*Todos*	*1*

	Exemplo: *Todos os ursos são perigosos.*					
10	*Sp*	*Det*	*Var*	*Real*	*Um*	*0*
	Exemplo: *Todo urso tem seu próprio lugar para dormir.*					
11	*Sp*	*Indet*	*Con*	*Real*	*Um*	*0*
	Exemplo: *Todos os garotos entraram em um bote.*					
12	*Sp*	*Indet*	*Var*	*Real*	*Um*	*0*
	Exemplo: *Cada garoto entrou em um outro bote.*					
13	*Ge*	*Indet*	*Var*	*Real*	*Mult*	*1+* [CARD=2]
	Exemplo: *É mais fácil manter dois cachorros que um.*					
14	*Ge*	*Indet*	*Var*	*Real*	*Um*	*0+* [CARD=1]
	Exemplo: *É mais difícil manter um papagaio que dois.*					
15	*Ge*	*Refer*	*Con*	*Real*	*Um*	*1*
	Exemplo: *O comportamento de uma multidão é difícil de ser previsto.*					
16	*Sp*	*Det*	*Con*	*Real*	*Um*	*1*
	Exemplo: *O policial dispersou a multidão.*					
17	*Sp*	*Indet*	*Con*	*Real*	*Muitosy*	*2*
	Exemplo: *O policial já tinha dispersado muitas multidões.*					
18	*Ge*	*Refer*	*Varia*	*Real*	*Vários*	*2*
	Exemplo: *Uma organização guarda-chuva consiste de várias organizações.*					
19	*Gener*	*Det*	*Con*	*Real*	*Todos*	*3*
	Exemplo: *O presidente visitou todas as organizações guarda-chuva.*					

Quadro 1: Exemplos de classificação dos conceitos nominais.

Com base no Quadro 1, Helbig (2006) conclui que:

a) **Conceitos individuais** (do inglês, *individual concepts*): conceitos que designam indivíduos com identificação fixa são caracterizados por meio dos seguintes atributos e valores: [GENER=*sp*], [VARIA=*con*] e [ETYPE=*0*] (Quadro 1, linhas 1 a 3). A caracterização desses conceitos, com exceção do atributo ETYPE, também é encontrada no caso dos conceitos coletivos.

b) **Conceitos genéricos** (do inglês, *generic concepts*): conceitos genéricos são especificados pelos atributos e valores [GENER=*ge*], [REFER=*refer*] e [VARIA=*con*]. O tipo de extensionalidade é geralmente [ETYPE=*0*], porque a descrição, no nível pré-extensional, de um conceito genérico normal B é um elemento prototípico do conjunto <todos os B>; com exceção dos conceitos coletivos, como <multidão>, que têm o atributo ETYPE com o valor > 0 (linha 15). Os valores dos atributos [REFER=*refer*] e [VARIA=*con*] são assim especificados porque um conceito genérico não determina a referência; esse conceito é relacionado a um elemento prototípico não-especificado que não varia (linha 8).

c) **Conceitos coletivos** (do inglês, *collective concepts*): são conceitos caracterizados como conceitos individuais no nível intensional, mas têm um tipo de extensionalidade maior que zero (linhas 5, 15 e 16). Os conceitos descritos por um nome coletivo podem ser [ETYPE=*2*] (linha 17) e [ETYPE=*3*] (linha 19).

d) **Conceitos parametrizados** (do inglês, *parameterized concepts*): conceitos que carregam o par atributo-valor [VARIA=*var*] são chamados entidades parametrizadas porque eles desempenham papel similar a variáveis quantificadoras (linhas 4, 10 e 12). Distinguem-se as entidades independentes e dependentes no nível pré-extensional, que são respectivamente caracterizadas pela presença e ausência de uma relação DPND (x depende de y) em sua especificação.

e) **Pluralidades generalizadas** (do inglês, *pluralities concepts*): conceitos que expressam construções plurais genéricas. Esses conceitos devem ser distinguidos dos conceitos genéricos clássicos porque não há um elemento prototípico em jogo. Nos casos descritos nas linhas 13 e 14, os atributos e seus valores [REFER=*indet*] e [VARIA=*var*] expressam que os conceitos correspondentes significam quaisquer dois cachorros ou qualquer papagaio.

Os tipos de entidades [SORT=*ent*] podem ser agrupados em função dos atributos relevantes a sua caracterização. Segundo a hierarquia da Figura 32, os atributos FACT e GENER são relevantes para os tipos de entidades *o* (objetos)*, si* (situação)*, t* (tempo) e *l* (local), os quais constituem o subtipo *osi-tl-lay.* Esses atributos são os únicos que caracterizam os elementos do subtipo *si-lay*. Já os elementos *o, t,* e *l,* que constituem o subtipo *o-tl-lay*, caracterizam-se pelos atributos QUANT, REFER, CARD, ETYPE e VARIA, além daqueles herdados do supertipo *osi-tl-lay* (FACT e GENER).

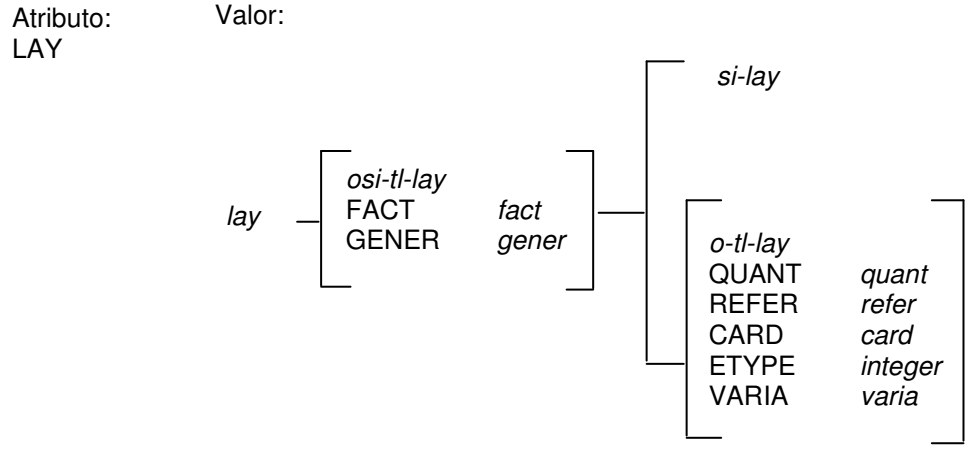


Figura 32: Organização dos tipos [SORT=*ent*] em função do atributo complexo LAY.

4.2.3. Os meios de estruturação dos conceitos

4.2.3.1. Relações e funções

Os meios de representação de uma estrutura conceitual, isto é, dos arcos que ligam os nós de uma rede semântica, são dados por relações e funções, as quais estão sistematizadas na Figura 33[40].

Na Figura 33, os meios de representação estão divididos, no topo da hierarquia, segundo suas associações com os níveis intensional e pré-extensional. Enquanto as relações e funções no nível intensional são usadas para descrever objetos e situações, os meios de representação no nível pré-extensional são empregados para descrever relações de conjunto e para interligar os níveis intensional e pré-extensional.

Nesse cenário, as relações lexicais desempenham papel essencial, pois conectam os significados das palavras, e, uma vez que as relações lexicais sempre conectam semanticamente conceitos genéricos, elas se associam ao nível intensional.

Com relação à caracterização dos conceitos, em especial, distinguem-se os meios para a representação da (i) descrição interna (estrutural ou qualitativa) dos objetos, denominada **descrição intraobjetiva** (do inglês, *intraobjective description*), e para a (ii) descrição do relacionamento entre diferentes conceitos, denominada **descrição interobjetiva** (do inglês, *interobjective description*) (HELBIG, 2006).

[40] As linhas pontilhadas indicam relações interníveis.

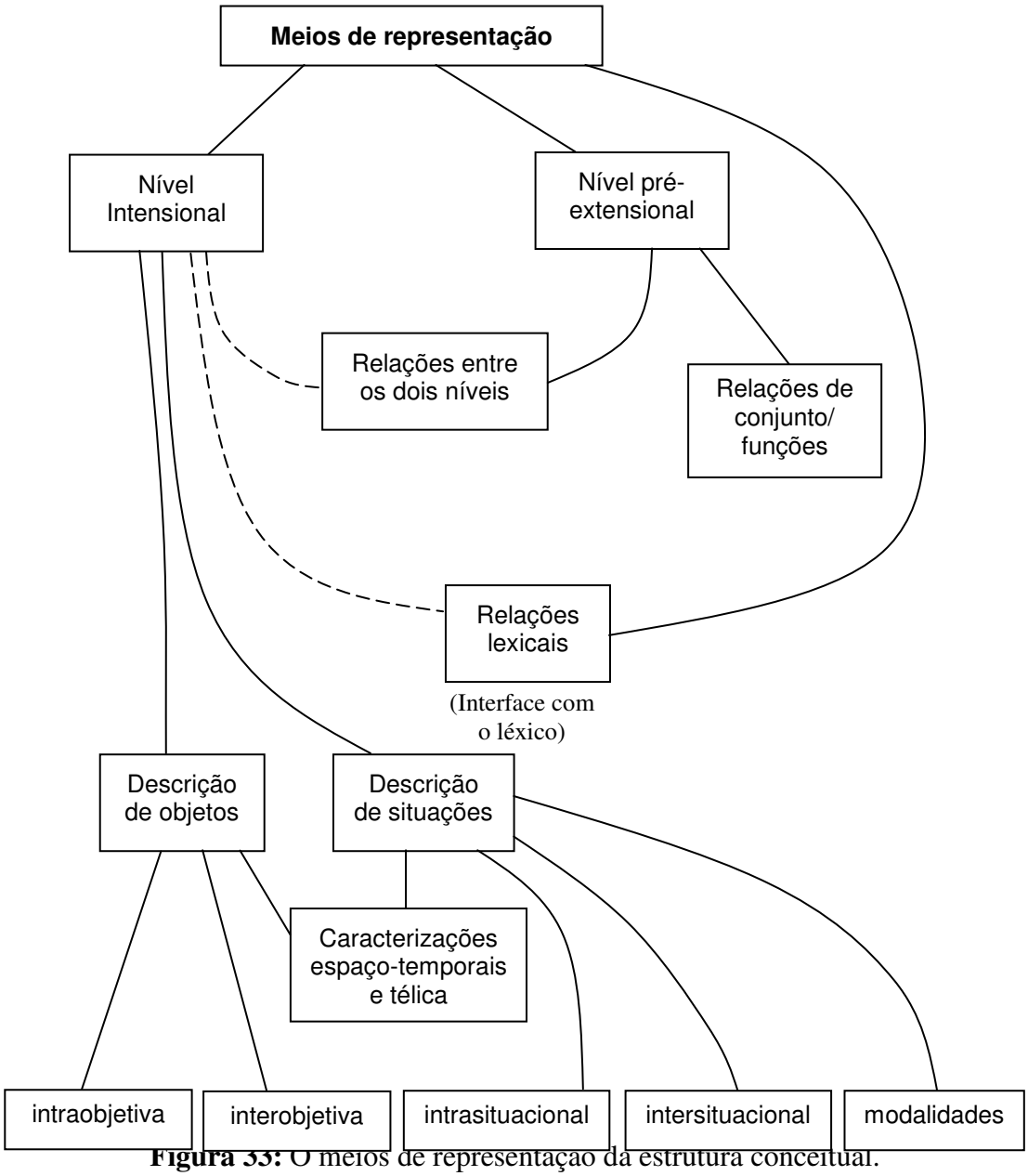


Figura 33: O meios de representação da estrutura conceitual.

Além disso, a Figura 33, baseada em Helbig (2006, p. 440), mostra que há meios para a caracterização espaço-temporal dos objetos e para a caracterização télica, que diz respeito à origem e função dos objetos. Vale ressaltar que a relação mais importante na área dos objetos é a relação SUB (do inglês, *subordination*) (subordinação ou inclusão), que estabelece a estrutura hierárquica dos objetos conceituais. Os meios de representação para a descrição dos objetos conceituais serão enfatizados na próxima Seção.

De modo similar, os meios descritivos para as situações podem ser divididos em **relações intrasituacionais** (do inglês, *intrasituational*) e **intersituacionais** (do inglês,

intersituational). As situações também são organizadas hierarquicamente. Tal hierarquia é especificada pela relação SUBS (do inglês, *subordination for situations*).

Um outro grupo de meios de representação é dado pelas relações entre os níveis intensional e pré-extensional.

Por fim, as relações lexicais são usadas para conectar diretamente os itens lexicais.

4.2.3.2. Encapsulamento de conceitos

Para representar o escopo do significado de um conceito dentro de uma representação do conhecimento, são necessários meios de expressão específicos. De outra forma, não seria possível especificar qual arco realmente define certo conceito e qual é mero ponteiro para o conceito em questão.

No MultiNet, a limitação do escopo do significado é especificada por meio de um encapsulamento de conceitos, o qual foi ilustrado na Figura 29 (pág. 75). Os diferentes componentes contidos em tal cápsula conceitual são graficamente representados como um retângulo dividido em três partes que correspondem aos diferentes tipos de significado ou de conhecimento.

4.2.3.3. Relacionamentos axiomáticos

Uma representação do conhecimento fornece também regras axiomáticas ou axiomas que permitem a realização de inferências. As relações e funções estão ligadas a regras axiomáticas. Uma grande parte dessas regras de inferência não está conectada a elementos de uma língua particular. Uma relação ou função pode assim representar toda uma classe de relações. Por exemplo:

(12) $(x\ \text{CAUS}\ y) \rightarrow \neg\ (y\ \text{ANTE}\ x)$

A regra (12) determina a conexão entre a relação de causalidade CAUS e o antecedente temporal ANTE. Essa regra afirma que se x causa y, então y não pode anteceder x temporalmente.

As regras que não contêm representações de conceitos como argumento, por exemplo, a regra em (12), são chamadas no MultiNet **Axiomas-R** (do inglês, *R-Axioms*).

Há também axiomas que se constrõem com expressões específicas da língua, como o axioma descrito em (13):

(13) (v SUB give1.1) ∧ (v AGT a) ∧ (v OBJ o) ∧ (v ORNT d) →
∃ w (w SUBS receive1.1) ∧ (w OBJ o) ∧ (w AVRT a) ∧ (w EXP d)[41]

A regra em (13) estabelece a ligação entre as unidades lexicais give1.1 e receive1.1 e, ao mesmo tempo, caracteriza a mudança das relações temáticas de ambas as ações.

Vale ressaltar, por fim, que o MultiNet também apresenta alguns axiomas que são considerados regras *default*, ou seja, que cobrem apenas casos típicos; essas regras são aplicadas caso não haja informação conflitante.

(14) (k1 PARS k2)∧ (k2 ORIGM s) → (k1 ORIGM s) [default]

A regra (14), por exemplo, especifica o seguinte: se k2 consiste de certo material, então a parte k1 de k2 também consiste desse material. Entretanto, essa regra é uma afirmação apenas possível, já que um carro de plástico, por exemplo, pode ter rodas de borracha. Os axiomas previstos pelo MultiNet não são, no entanto, tratados neste trabalho.

A seguir, discute-se, ainda que brevemente, a questão da divisão entre conhecimento lingüístico e conhecimento de mundo. Essa discussão faz-se pertinente porque, no âmbito do PLN, o conhecimento sobre as unidades lexicais e o conhecimento sobre os conceitos a elas subjacentes são comumente divididos em **conhecimento lingüístico** e **conhecimento de mundo** (enciclopédico ou conceitual), respectivamente. Essa distinção, em princípio metodológica, é relevante para o processo de interpretação das línguas naturais.

4.3. Representação do conhecimento e léxico: a relação entre conhecimento de mundo e conhecimento lingüístico no MultiNet

A distinção entre conhecimento lingüístico e conhecimento de mundo é uma questão antiga nos estudos da linguagem, sendo investigada por lingüistas, filósofos, professores de língua, cientistas cognitivos, entre outros. A distinção tem sido defendida por uns e atacada por outros e, segundo Peeters (2001), tanto os defensores quanto os acusadores têm desenvolvido fortes argumentos para subsidiar suas hipóteses.

[41] AGT = agente [x é realizado por y]; OBJ = objeto neutro [o objeto x participa passivamente da situação y]; SUBS = relação de subordinação para situações [a situação x é uma especialização da situação y]; AVRT = relação distanciando-se [x está se distanciando de y]; EXP = experienciador [x é experienciado por y]; ORNT = XXXXXXX.

Do ponto de vista do PLN, a separação entre esses dois tipos de conhecimento é operacionalmente relevante, pois torna o desenvolvimento de léxicos computacionais e de bases de conhecimento (conceitual) tarefa modular (CAVAZZA, ZWEIGENBAUM, 1995). Diante de tal relevância, essa distinção é considerada neste trabalho. No MultiNet, como será enfatizado a seguir, a relação entre o conhecimento conceitual e o lingüístico é feita pelos traços semânticos. Antes, porém, uma breve descrição das duas abordagens de tratamento da distinção entre conhecimento lingüístico e conhecimento de mundo é apresentada.

4.3.1. Abordagem em um nível

Diz-se que aqueles que não acreditam em tal divisão abordam a questão do significado dos itens lexicais em apenas um nível (do inglês, *one-level approach*). Na área da Lingüística Cognitiva, autores como Goddard (1998), Haiman (1980), Fillmore (1977), Langacker (1987), Lakoff (1987), Talmy (1988), entre outros, não vêem razões para a divisão, pois para eles o conhecimento sobre todas as coisas está integrado na mente de tal modo que qualquer divisão desse conhecimento não faria sentido. Para esses autores, a linguagem serve para categorizar o mundo, sendo o significado lingüístico indissociável do conhecimento de mundo e, por isso, não se pode postular a existência de um nível estrutural ou sistêmico de significação distinto de um nível do conhecimento de mundo. Segundo Geeraerts (1988), defender a hipótese da indissociabilidade dos conhecimentos, como o faz a Lingüística Cognitiva, é reflexo de se estudar os conceitos lexicais como parte integrante da cognição humana e não como parte de uma estrutura da língua autônoma dentro da cognição. Dentro dessa abordagem, Langacker (1987) e Lakoff (1987) enfatizam que a distinção entre os conhecimentos lingüístico e não-lingüístico é apenas artificial (técnica) ou, segundo Haiman (1980), uma "profissão de fé".

4.3.2. Abordagem em dois níveis

Aqueles que acreditam em tal distinção, pautando-se em uma abordagem em dois níveis (do inglês, *two-level approach*), fazem-no por meio de uma analogia com a Fonética e Fonologia. Os seres humanos aprendem a reconhecer uma variedade quase infinita de sons, mas, em qualquer língua, somente alguns deles carregam ou transmitem significado. Esses seriam os elementos verdadeiramente lingüísticos. De modo similar, a variedade de material semântico é virtualmente infinita, mas somente um número limitado desse material é lingüisticamente verdadeiro e interage sistematicamente com outros aspectos do sistema lingüístico (CRUSE, 2004). O vasto conhecimento detalhado do mundo, que os falantes inevitavelmente

processam, é, segundo a abordagem em dois níveis, uma propriedade dos conceitos, que são extralingüísticos.

Mais especificamente, autores como Katz e Fodor (1963), Frawley (1981), Wierzbicka (1995), Raskin (1985), entre outros, condenam a impossibilidade da dicotomia. Para Katz e Fodor (1963), em especial, a descrição semântica de um item lexical divide-se em uma parte sistêmica, que reflete "quaisquer relações semânticas sistemáticas existentes entre aquele item (lexical) e o resto do vocabulário da língua", e uma parte assistêmica, específica de cada item lexical, que reflete "o que é idiossincrático quanto ao significado daquele item". Na parte sistêmica, fala-se em "marcadores semânticos", os quais corresponderiam ao conhecimento lexical (lingüístico ou semântico). Na parte idiossincrática do significado lexical, fala-se em "distinguidores", que corresponderiam ao conhecimento conceitual (extralingüístico). Sob esse ponto de vista, a determinação do conhecimento lingüístico associa-se à sintaxe. Por exemplo, saber que *granito* é [concreto] e [-animado] permite construir sentenças como *Eu vendo/ vejo/ limpo granito*, mas não sentenças como *Eu conjugo/ ensino granito*. No caso, traços como [concreto] e [-animado] são *marcadores* e restringem as combinações do item *granito.* O fato de *granito* ser formado por quartzo, feldspato e mica não é relevante para o uso correto do item; essas informações sobre sua constituição são, portanto, *distinguidores*. Assim, os traços permitem a descrição das restrições selecionais e subcategorizações das unidades lexicais.

Uma possibilidade de se entender a relação entre o componente semântico (do léxico) e a base de conhecimento pode ser encontrada em Allan (2001). Segundo o autor, uma entrada típica no LM (ou "léxico hipotético") está representada na forma de uma rede que interliga três tipos de especificações, as quais correspondem às informações: (i) formal (isto é, grafofonológicas), (ii) morfossintática e (iii) semântica de um item em questão. Vale ressaltar apenas que Allan (2001) unifica a informação morfológica contida na dimensão lexemática e a informação sintática contida na dimensão lemática em apenas um tipo e especificação - a morfossintática. A informação conceitual ou enciclopédica, armazenada na base de conhecimento ou simplesmente enciclopédia, seria uma aresta adicional na rede da entrada, como está ilustrado na Figura 34.

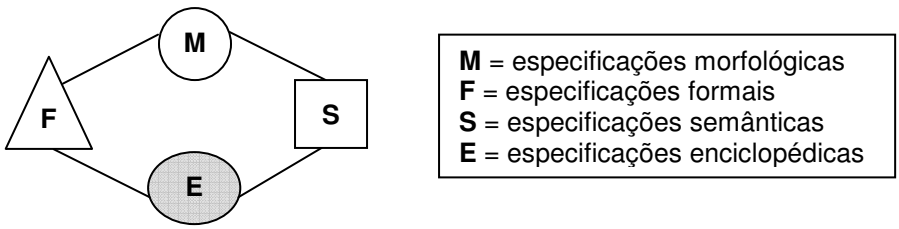


Figura 34: Entrada lexical enriquecida com informação enciclopédica.

Não há, no entanto, consenso a respeito da dicotomia entre o conhecimento de mundo e conhecimento lingüístico. Peeters (2001), aliás, no livro intitulado '*A interface léxico-enciclopédia'* (do inglês, *The lexicon-encyclopedia interface*), inteiramente dedicado à questão, reforça o fato de que uma resolução, se possível, está longe de ser alcançada (PEETERS, 2001).

4.3.3. O MultiNet e o componente semântico do léxico

No Quadro 2, apresenta-se o conjunto básico dos traços usados pelo MultiNet para a caracterização dos objetos[42].

Nome	Significado	Exemplos de valores	
		+	-
ANIMAL	Animal	Raposa	Pessoa
ANIMATE	Ser vivo	Árvore	Pedra
ARTIF	Artefato	Casa	Árvore
AXIAL	Objeto com eixos característicos	Caneta	Esfera
GEOGR	Objeto geográfico	Os Alpes	Mesa
HUMAN	Humano	Mulher	Macaco
INFO	Objeto de informação	Livro	Grama
INSTIT	Instituição	ONU	Maçã
INSTRUT	Instrumento	Martelo	Montanha
LEGPER	Pessoa jurídica ou natural	Firma	Animal
MENTAL	Objeto mental ou situação	Prazer	Extensão
METHOD	Método	Procedimento	Livro
MOVABLE	Objeto que se move	Carro	Floresta
POTAG	Agente potencial	Motor	Mensageiro
SPATIAL	Objeto com extensão espacial	Mesa	Idéia
THCONC	Conceito teórico	Matemática	Prazer

Quadro 2: Os traços para a caracterização dos objetos.

Segundo Helbig (2006), a informação fornecida pelos traços permite uma descrição diferenciada tanto para investigações lexicográficas como para o processamento automático das línguas naturais. Vale ressaltar, no entanto, que todo conjunto pré-definido de traços semânticos, independentemente de seu refinamento, não é inteiramente adequado à tarefa de descrição da semântica das línguas naturais. Naturalmente, sempre há casos para os quais os meios representacionais são muito rústicos para a descrição das restrições de seleção.

[42] Os nomes dos traços serão mantidos em inglês.

Todavia, as aplicações de PLN necessitam de classes como os tipos e os traços, já que não há (ainda) a possibilidade de inserção de todo o conhecimento humano em tais sistemas.

A seguir, os traços expostos no Quadro 2 são brevemente descritos.

- ANIMAL: caracteriza animais.
- ANIMATE: caracteriza entidades vivas, isto é, seres humanos, animais e plantas.
- ARTIF: caracteriza artefatos como <carro> e <mesa>.
- AXIAL: caracteriza que o objeto possui um eixo específico.
- GEOGR: caracteriza geograficamente objetos como <os Alpes> e <Paris>.
- HUMAN: caracteriza seres humanos.
- INFO: caracteriza fontes abstratas de informação como <mensagem> e fontes concretas de informação como <livro>.
- INSTIT: caracteriza instituições como <universidade> e <corte>.
- INSTRU: caracteriza instrumentos típicos como <violino> e <martelo>.
- LEGPER: caracteriza pessoas naturais e legais (como <firma>), às quais grupos de pessoas e instituições pertencem.
- MENTAL: caracteriza estados mentais abstratos e eventos como <raiva> e <ansiedade>, que podem ser experienciados ou sentidos.
- METHOD: caracteriza métodos abstratos como <pasteurização>.
- MOVABLE: caracteriza objetos que podem ser movidos. Essas entidades podem ser usadas como objetos de atividades de transporte como <entregar> e <trazer>.
- POTAG: caracteriza agentes que possuem o "poder" de desencadear ou realizar uma atividade.
- SPATIAL: caracteriza objetos que têm uma extensão espacial; eles podem ser <vistos> ou <deformados>.
- THCONC: caracteriza conceitos teóricos, como <lingüística> e <transitividade>, que são construtos mentais dos seres humanos.

Como já apontado na Figura 28, há regularidades no relacionamento entre os tipos e os traços. Helbig (2006, p. 286) fornece um exemplo de parte das dependências entre os tipos e traços (Figura 35).

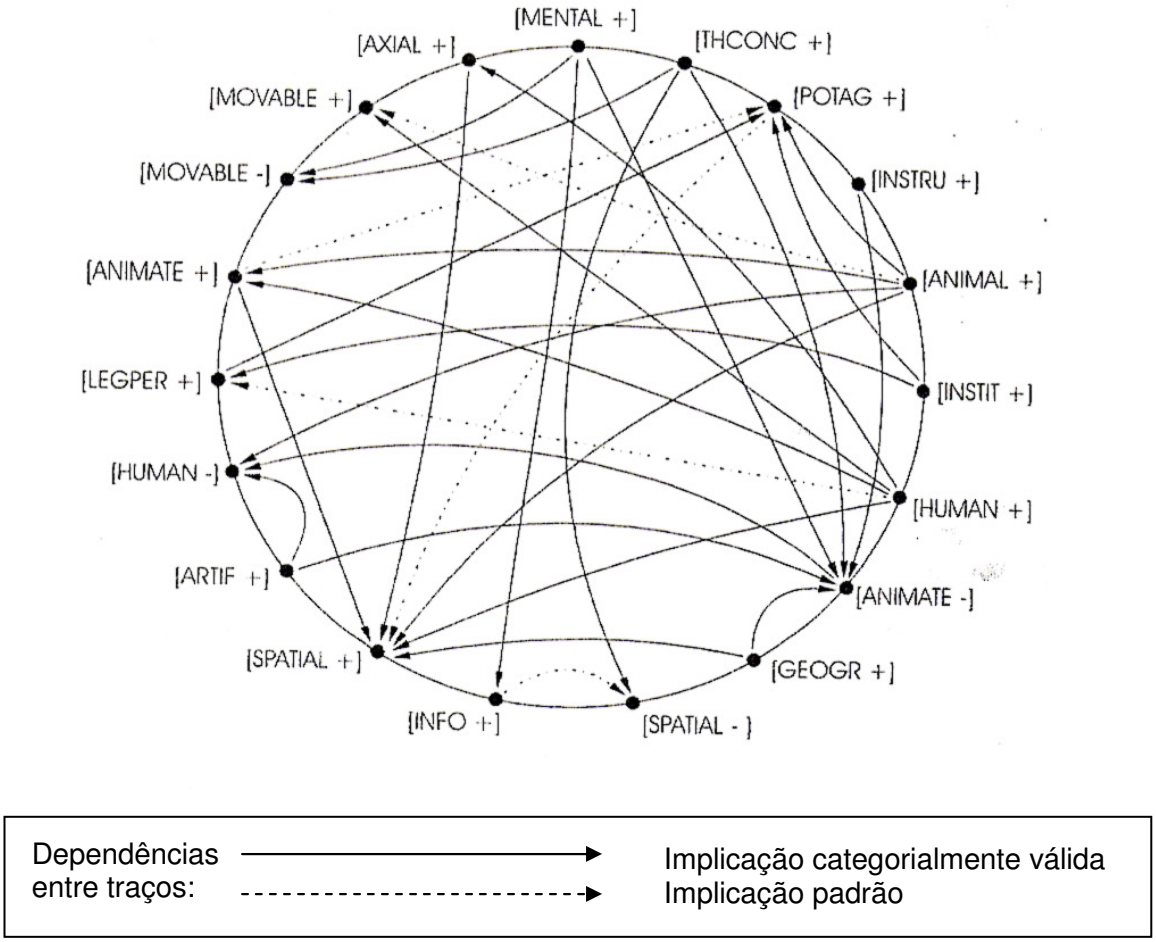


Figura 35: Exemplos de dependências entre tipos e traços.

A Figura 35, mostra que seres humanos [HUMAN+] e animais [ANIMAL+] são sempre seres vivos [ANIMATE+]. Todos os objetos [ANIMATE-] são sempre especificados pelo traço [ARTIF+] e assim por diante.

4.4. Síntese da Seção IV

Nesta Seção, foram apresentados os critérios segundo os quais o paradigma de representação do conhecimento denominado MultiNet foi proposto, seus principais recursos representacionais e os recursos específicos para a ligação entre a representação do conhecimento semântico-conceitual e o léxico.

A *Multilayered Extended Semantic Networks* ou simplesmente MultiNet é um dos poucos modelos que se enquadram no paradigma de RC baseado em redes semânticas. Como conseqüência desse enquadramento, ao se adotar o MultiNet como modelo de representação

formal, adota-se também a abordagem cognitiva do significado. Assim, considera-se que os conceitos estão organizados na mente e que um conceito tem como base um referente prototípico. Assume-se também o formalismo de redes semânticas.

O MultiNet distingue-se dos demais paradigmas de RC principalmente por dois critérios: o da homogeneidade e o da adequação cognitiva. Pelo critério da homogeneidade, os meios de representação do MultiNet são capazes de expressar conceitos lexicalizados e não-lexicalizados (expressos por meio de sintagmas e sentenças) da mesma forma, especificados tanto no nível do significado da expressão (isto é, das expressões em isolado) quanto no do enunciado (isto é, das expressões em contexto). Pelo critério da adequação cognitiva, todo conceito tem uma representação única por meio da qual toda a informação a ele associada torna-se acessível.

Quanto à caracterização dos conceitos, salienta-se que o MultiNet fornece um conjunto de recursos que permite uma representação refinada desses elementos. Os meios de representação da estrutura conceitual, responsáveis pela conexão entre os conceitos, que são os nós da rede, pautam-se em um conjunto pré-definido de 110 relações e funções semânticas. Já os meios de classificação, responsáveis pela caracterização dos conceitos, caracterizam-se por sua inclusão em um espaço multidimensional de atributos, os quais buscam descrever várias facetas do significado.

Devido também ao critério da interoperabilidade, o MultiNet tem sido usado no âmbito do PLN, principalmente como interlíngua semântica de interfaces de busca (em língua natural) em bases de dados (HELBIG, 2006).

Devido ao fato de os meios representacionais do MutiNet serem independentes de língua, eles podem ser usados para representar, no caso deste trabalho, os conceitos que compõem a interlíngua para o alinhamento dos conceitos lexicalizados no AmE e no PB.

A seguir, apresentam-se os meios representacionais fornecidos pelo MultiNet para o tratamento específico dos objetos concretos discretos.

Seção V

O MultiNet e a representação dos objetos

5.1. A caracterização semântica dos objetos

O tipo de conceito denominado **objeto** (abreviado para *o*) engloba os subtipos **concretos** e **abstratos**, como <casa> e <teoria>, respectivamente (cf. págs. 69-72). Como mencionado na Seção I deste trabalho, busca-se investigar os conceitos que pertencem à classe dos objetos concretos discretos. Assim, enfatizam-se os meios representacionais fornecidos pelo MultiNet para o tratamento específico desse tipo de conceito. Vale ressaltar que os meios classificatórios (tipos, traços e atributos) foram descritos na Seção anterior.

Nesta Seção, apresentam-se as principais relações e funções previstas para a caracterização semântica desses conceitos, as quais estão sistematizadas na Figura 36 (HELBIG, 1997, 2006).

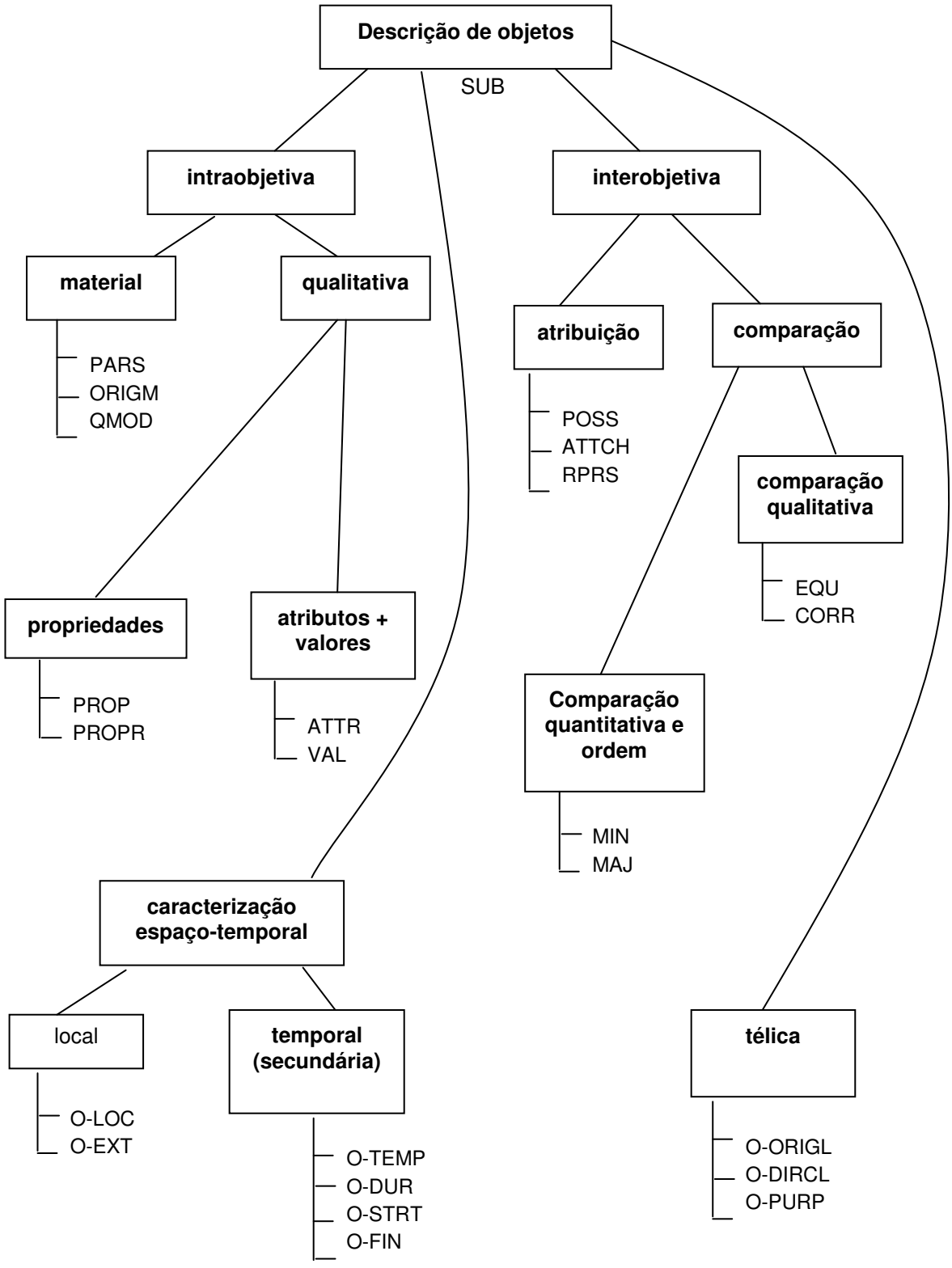


Figura 36: Relações e funções para a caracterização semântica dos objetos concretos.

5.1.1. SUB – relação de subordinação

A relação mais importante que se estabelece entre os conceitos que representam objetos conceituais é a de subordinação (SUB). Essa relação pode ocorrer entre dois conceitos genéricos ou entre um conceito individual e um genérico[43]. A Figura 37, elaborada com base em Helbig (2006, p. 46), ilustra parte de uma hierarquia de objetos conceituais.

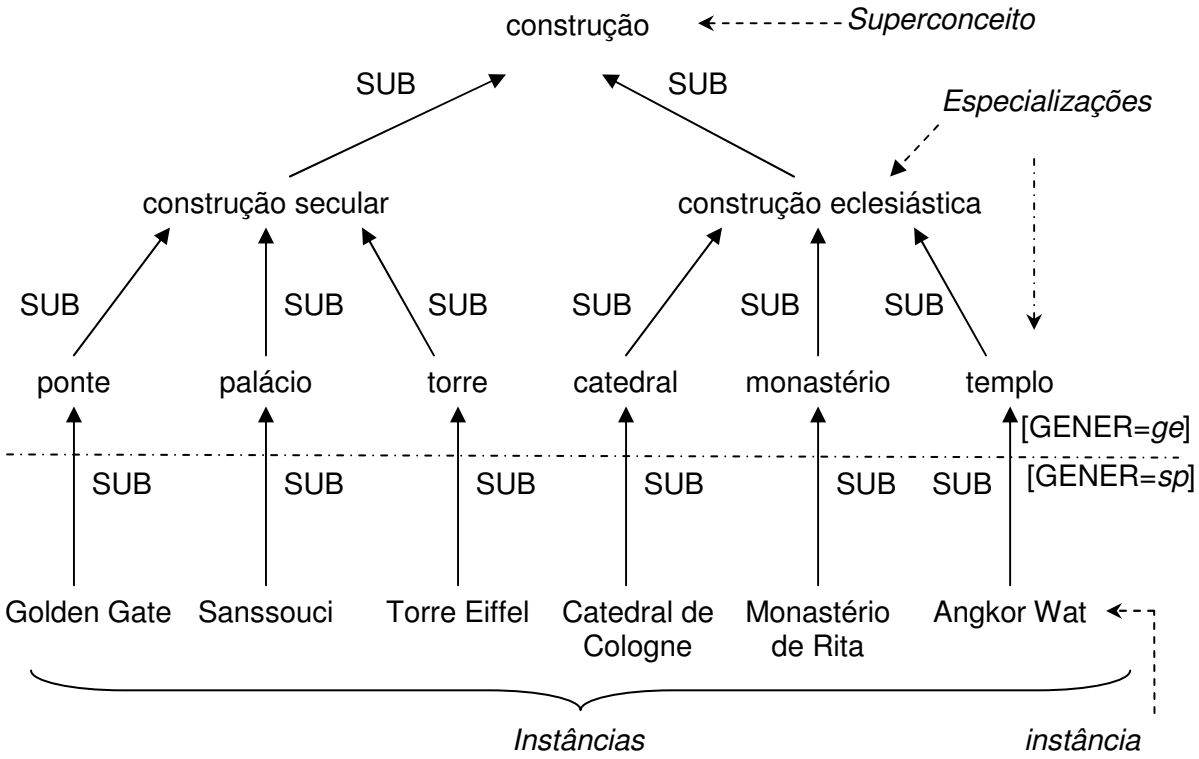


Figura 37: Um exemplo de parte de uma hierarquia de objetos conceituais.

Todos os conceitos subordinados a um conceito superordenado formam uma hierarquia conceitual e, portanto, constituem uma árvore parcial de uma rede semântica. O nó localizado no topo da hierarquia é exatamente o **superconceito**. Os nós terminais da árvore são chamados **instâncias**. Todos os nós de uma hierarquia, com exceção do superconceito, são especializações do superconceito. A hierarquia organizada segundo a relação SUB está conectada a mecanismos de herança por meio dos quais informações do superconceito são transferidas para as suas especializações. Nesse cenário, dois mecanismos de herança são especificados (Figura 38).

[43] Ver estrutura *qualia* do modelo do Léxico Gerativo (PUSTEJOVSKY, 1996). Mais especificamente, o *quale* FORMAL.

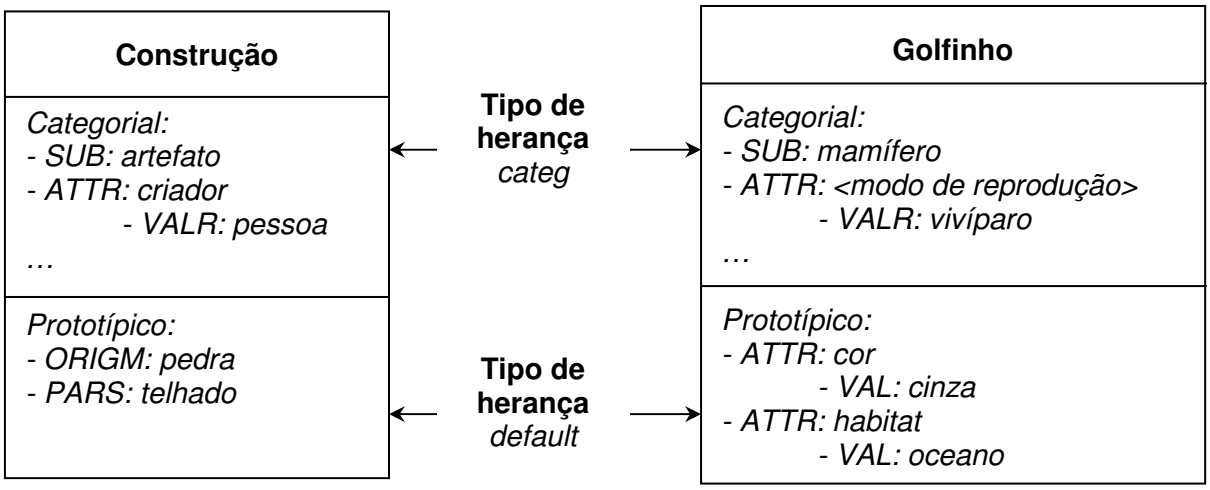


Figura 38: Dois mecanismos de herança.

De um lado, há o conhecimento categorial, que é herdado sem exceções por todos os subconceitos da hierarquia, isto é, por todas as especializações e instâncias. A essa herança é dado o nome de **monotônica** (ROCA, 2000). As informações, por exemplo, de que uma construção é um artefato e tem um criador pertencem ao conhecimento categorial.

Por outro lado, encontra-se o conhecimento conceitual que é típico de parte dos objetos de uma categoria; tal como o fato, por exemplo, de uma construção ser de pedra. Uma construção pode ser construída de madeira, de tijolos, etc. Tal conhecimento é tido como prototípico e, conseqüentemente, caracteriza-se como uma "hipótese *default*", herdada por meio de regras especiais, o que é muito econômico para uma representação do conhecimento. Ao contrário do conhecimento categorial, o conhecimento prototípico, herdado por especializações e instâncias, pode ser sobrescrito se houver informação contraditória disponível em um nível inferior da hierarquia (p.ex.: ORIGM →madeira). A esse mecanismo de herança é dado o nome **não-monotônico** (ROCA, 2000).

5.1.2. Caracterização intraobjetiva

5.1.2.1. Caracterização material

O material e a estrutura física dos referentes desempenham papel fundamental na caracterização do conceito correspondente. Para a caracterização de tais objetos conceituais, o MultiNet fornece as seguintes relações: PARS para a descrição da relação parte-todo, ORIGM

para a descrição do material ou substância do qual o objeto é feito e QMOD para a especificação de uma quantidade de objetos[44].

a) **PARS – relação parte-todo:** para objetos concretos, PARS é usada para especificar os componentes físicos que constituem esses objetos. Essa relação pode ser traduzida para (x PARS y) → [x é parte de y]. Por exemplo, essa é a relação que ocorre entre <x: porta> e <y: casa>. Quanto ao tipo de conhecimento (atributo K-TYPE) do arco rotulado por PARS, salienta-se que o arco, ao partir de x e ao chegar em y, é especificado como **proto** (de prototípico).
b) **ORIGM – relação substância de origem:** a expressão (x ORIGM y) especifica a origem material (ou substancial) de x a partir de y; pode ser traduzida para (x ORIGM y) → [x consiste de y].
c) **QMOD – relação modificação quantitativa:** a relação QMOD descreve o aspecto quantitativo da composição do material; pode ser traduzida para (d QMOD q) → [d é quantitativamente determinado por q].

5.1.2.2. Caracterização qualitativa

a) **PROP – relação objeto-propriedade:** a asserção (x PROP y) estabelece uma conexão entre um objeto x e uma propriedade y. A relação PROP pode ser traduzida para (x PROP y) → [x tem a propriedade y].
b) **PROPR – relação plural-propriedade:** a asserção (g PROPR r) é usada para caracterizar uma pluralidade[45] g por meio de uma propriedade r. A relação PROPR (g PROPR r) pode ser traduzida para → [a pluralidade g é caracterizada pela propriedade r].
c) **ATTR – relação designação atributo-objeto:** a relação (o ATTR m) especifica que m é uma característica atributiva de o, pertencente ao conhecimento imanente. A relação ATTR pode ser traduzida para (o ATTR m) → [o tem um atributo m].
d) **VAL – relação atributo-valor:** a expressão (x VAL v) estabelece uma relação entre um atributo, que é específico de um objeto o e um conceito v, que é o valor desse atributo em relação a o. VAL pode ser traduzida para (x VAL y) → [o atributo (específico) x tem o valor y].

---

[44] Ver estrutura *qualia* do modelo do Léxico Gerativo (PUSTEJOVSKY, 1996). Mais especificamente, o *quale* CONSTITUTIVO.
[45] No nível pré-extensional, uma pluralidade é um conjunto. O valor do atributo CARD é ≥ 2.

5.1.3. Caracterização interobjetiva

A caracterização interobjetiva especifica o relacionamento entre diferentes objetos. Tal caracterização prevê dois tipos de relações: atribuição e comparação.

5.1.3.1. Atribuição

a) **POSS – relação de posse:** a asserção (o1 POSS o2) expressa a conexão entre o possuidor o1 e a coisa possuída o2; pode ser traduzida para (o1 POSS o2) → [o1 é possuidor de o2].
b) **ATTCH – relação de vinculação de um objeto a outro:** a asserção (o1 ATTCH o2) indica que o1 é situacionalmente associado ou vinculado a o2; pode ser traduzida para (o1 ATTCH o2) → [o objeto o1 é associado a o2].
c) **RPRS – relação de forma de representação ou manifestação de um objeto**: a asserção (o1 RPRS o2) especifica que o1 ocorre (no texto) na forma de representação ou manifestação o2. Pode ser traduzida para (o1 RPRS o2) → [o1 ocorre na forma o2].

5.1.3.2. Comparação

A comparação pode ser quantitativa e qualitativa. Para a primeira, o MultiNet fornece as seguintes relações:

a) **MIN – relação "menor que":** a asserção (q1 MIN q2) significa que a quantidade q1 é menor que o número ou quantidade q2. Pode ser traduzida para (q1 MIN q2) → [q1 é menor que q2].
b) **MAJ – relação "maior que":** a asserção (q1 MAJ q2) significa que o número ou quantidade q1 é maior que o número ou quantidade q2. Pode ser traduzida para (q1 MAJ q2) → [q1 é maior que q2].

Para a caracterização da comparação qualitativa, o MultiNet fornece as seguintes relações:

a) **EQU – relação de equivalência:** a expressão (e1 EQU e1) significa que as entidades e1 e e2 são iguais nos planos intensional e pré-extensional; pode ser traduzida para (e1 EQU e2) → [e1 é semanticamente equivalente a e2].
b) **CORR - relação de correspondência qualitativa ou quantitativa:** a asserção (e1 CORR e2) é usada para especificar uma correspondência qualitativa ou quantitativa entre duas entidades. Essa relação pode ser traduzida para (e1 CORR e2) → [e1 corresponde a e2].

5.1.4. Caracterização espaço-temporal[46]

5.1.4.1. Caracterização espacial

A asserção (o LOC l) especifica que o objeto o está em um local l. Formalmente, essa relação pode ser traduzida para (o LOC l) → [o está localizado em l].

5.1.4.2. Caracterização temporal

A caracterização temporal dos objetos divide-se em quatro relações, TEMP, DUR, STRT e FIN, as quais, aliás, podem ser descontínuas.

a) **TEMP – relação de tempo:** a asserção (o TEMP t) especifica o intervalo de tempo ou o momento da existência de o. Essa relação pode ser traduzida para (o TEMP t) → [o existe no momento t].
b) **DUR – relação extensão de tempo ou duração:** a asserção (o DUR t) especifica a duração da existência de o; pode ser traduzida para (o DUR t) → [o existe durante o intervalo de tempo t].
c) **STRT – relação tempo de início:** a asserção (o STRT t) especifica o tempo em que o começa a existir; pode ser traduzida para (o STRT t) → [o existe desde t ou começou em t].
d) **FIN – relação tempo de término:** a asserção (o FIN t) especifica o fim da existência de o; pode ser traduzida para (o FIN t) → [o é terminado em t ou tem o fim em t].

5.1.5. Caracterização télica[47]

No MultiNet, a caracterização télica de um objeto pode ser feita por meio de três relações, ORIGL, DIRCL e PURP.

a) **ORIGL – relação local de origem:** a asserção (o ORIGL l) especifica o local de origem l de um objeto o; pode ser traduzida para (o ORIGL l) → [o é caracterizado pelo local de origem l].

[46] A Figura 36 mostra que os rótulos das relações que descrevem a caracterização espaço-temporal (e télica) dos objetos são antecedidos por um O maiúsculo (p.ex.: O-LOC). Essa notação indica que essas relações se estabelecem no âmbito dos objetos e não no âmbito das situações. Tendo em vista que, neste trabalho, enfatizam-se apenas os meios representacionais relativos aos objetos, o O distintivo da nomeação das relações foi excluído.
[47] Ver *quale* TÉLICO do Léxico Gerativo (PUSTEJOVSKY, 1996).

b) **DIRCL – relação local de destino (direção):** a asserção (o DIRCL l) especifica o local de destino ou direção espacial dos objetos; pode ser traduzida para (o DIRCL l) → [o é espacialmente descrito por l].
c) **PURP – relação de propósito:** a asserção (o PURP p) significa que um objeto tem a função p. Por exemplo, essa é a relação que ocorre entre <a ferramenta> e <remover rodas> em *A ferramenta é usada para remover rodas.* A relação PURP pode ser traduzida para (o PURP p) → [o tem a função ou finalidade p]. Quanto ao K-TYPE do arco rotulado por PURP, salienta-se que ele parte de o e chega em p como *situa*.

Na próxima Seção, além dos meios representacionais para a caracterização semântica dos objetos concretos, apresentam-se os meios representacionais do nível pré-extensional. Esses meios não são exclusivos da caracterização dos objetos concretos, mas também se aplicam a eles.

5.2. As relações e funções do nível pré-extensional

Os meios representacionais do nível pré-extensional dividem–se em relações e funções de conjunto e relações de conexão com o nível intensional. Tais meios estão sistematizados na Figura 39.

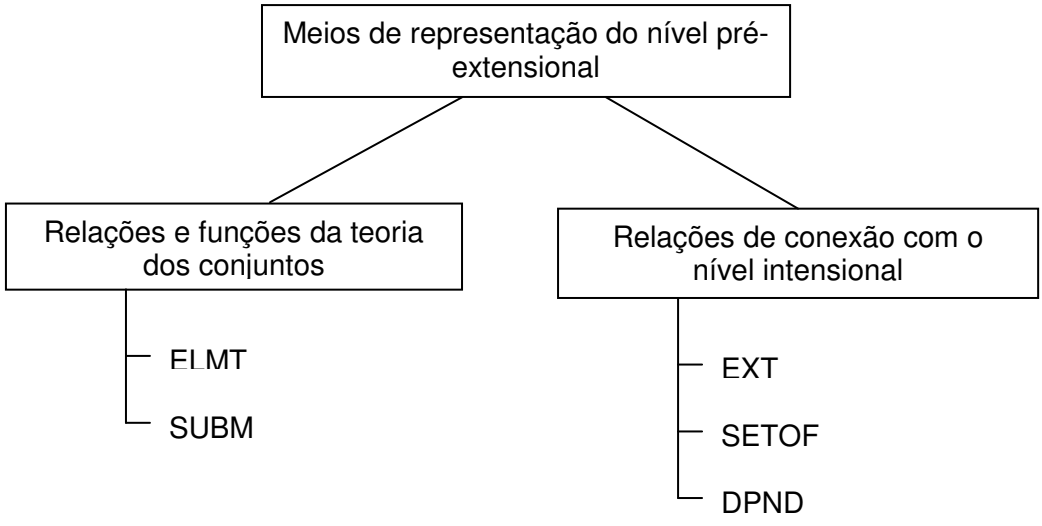


Figura 39: As relações e funções do nível pré-extensional.

5.2.1. As relações e funções no domínio dos conjuntos

As relações e funções no âmbito dos conjuntos de elementos pertencem exclusivamente ao nível pré-extensional e são divididas em ELMT (relação "elemento de um conjunto") e SUBM (relação de "subordinação de conjuntos"). A asserção (e ELMT g) especifica que a

entidade e é um membro de um conjunto. Essa relação acontece, por exemplo, entre <Lucas> e <filhos de Sônia> em [*Lucas*]^{arg1} *é um dos* [*filhos de Sônia*]^{arg2}. A relação ELMT pode ser traduzida para (e ELMT g) → [e é um elemento de g]. A asserção (g1 SUBM g2) especifica que g1 está completamente contida, mas não é idêntica a g2. Essa relação ocorre, por exemplo, entre <várias laranjas> e <quatro> em *Mariana comprou* [*várias laranjas*]^{arg2}, [*quatro*]^{arg1} *estavam podres.* A relação SUBM pode ser traduzida para (g1 SUBM g2) → [g1 é um conjunto parcial de g2].

5.2.2. As relações de conexão entre os níveis intensional e pré-extensional

As relações responsáveis por conectar os níveis intensional e pré-extensional são EXT (relação entre elementos dos dois níveis), SETOF (relação entre um elemento do nível pré-extensional e um conceito genérico) e DPND (relação de dependência entre elementos pré-extensionais). A asserção (e1 EXT e2) estabelece a conexão entre uma representação do nível pré-extensional e o seu correspondente no nível intensional. Essa relação pode ser traduzida para (x EXT y) → [x tem extensão y]. A asserção (g SETOF c) estabelece a conexão entre um conjunto, isto é, uma entidade do nível pré-extensional, e um conceito genérico; significa que todos os elementos de g são extensões do conceito c no nível intensional. Essa relação pode ser traduzida para (g SETOF c) → [g é um conjunto-extensão de c]. A asserção (e1 DPND e2) caracteriza que o elemento é do nível pré-extensional e dependente do elemento e2, também do nível pré-extensional. É o que acontece, por exemplo, entre <todo estudante> e <uma monografia> em *Todo estudante escreveu uma monografia.* Essa relação pode ser traduzida para (x DPND y) → [y depende de x].

5.3. Síntese da Seção V

Nesta Seção, foram descritos os meios para a estruturação de conceitos fornecidos pelo MultiNet para a caracterização específica dos objetos. Tais meios são de dois tipos, os de caracterização semântica e os de caracterização pré-extensional. Os primeiros, por sua vez, podem ser de dois tipos, os de caracterização interobjetiva e intraobjetiva.

Na próxima Seção, diante da seleção do conjunto de conceitos lexicalizados no AmE e do critério de delimitação dos mesmos, explicita-se que conjunto de relações de caracterização interobjetiva e intraobjetiva é considerado neste trabalho.

Neste ponto, no entanto, já é possível afirmar que os meios representacionais do nível pré-extensional não serão considerados. Isso se deve ao fato de que tais meios são importantes, por exemplo, para a interpretação de conceitos como <todos os x exceto y> ou <três deles>, os quais não fazem parte desta pesquisa.

Na próxima Seção, identificam-se os "objetos concretos discretos", foco da investigação empírica deste trabalho, e suas expressões no PB.

Seção VI

A identificação de parte dos "objetos concretos discretos" e sua expressão do PB

6.1. Delimitação do tipo conceitual

Dentre os vários tipos conceituais, focalizam-se, neste trabalho, os do tipo **objetos concretos discretos**.

Segundo a classificação de tipos do MultiNet, eles são do tipo ([SORT=*d*]) e, segundo a classificação de Lyons (1977), são entidades de 1ª ordem. Assim, esses conceitos intuitivamente categorizam referentes perceptíveis pelos sentidos, localizados no tempo e no espaço, que são contáveis e indivisíveis. Quanto à expressão lingüística, os objetos concretos discretos realizam-se por meio de expressões nominais simples, compostas e complexas (LYONS, 1977; JACKENDOFF, 2002).

A escolha dessa classe de conceitos justificou-se pelo fato de poderem ser hierárquica e formalmente sistematizados, possibilitanto assim um teste e aplicação do modelo para futuras aplicações em domínios mais complexos.

6.2. Delimitação do domínio conceitual

Partindo-se do princípio de que os conceitos não estão isolados na mente, mas sim organizados, tomou-se, como ponto de partida para o processo de alinhamento léxico-conceitual, a escolha de um domínio ou campo conceitual específico. Na literatura lingüística, mais especificamente nas pesquisas relacionadas à Lingüística Cognitiva, encontram-se vários domínios conceituais bastante estudados como, por exemplo, o domínio das cores, dos utensílios domésticos, dos móveis, entre outros.

Neste trabalho, selecionou-se o domínio dos "veículos com rodas". A escolha desse domínio não se justifica por questões teóricas, mas sim práticas: a delimitação ser bem-definida e a extensão ser reduzida. Essas características facilitaram as tarefas de análise e representação léxico-conceitual.

Mas, afinal, quais são os veículos com rodas? Segundo a WordNet de Princeton, um veículo com rodas é "um tipo de veículo que se move sobre rodas e serve para transportar pessoas ou coisas". Esse, então, será o domínio conceitual que se enfoca neste trabalho.

6.3. Seleção do conjunto de conceitos expressos no AmE

Tomou-se como ponto de partida para este trabalho, a base da WN.Pr. Vale ressaltar que essa base está em constante modificação. Atualmente, ela está disponível em sua versão 3.0[48]. No momento da elaboração do projeto deste trabalho, a versão que estava disponível era a 2.1. Assim, devido à necessidade de congelamento dos dados para análise, optou-se por manter a versão 2.1 como fonte de informação.

A escolha da WN.Pr como fonte de conceitos teve várias motivações. A primeira delas foi o fato de que os *synsets* são construídos para codificar conceitos "lexicalizados" no AmE, posto que são formados por unidades lexicais que devem expressar o mesmo conceito. A segunda motivação diz respeito ao fato de que os *synsets* da WN.Pr estão organizados em função das categorias sintáticas e de campos conceituais, tais como "artefatos", "pessoas", "objetos naturais", "alimentos" , etc., o que permite selecionar apenas o conjunto dos conceitos do domínio dos "veículos com rodas", expressos por nomes. A terceira motivação liga-se ao fato de que os conceitos estão organizados em função de várias relações léxico-conceituais no interior de cada campo conceitual; assim, o conjunto de conceitos extraídos da WN.Pr pode ser definido como os conceitos constitutivos da interlíngua utilizada no processo de alinhamento. A quarta motivação está relacionada ao fato de que, ao (i) escolher um conjunto de conceitos já armazenados na WN.Pr e (ii) identificar as lexicalizações dos mesmos no PB e as relações entre os conceitos lexicalizados no AmE e no PB, pode-se contribuir diretamente para o alinhamento das bases da WN.Br e WN.Pr. Outra motivação para a escolha da WN.Pr como ponto de partida foi o fato de que esta é uma rede semântica e, por isso, os conceitos nela contidos podem ser reestruturados em termos do MultiNet.

Para a efetiva seleção dos conceitos, foram compilados todos os *synsets* hipônimos do *synset* {wheeled vehicle} da WN.Pr (versão 2.1).

Vale ressaltar, neste ponto, que, dentre os conceitos hipônimos de {wheeled vehicle}, foram identificados 12 conceitos não-lexicalizados no AmE. Essa identificação foi feita com base na verificação da presença das expressões que codificam tais conceitos nos seguintes dicionários monolíngües do AmE: *'Cambridge Dictionary of American English'* (LANDAU, 2000) e o

[48] http://wordnet.princeton.edu/

'Longman Dictionary of Contemporary English Online'[49] (LDOCE-*Online*) (SUMMERS, 2005). Em outras palavras, expressões que representam tais conceitos não constavam como entrada ou subentrada nos referidos dicionários e, por isso, foram consideradas expressões não-lexicalizadas. Tais conceitos, no entanto, foram mantidos na análise desenvolvida neste trabalho, posto que auxiliam na organização conceitual. Até mesmo possíveis expressões no PB de tais conceitos foram identificadas. A ausência de expressões lexicais no PB para esses conceitos, entretando, não caracteriza lacuna lexical do PB.

Assim, dos 217 conceitos selecionados da WN.Pr, 205 são lexicalizados e constituem o conjunto efetivamente analisado neste trabalho. Os 217 conceitos estão listados no Quadro 3, sendo que os não-lexicalizados estão destacados com letras maiúsculas.

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
=> safety bicycle; safety bike
=> velocipede
=> boneshaker
=> car; railcar; railway car; railroad car
=> baggage car; luggage van
=> cabin car; caboose
=> club car; lounge car
=> freight car
=> boxcar
=> stockcar
=> cattle car
=> coal car
=> flatcar; flatbed; flat
=> gondola car; gôndola
=> refrigerator car
=> tank car; tank
=> guard's van
=> handcar
=> mail car
=> passenger car; coach; carriage
=> dining car; diner; dining compartment; buffet car
=> NONSMOKING CAR
=> parlor car; parlour car; drawing-room car; palace car; chair car
=> Pullman; Pullman car
=> sleeping car; sleeper; wagon-lit
=> smoking car; smoking carriage; smoking compartment
=> slip coach; slip carriage
=> tender
=> handcart; pushcart; cart; go-cart
=> applecart
=> barrow; garden cart; lawn cart; wheelbarrow

[49] O *'Longman Dictionary of Contemporary English Online'* é a versão *online* do cd-rom que acompanha a 4ª edição do dicionário impresso *'Longman Dictionary of Contemporary English'*, publicada em 2005. A versão *online* pode ser consultada no endereço http://www.ldoceonline.com/.

=> hand truck; truck
=> laundry cart
=> serving cart
=> pastry cart
=> tea cart; teacart; tea trolley; tea wagon
=> shopping cart
=> HORSE-DRAWN VEHICLE
=> carriage; equipage; rig
=> barouche
=> brougham
=> buckboard
=> buggy; roadster
=> cab; cabriolet
=> caroche
=> chaise; shay
=> chariot
=> clarence
=> coach; four-in-hand; coach-and-four
=> stagecoach; stage
=> droshky; drosky
=> gharry
=> gig
=> hackney; hackney carriage; hackney coach
=> four-wheeler
=> remise
=> hansom; hansom cab
=> landau
=> post chaise
=> stanhope
=> surrey
=> trap
=> troika
=> chariot
=> limber
=> sulky
=> motor scooter; scooter
=> scooter
=> SELF-PROPELLED VEHICLE
=> ARMOURED VEHICLE
=> armored car; armoured car
=> armored car; armoured car
=> ARMORED PERSONNEL CARRIER; ARMOURED PERSONNEL CARRIER; APC
=> assault gun
=> tank; army tank; armored combat vehicle; armoured combat vehicle
=> panzer
=> tank destroyer
=> carrier
=> forklift
=> locomotive; engine; locomotive engine; railway locomotive
=> DIESEL LOCOMOTIVE
=> DIESEL-ELECTRIC LOCOMOTIVE; DIESEL-ELECTRIC
=> DIESEL-HYDRAULIC LOCOMOTIVE; DIESEL-HYDRAULIC
=> dinky; dinkey
=> ELECTRIC LOCOMOTIVE
=> iron horse
=> pilot engine
=> shunter
=> steam locomotive
=> switch engine; donkey engine
=> tank engine; tank locomotive

```
      => traction engine
   => motor vehicle; automotive vehicle
      => amphibian; amphibious vehicle
         => swamp buggy; marsh buggy
      => bloodmobile
      => car; auto; automobile; machine; motorcar
         => ambulance
            => funny wagon
         => beach wagon; station wagon; wagon; beach waggon; station waggon; waggon
         => bus; jalopy; heap
         => cab; hack; taxi; taxicab
            => gypsy cab
            => minicab
         => compact; compact car
         => convertible
         => coupe
         => cruiser; police cruiser; patrol car; police car; prowl car; squad car
            => panda car
         => electric; electric automobile; electric car
         => gas guzzler
         => hardtop
         => hatchback
         => hot rod; hot-rod
         => jeep; landrover
         => limousine; limo
            => berlin
         => loaner
         => minicar
         => Model T
         => pace car
         => racer; race car; racing car
            => stock car
         => roadster; runabout; two-seater
         => sedan
            => brougham
         => sports car; sport car
         => sport utility; sport utility vehicle; S.U.V.; SUV
         => STANLEY STEAMER
         => subcompact; subcompact car
         => touring car; phaeton; tourer
         => used-car; secondhand car
      => doodlebug
      => four-wheel drive; 4WD
      => go-kart
      => golfcart; golf cart
      => hearse
      => motorcycle; bike
         => minibike; motorbike
            => moped
         => trail bike; dirt bike; scrambler
      => snowplow; snowplough
      => truck; motortruck
         => dump truck; dumper; tipper truck; tipper lorry; tip truck; tipper
         => fire engine; fire truck
            => ladder truck; aerial ladder truck
         => garbage truck; dustcart
         => lorry; camion
         => pickup; pickup truck
            => technical
         => sound truck
```

```
            => tow truck; tow car; wrecker
            => trailer truck; tractor trailer; trucking rig; rig; articulated lorry; semi
               => tandem trailer
            => transporter; car transporter
            => van
               => bookmobile
               => delivery truck; delivery van; panel truck
               => laundry truck
               => milk float
               => moving van
               => passenger van
                  => minivan
               => police van; police wagon; paddy wagon; patrol wagon; wagon; black Maria
      => RECONNAISSANCE VEHICLE; SCOUT CAR
      => recreational vehicle; RV; R.V.
         => camper; camping bus; motor home
            => van; caravan
         => dune buggy; beach buggy
      => streetcar; tram; tramcar; trolley; trolley car
         => horsecar
      => tracked vehicle
         => Caterpillar; cat
         => half track
         => snowmobile
            => Sno-cat
      => tractor
         => bulldozer; dozer
            => angledozer
         => skidder
      => weapons carrier
   => skateboard
   => trailer; house trailer
      => camper trailer
      => mobile home; manufactured home
   => tricycle; trike; velocipede
      => pedicab; cycle rickshaw
   => unicycle; monocycle
   => wagon; waggon
      => bandwagon
      => cart
         => dogcart
         => dumpcart
            => tumbrel; tumbril
         => horse cart; horse-cart
            => dray; camion
         => jaunting car; jaunty car
         => jinrikisha; ricksha; rickshaw
         => oxcart
         => pony cart; ponycart; donkey cart; tub-cart
         => water cart
            => watering cart
      => chuck wagon
      => covered wagon; Conestoga wagon; Conestoga; prairie wagon; prairie schooner
      => ice wagon; ice-wagon
      => lorry
      => milk wagon; milkwagon
      => tramcar; tram
      => wain
   => wagon; coaster wagon
```

Quadro 3: As expressões no AmE dos conceitos do tipo <wheeled vehicle>.

No Quadro 3, o símbolo "=>" indica o *synset*, cujas unidades lexicais estão separadas por ponto-e-vírgula, e as diferenças de adentramento do símbolo "=>" indicam os diferentes níveis em que os conceitos se encontram na hierarquia.

6.4. Especificação dos conceitos lexicalizados no AmE

6.4.1. Critérios teóricos

6.4.1.1. Nível de significado

Para a identificação e alinhamento dos conceitos lexicalizados no AmE e no PB, toma-se como ponto de partida um conjunto de conceitos (representados pelos *synsets*) armazenado na base da WN.Pr. Os *synsets* dessa base codificam ou representam conceitos genéricos, como <casa> e <bicicleta>, delimitados na dimensão do **significado da expressão**. Assim, os *synsets* capturam o potencial de significação das unidades lexicais que o compõem.

Para delimitar o nível do significado da expressão, considera-se a sentença declarativa (15) *Eu quero sua bicicleta*. Sem grandes esforços, o leitor do português brasileiro reconhece a sentença (15) como sendo pertencente a sua língua, interpretando-a e provavelmente imaginando até mesmo uma situação comunicativa em que ela pode ser usada. Neste ponto, a pergunta que se faz é: qual é o significado dessa sentença? A partir do momento em que o leitor a interpreta, ele sabe qual é o seu significado. Entretanto, saber o seu significado não implica saber descrevê-lo. Isso, aliás, acontece com quase todo o nosso conhecimento. Por exemplo, pode-se decorar uma música, mas isso não implica saber descrever sua melodia. Descobrir, desvendar o conhecimento semântico dos itens lexicais e sentenças e descrever ou revelar sua natureza são objetivos da Semântica (ALLAN, 2001; LÔBNER, 2002; CRUSE, 2004). Um procedimento possível para se determinar o significado da sentença (15) é começar pela identificação do significado das unidades que a compõem. Uma distinção óbvia a se fazer é aquela entre o significado das palavras e o das sentenças e entre o significado gramatical e o lexical.

O pronome pessoal *eu,* por exemplo, é usado para se referir àquele que diz, mais precisamente, àquele que produz a sentença. A função do pronome *eu*, então, é identificar o emissor da sentença. De modo similar, o pronome possessivo *sua* expressa algum tipo de relação entre a entidade "possuída" e o seu "possuidor" (no caso, o receptor). No caso de (15), *sua* indica que a bicicleta desejada é a do receptor(a). Nesses casos, diz-se que tais elementos

são funcionais (do inglês, *functional words*) ou gramaticais (do inglês, *grammatical words*) e expressam significado gramatical, tendo como principal função articular a estrutura gramatical das sentenças (LYONS, 1977; PERINI, 1999; CRUSE, 2004).

Os elementos *quero* e *bicicleta*, por sua vez, portam a maior carga de informação da sentença, trazendo em si alguma representação do mundo (real ou imaginário) (LYONS, 1977; PERINI, 1999; CRUSE, 2004). Nesses casos, diz-se que eles são elementos lexicais e, conseqüentemente, expressam significado lexical.

Mas, afinal, qual é o significado do verbo *querer* e do nome *bicicleta* (15) (*Eu quero sua bicicleta*)? Nessa sentença, o verbo *querer* é usado com um objeto direto (*sua bicicleta*) e significa <ter vontade de, desejar>. Na sentença (15), o que é desejado é expresso por *sua bicicleta*, ou seja, uma expressão composta pelo pronome possessivo *sua* e pelo nome *bicicleta*. O nome *bicicleta* significa aproximadamente <veículo que contém duas rodas raiadas, um quadro, um selim, e é movido por pedais>.

Observa-se que a descrição do significado dos itens lexicais não é tarefa trivial, pois deve ser suficientemente específica para distingüi-los de todos os outros itens com significados diferentes. Por exemplo, não seria adequado, por um lado, descrever o significado de *bicicleta* genericamente como <veículo de duas rodas>, já que há outros veículos de duas rodas que não são bicicletas; por outro lado, não seria adequado descrevê-lo de modo tão específico como <veículo vermelho de duas rodas>, já que há bicicletas que não são vermelhas.

Se todas as informações relevantes forem "calculadas", obtém-se o significado da sentença (15), *Eu quero sua bicicleta*. Esse significado pode ser assim descrito: <o emissor deseja, por alguma razão, o veículo que contém duas rodas, um quadro, um selim, e é movido por pedais, pertencente ao receptor, no momento em que produz (15)>.

Neste ponto, ressalta-se que a sentença (15), como tal, não fornece informações, por exemplo, a respeito do emissor, do receptor e da bicicleta em questão. Tais informações só podem ser obtidas se a sentença (15) for usada em uma situação de comunicação concreta. O significado de palavras, sintagmas ou sentenças, divorciado de um contexto particular, constitui o nível de significação a que Lôbner (2002) e Cruse (2004) denominam significado da expressão (do inglês, *expression meaning* ou *sentence meaning*).

No nível do significado da expressão, a determinação do significado da expressão requer uma abstração do uso concreto das expressões. O que se busca descrever é o potencial de significado das expressões. O item lexical *bicicleta*, por exemplo, tem o potencial de ser usado para referir a tudo que possua aquelas propriedades características reunidas no significado dessa expressão.

Neste trabalho, para a delimitação dos conceitos, segue-se a mesma diretriz empregada na WN.Pr, em que estes são especificados no nível do significado da expressão, independentemente de um contexto específico.

6.4.1.2. Tipo de significado

Outro aspecto importante dos *synsets* é o fato de que suas unidades lexicais constitutivas possuem a mesma denotação, independentemente de diferenças dialetais, conotativas, entre outras, que elas possam ter. Dessa forma, a noção de **significado descritivo** é relevante ao se lidar com o construto "*synset*".

Esse tipo de significado recebe várias denominações: significado descritivo (do inglês, *descriptive meaning*), referencial (do inglês, *referential*), lógico (do inglês, *logical*) e proposicional (do inglês, *propositional*) (CRUSE, 2004). Adota-se, aqui, a denominação de Lyons (1981) – significado descritivo – por ser mais adequada à descrição do significado dos itens lexicais. De acordo com Leech (1976), Lyons (1981), Lôbner (2002) e Cruse (2004, 2006), o significado descritivo de uma unidade lexical é a parte de seu significado que restringe a que ela se refere e que determina o valor de verdade de uma proposição que a contém. Assim, esse significado inclui o **significado proposicional**, mas pode ser estendido para incluir características ou traços prototípicos, como o "latido dos cachorros", em vez de simplesmente incluir traços logicamente necessários, como o fato de um cachorro ser um animal. As dimensões mais importantes do signficado descritivo são (i) qualidade (isto é, o que distingue, por exemplo, *amarelo* de *azul*, *cachorro* de *gato*, etc.), (ii) especificidade (isto é, o que distingue, por exemplo, *cachorro* de *animal*, etc.), (iii) intensidade (isto é, o que distingue, por exemplo, pares como *medo: terror* e *desagradar: odiar*, etc.) e (iv) vagueza (isto é, o que distingue, por exemplo, *grande* e *pequeno,* etc.) (VILELA, SILVA, 2004; CRUSE, 2006).

A esse tipo de significado, opõe-se o tipo **não-descritivo** (do inglês, *non-descriptive meaning).* A principal distinção entre o significado descritivo e o não-descritivo reside no fato de que este, ao contrário daquele, não pode ser explicado em termos de verdade (LEECH, 1976).

6.4.1.3. O conjunto de relações consideradas

Como mencionado nesta Seção, considera-se apenas o nível do significado da expressão na delimitação dos conceitos; a especificação dos mesmos é feita, então, independentemente de um contexto específico. Portanto, além da relação de SUB, apenas as relações pertinentes à

caracterização intraobjetiva são consideradas neste trabalho, porque estas são responsáveis por especificar as características intrínsecas desse tipo de conceito. Dessa forma, o conjunto fica restrito às seguintes relações: PARS, ORIGM, QMOD, PROP, PROPR, ATTR e VAL.

Desse elenco, as únicas relações intraobjetivas a não serem utilizadas são PROPR, responsável por relacionar "pluralidades" (enquanto objetos conceitos) a propriedades relacionais; VAL, que estabelece a relação entre um atributo de um objeto e o seu valor, e QMOD, que relaciona o material constitutivo de um objeto a sua quantidade. A exclusão da relação PROPR deve-se ao fato de não se analisar conceitos plurais neste trabalho. A exclusão da relação VAL justifica-se por esta caracterizar os objetos conceituais de modo indireto, pois ela especifica, na verdade, o valor de um atributo de um objeto conceitual; o mesmo se aplica à exclusão da relação QMOD.

Dentre as relações interobjetivas, a relação PURP é a única a ser utilizada. A decisão de se utilizar tal relação para a caracterização dos objetos baseia-se principalmente em Pustejovsky (1996, 2001). Os conceitos estudados neste trabalho são do tipo concreto e discreto e, como se vê a seguir, são do subtipo dos artefatos. Segundo Pustejovsky, a relação PURP ("relação de propósito") é relevante para a caracterização dos conceitos desse subtipo, como <car> e <hammer>.

6.4.2. Proposição do *template conceitual*

Para a delimitação dos conceitos lexicalizados, elaborou-se um *template* (ou formulário padrão) conceitual composto por 6 campos, sendo que, para cada conceito selecionado na WN.Pr, preencheu-se o *template* conceitual correspondente. O Quadro 4 ilustra o modelo de *template* utilizado.

Tal modelo tem basicamente a finalidade de facilitar e controlar a tarefa manual de delimitação dos conceitos. Tal tarefa poderia ter sido feita com o auxílio de um editor, ou seja, de uma ferramenta computacional de auxílio à inserção e manipulação de dados. No entanto, a definição por um editor que suportasse o formato "multinet" dos dados foi uma tarefa complexa, realizada apenas nas últimas etapas de desenvolvimento deste trabalho. Conseqüentemente, o uso do *template* foi fundamental ao longo de todo o trabalho para garantir a organização do conhecimento conceitual. O mesmo pode ser dito para o o modelo de *template* lexical, descrito na seqüência.

UniC:	*Identificador da unidade conceitual; símbolo mnemônico*	
Glosa	*Definição lexicográfica informal de uma unidade conceitual*	
Tipo conceitual:	*Tipo conceitual* (*MultiNet*)	
Traços semânticos:	*Traços semânticos*	
Atributos multidimensionais	**REFER:**	*Determinação da referência de uma unidade conceitual*
	ETYPE:	*Caracterização do tipo de extensionalidade de uma unidade conceitual*
	VARIA:	*Caracterização da variabilidade de uma unidade conceitual*
	GENER:	*Caracterização da generalidade de uma unidade conceitual*
	FACT:	*Caracterização da factividade de uma unidade conceitual*
Relações:	**SUB:**	*Relação de subordinação entre unidades conceituais*
	PARS:	*Relação de parte-todo entre unidades conceituais*
	ORIGM:	*Relação de material de origem entre unidades conceituais*
	PROP:	*Relação de propriedade entre unidades conceituais*
	ATTR:	*Relação atributiva entre unidades conceituais*
	PURP	*Relação de propósito entre unidades conceituais*

Quadro 4: O *template* conceitual.

De certa forma, o *template* proporcionou uma organização prévia à representação formal com base no MultiNet. Mais especificamente, o *template* é composto pelos seguintes campos de informação, denominados respectivamente: "Unidade Conceitual" (UniC), "Glosa", "Tipo-conceitual", "Traços semânticos", "Atributos multidimensionais" e "Relações". Os quatro últimos refletem especificamente os meios representacionais do MultiNet. A seguir, os critérios e fontes utilizados para a especificação de cada campo são descritos.

a) **Especificação do campo UniC**

A identificação da Unidade Conceitual ou UniC é feita por rótulos na língua inglesa e entre os sinais < >. Vale ressaltar, no entanto, que estes são apenas recursos mnemônicos e, por conseguinte, poderiam ser meros códigos como C1 e C2 ou outros. Mais especificamente, os identificadores das UniCs baseiam–se na primeira unidade constitutiva dos *synsets* da WN.Pr. Por exemplo, se o conceito é codificado pelo *synset* {car, auto, automobile, machine, motorcar}, a UniC considerada é <car>.

b) **Especificação do campo Glosa**

Como mencionado, a glosa é uma definição informal dos conceitos implicitamente codificados em cada *synset*. Diz-se "informal" porque, assim como previsto na WN.Pr, sua elaboração não segue necessariamente os padrões definicionais propostos pela Lexicografia.

Para elaboração das glosas (no caso, em PB), partiu-se primeiramente das prórprias glosas (em inglês) armazenadas na WN.Pr. As glosas consideradas bem-formadas foram diretamente traduzidas para o PB. Caso contrário, novas glosas foram elaboradas com base nas definições dos dicionários monolíngues do AmE e na análise das ocorrências em *corpus* das unidades lexicais do *synset* que codifica o conceito que está sendo definido.

Os dois dicionários escolhidos como fonte para essa tarefa foram o *'Cambridge Dictionary of American English'* (LANDAU, 2000) e o *'Longman Dictionary of Contemporary English Online'* (LDOCE-*Online*) (SUMMERS, 2005). A escolha pelo primeiro pautou-se no fato de ele, além de ser uma obra de referência, é específico da variante norte-americana do inglês. A escolha pelo segundo foi motivada principalmente pelo fato de o LDOCE-*Online* ser considerado um dos dois principais dicionários da língua inglesa, o qual engloba as duas variantes, a norte-americana e a britânica.

Além dos dicionários, a elaboração das glosas, ou seja, a proposição de definições informais para os conceitos lexicalizados no AmE, contou também com a observação das ocorrências das unidades lexicais em *corpora* do AmE. Para tal observação, foram utilizados textos em AmE disponíveis na *Web*, considerando, assim, a *Web* como um *corpus* (KILGARIFF, GREFENSTETTE, 2003). A consulta a tais textos foi feita por meio do portal WebCorp[50], que pode ser definido, em linhas gerais, como um conjunto de ferramentas que permite o acesso à *Web* como um *corpus*, ou seja, como uma coleção de textos a partir dos quais fatos sobre a língua podem ser observados e extraídos (MORLEY, 2006). Mais especificamente, o WebCorp gera, a partir de uma unidade lexical e alguns critérios de busca, uma concordância, ou seja, uma lista das ocorrências no *corpus* do item determinado pelo usuário na busca, acompanhado do seu co-texto.

No que diz repeito aos recursos disponíveis no WebCorp, três deles foram usados neste trabalho. O primeiro é o recurso de restringir as buscas ao gênero informativo (subgênero jornalístico); no caso, dentre as opções fornecidas pelo WebCorp, optou-se por restringir as consultas aos textos dos jornais norte-americanos, delimitados pelo próprio portal (*'New York Times'*, *'Washington Post'* e *'SunTimes*'). O segundo recurso foi o que permite visualizar as

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

sentenças completas em que item de busca aparece no *corpus*. O terceiro e último recurso foi o de restringir a visualização das concordâncias para apenas uma por página pesquisada.

c) **Especificação dos campos Tipo conceitual e Traços semânticos**

A classificação completa dos subtipos do tipo [SORT=*co*] em função dos traços não está inteiramente disponível. A Figura 28 (pág. 73) ilustra a parte dessa classificação fornecida pelos autores do MultiNet. Dessa forma, considera-se, neste trabalho, que o conceito <veículos com rodas> é do subtipo **[*mov-art-discrete*]**. Segundo a hierarquia da Figura 28, os conceitos do subtipo **[*mov-art-discrete*]** são especificados pelos traços [ARTIF+], [INSTRU+] e [MOVABLE+].

Assim, para os hipônimos de <veículos com rodas>, os valores **[*mov-art-discrete*]** e [ARTIF+], [INSTRU+] e [MOVABLE+] dos campos "Tipo conceitual" e "Traços semânticos", respectivamente, são fixos, ou seja, todos os hipônimos de <veículos com rodas> têm o mesmo tipo conceitual e, conseqüemente, os mesmos traços semânticos.

d) **Especificação do campo Atributos multidimensionais**

Quanto aos atributos multidimensionais, ressalta-se que os atributos QUANT e CARD não foram considerados na delimitação dos conceitos porque são específicos para a caracterização dos conceitos subjacentes a sintagmas quantificados do tipo <vários alunos>.

Além disso, ressalta-se que os conceitos codificados pelos *synsets* provenientes da WN.Pr, como <carro>, são classificados como conceitos genéricos. De acordo com os pressupostos do MultiNet, tais conceitos são especificados pelos seguintes pares de atributo-valor: [GENER=*ge*], [REFER=*refer*], [VARIA=*con*] e [FACT=*real*].

O valor *ge* do atributo GENER indica exatamente a natureza genérica do conceito. O valor *refer* do atributo REFER indica que esse tipo de conceito não tem referência específica; ele é relacionado a um elemento prototípico não-especificado[51]. O valor *con* do atributo VARIA indica que esse tipo de conceito não varia no nível pré-extensional. Já o valor *real* do atributo FACT indica que esses conceitos genéricos fazem referência a objetos reais.

Por fim, salienta-se que o tipo de extensionalidade dos conceitos genéricos é geralmente [ETYPE=*0*], posto que o extensional (descrição no nível pré-extensional) de um conceito genérico X é um elemento prototípico do conjunto <todos os X>.

[51] O valor do atributo REFER fica, na verdade, subespecificado.

Portanto, todos os conceitos aqui estudados são delimitados em função dos atributos GENER, REFER, VARIA, FACT e ETYPE, os quais são especificados com os seguintes valores: *ge, refer, con, real* e *0,* respectivamente.

e) **Especificação do campo Relações**

A especificação do campo "Relações" foi feita em um primeiro momento com base nas relações fornecidas pela WN.Pr. Como mencionado, a WN.Pr é também uma rede semântica e, por isso, os conceitos, codificados pelos *synsets*, estão organizados em funçao de várias relações léxico-conceituais. Para os conceitos concretos discretos, as relações relevantes da WN.Pr são: hiponímia/ hiperonímia (SUB) e meronímia/ holonímia (PARS). Dessa forma, tais relações puderam ser obtidas diretamente da base da WN.Pr. Durante a delimitação do conceito, no entanto, a relação PURP – também foi especificada com base nas informações da WN.Pr.

6.5. Identificação e compilação das expressões lingüísticas do PB

Antes de os critérios teóricos que fundamentaram a identificação das expressões lingüísticas do PB serem apresentados, salienta-se que as lexicalizações do AmE foram diretamente importadas da WN.Pr; outras possíveis lexicalizações do AmE não foram investigadas, pois tal investigação não fazia parte dos objetivos deste trabalho.

6.5.1. Critérios teóricos

6.5.1.1. A expressão lexical dos objetos

Começa-se esta Subseção com a afirmação de Helbig (2006, p. 17) de que um conceito pode ser intensionalmente caracterizado por três componentes: (i) uma palavra ou grupo de palavras que o designa e representa externamente (isto é, em uma comunicação em língua natural); (ii) uma coleção de relações com outros conceitos; (iii) um padrão complexo de origem perceptual. Nesse sentido, um "conceito lexicalizado" é definido como aquele expresso por uma palavra simples, delimitada, do ponto de vista gráfico, por espaços em branco. No entanto, como bem salienta Helbig (2006), essa definição depende da decisão do que se colocar no léxico, ou seja, do que se considerar como "unidade pertencente ao léxico".

A questão da delimitação da **unidade léxica** ou **lexical** (isto é, unidades armazenadas no léxico) (JACKENDOFF, 2002) tem sido discutida em diferentes perspectivas e a partir de

diferentes pontos de conflito. Muito já se escreveu, aliás, sobre a unidade lexical. Os limites de tais unidades e as suas relações têm suscitado inúmeros estudos sob diferentes perspectivas e uma superabundância de designações que está longe de solucionar o problema da delimitação.

Evidentemente, as discussões sobre a delimitação dos elementos pertencentes ao léxico encontram sua expressão mais viva na Lexicografia, em que as decisões devem ser feitas a cada passo do trabalho.

Neste trabalho, o primeiro critério para a delimitação das unidades que expressam um "conceito lexicalizado" advém da Lexicografia tradicional. Considera-se uma expressão como "lexicalizada" quando esta ocorre como entrada em um dos dicionários monolíngües do PB selecionados para esta pesquisa. Isso quer dizer que, se a expressão *x* consta em pelo menos um dos dicionários monolíngües como entrada, então esta é considerada lexicalizada. Esse critério, aliás, foi adotado no desenvolvimento da *wordnet* para o basco (AGIRRE et al., 2006).

A adoção desse critério tem várias implicações para a definição do que se entende por "unidade lexical".

A primeira delas é a representação das unidades lexicais pelas suas formas canônicas, ou seja, o infinitivo para os verbos, o masculino singular para os adjetivos e nomes e a forma completa para os determinantes e outros. Nessa concepção, as formas pertencentes ao paradigma flexional (p.ex.: formas *embalar, embalou, embalando*) são marcadas como realizações discursivas de um "item lexical" (no caso, *embalar*). Por outro lado, as formas pertencentes ao paradigma derivacional (p.ex.: *embalar, embaladeira, embalado*) são marcadas como itens lexicais distintos (LYONS, 1979). Essa, então, é a concepção de partida deste trabalho. Mas, afinal, quais são as unidades lexicais propriamente ditas?

A segunda implicação diz respeito ao fato de se considerar a palavra como "unidade semântica mínima do discurso", ou seja, como a menor unidade lingüística de que se tem consciência, já que as unidades significantes menores que a palavra não têm significação autônoma (ULLMANN, 1952; BIDERMAN, 1999).

Uma terceira implicação é a classificação das unidades lexicais em: (i) formas livres e (ii) formas dependentes, como os clíticos (ou seja, pronomes pessoais átonos, de uma só sílaba, que não têm acentuação própria e, por isso, dependem do acento da palavra que está imediatamente antes ou depois) e unidades instrumentais ou funcionais. As formas livres no PB são geralmente nomes, adjetivos, advérbios e verbos. As formas dependentes nada mais são do que instrumentos que articulam o discurso, sendo desprovidas de significação externa,

e englobam os atigos, as preposições, os pronomes, as conjunções, etc (BIDERMAN, 1999, 2001).

Assim, pode-se dizer que, sob o ponto de vista mórfico, os conceitos podem se materializar na língua por meio dos seguintes tipos de "unidades lexicais":

(a) unidades simples, como *te, me, com, casa, carro,ncantar, hoje,* etc.;
(b) unidades compostas, como *carro-bomba*, *carro-forte,* etc.;
(c) unidades complexas, que se dividem em "expressões idiomáticas" (do inglês, *idioms*), como *montanha-russa*,

Os conceitos concretos, prototipicamente expressos por unidades da categoria dos nomes, podem ser realizados no PB por meio de unidades dos três tipos.

As unidades simples caracterizam-se por uma seqüência gráfica separada por dois espaços em branco e, por isso, sua identificação não coloca problemas. Nesse grupo, encontra-se a maioria dos clíticos, bem como muitos outros pronomes, a maioria das palavras dependentes como artigos, preposições e conjunções, assim como um grande número de nomes, adjetivos, advérbios e verbos. As compostas são expressões formadas por várias unidades que são separadas por hífen. Por fim, as complexas são formadas por várias unidades separadas por espaços em branco. As expressões idiomáticas são exemplos paradigmáticos desse tipo de unidade.

As **expressões idiomáticas** são relativamente cristalizadas, sendo que seu significado não pode ser construído composicionalmente a partir do significado de suas partes. Além disso, os constituintes de uma expressão idiomática não podem ser substituídos por sinônimos. Esse é o caso, por exemplo, da unidade complexa *montanha-russa* (no inglês, *roller coaster*), cujo significado ("brinquedo encontrado em parques de diversão, constituído por uma armação de trilhos em que uma espécie de trem desliza com rapidez sobre aclives sucessivos e bruscos") não pode ser deduzido do significado de seus elementos constitutivos.

O segundo critério utilizado para a delimitação das unidades lexicais do PB é a noção de **colocação restrita** (do inglês, *restricted collocation*), utilizada no desenvolvimento da WN.Pr. Segundo Bentivogli e Pianta (2004), que se fundamentam em Cruse (1986) e Sag et al. (2002), as unidades complexas do tipo **colocação restrita** são constituídas por palavras que habitualmente co-ocorrem e possuem significado composicional (apesar de apresentar certo grau de coesão semântica). Dessa forma, nota-se que, além de não serem homogêneas, concebem-se as unidades lexicais como entidades que se caracterizam por não serem necessariamente composicionais. As colocações restritas, no entanto, apesar de composicionais, apresentam certo grau de coesão semântica, principalmente devido ao uso e,

por isso, tendem a limitar a substituição de seus componentes por formas (ou expressões) sinônimas (teste da substituição) e a inserção de elementos em seu interior (teste da inserção). Esse é o caso, por exemplo, da expressão do PB *ficha criminal.* No PB, a expressão *ficha penal*, em que se substitui *criminal* por *penal*, e *ficha/ muito/ criminal,* em que se insere o advérbio *muito* entre os constituintes, são pouco aceitáveis, o que dá indícios de que *ficha criminal* é um sintagma lexicalizado (ou colocação restrita).

A noção de "colocação restrita" amplia a categoria das unidades complexas, que passa a englobar as expressões que comumente são classificadas como "locuções" nos dicionários monolíngües do PB. Assim, não somente a existência como entrada, mas também como subentrada, em ao menos um dos dicionários do PB, faz uma expressão *x* ser considerada "lexicalizada".

Dessa forma, conclui-se que a concepção de "unidade lexical" revela um conjunto bastante heterogêneo, pois tais unidades têm caráter e massa muito diferentes, que vão desde um elemento simples até unidades sintagmáticas (BIDERMAN, 1997, 1999; BORBA, 2003). Em outras palavras, pode-se dizer que o conjunto das unidades lexicais não possui um carácter discreto, mas gradual e contínuo.

Baseando-se, então, na premissa de que os conceitos podem ser expressos por unidades dos tipos simples, composto e complexo (expressões idiomáticas e colocações restritas), conclui-se que uma "lacuna lexical", ou seja, caso em que não há unidades lexicais em uma dada língua que expressam um conceito lexicalizado em outra língua, ocorre em duas situações:

(i) quando um conceito, expresso por uma unidade lexical na língua *x*, for desconhecido para os falantes da língua *y*;

(ii) quando um conceito, expresso por uma unidade lexical na língua *x*, for conhecido para os falantes da língua *y*, porém, expresso por meio de uma combinação livre.

Enquanto as unidades simples, compostas e complexas (expressões idiomáticas e colocações restritas) são vistas como lexicalizações, as combinações livres não o são (CRUSE, 1986).

Bentivogli e Pianta (2003, 2004) propõem ainda, no âmbito de desenvolvimento da MultiWordNet[52], que a expressão lingüística de um conceito se estenda um pouco além das fronteiras das unidades lexicais propriamente ditas[53]. Os autores justificam tal proposta tendo em vista que a distinção entre uma unidade complexa e um sintagma livre nem sempre é

[52] Uma proposta alternativa e análoga à EuroWordNet.

[53] Essa proposta também foi empregada no desenvolvimento da base *wordnet* para o basco, a BalkaNet (AGUIRRE et al., 2006).

tarefa fácil. Além disso, ao se aplicar rigorosamente essa distinção, várias expressões comumente utilizadas em uma língua *x* para representar um conceito lexicalizado em uma língua *y* não são consideradas.

Assim, Bentivogli e Pianta propõem que, ao se especificar a expressão lingüística na língua *y* de um conceito lexicalizado na língua *x*, sejam considerados, além das unidades lexicais (simples, compostas e complexas), os chamados **sintagmas livres recorrentes** (SLRs) (do inglês, *recurrent free phrases*) típicos da língua *y*. Um exemplo desse tipo de sintagma é a expressão do PB *carro envenenado,* que expressa o conceito <carro velho modificado para obter melhor desempenho>, lexicalizado no AmE por *hot rod*. Nesse caso, o referido conceito não é lexicalizado no PB, no entanto, há um SLR comumente usado para expressar tal conceito.

Os sintagmas livres recorrentes apresentam as seguintes características principais:

a) os SLRs são combinações livres, ao contrário das colocações restritas;
b) os SLRs são sintagmas, isto é, constituintes sintáticos cujo núcleo é, no caso, um nome; p.ex.: *campo de grãos* e *rolo de papel higiênico* (em PB);
c) os SLRs são relativamente freqüentes;
d) os SLRs são salientes; assim, o lingüista (falante nativo) é capaz de identificar por meio da intuição se determinado SLR tem potencial para capturar um conceito específico; a saliência não necessariamente está relacionada a freqüência.

Outra característica do que se denominou SLR é a concorrência entre o determinante de caráter adjetival e o constituído por uma preposição seguida de substantivo. Por exemplo, *veículo motorizado* e *veículo a motor*. Essa concorrência implica que o sintagma, apesar de freqüente, ainda não se lexicalizou.

Os SLRs são importantes para o tratamento computacional das "lacunas lexicais", uma vez que provêem expressões correspondentes para conceitos não lexicalizados, o que pode contribuir, por exemplo, para a tarefa de tradução automática. Além disso, como salientam Bentivogli e Pianta (2003, 2004), os SLRs também são relevantes para a tarefa de alinhamento lexical de *copora* paralelos.

Dessa forma, optou-se por considerar, além das unidades simples, compostas e complexas, os SLRs na tarefa de identificação das expressões lingüísticas do PB. A identificação dos SLRs é particularmente relevante para o PLN em duas situações: quando (i) expressam conceitos que não são lexicalizados em uma língua *x*; e quando (ii) são sinônimos de uma unidade lexical da língua *x*, fornecendo informação adicional sobre as várias possibilidades de expressão do conceito na língua *x*.

6.5.1.2. A montagem de synsets

As unidades lexicais que expressam um mesmo conceito no PB foram organizadas em *synsets.* Para a montagem dos conjuntos de sinônimos, considerou-se a noção de sinonímia contextual, a mesma adotada na montagem dos *synsets* da WN.Pr. Segundo essa noção, duas expressões são sinônimas em um contexto lingüístico C se a substituição de uma pela outra em C não altera o conteúdo da proposição expressa por C; assim, os lexemas precisam ser comutáveis em apenas um contexto para serem considerados sinônimos. Mais especificamente, a sinonímia é uma relação simétrica, ou seja, se x é "semanticamente similar" a y, então y é "semanticamente similar" a x (CRUSE, 1986).

Além do teste da substituição das formas candidatas à sinonímia em um contexto C, a montagem dos *synsets* formados pelas lexicalizações do PB seguiu o mesmo tipo de teste que guiou a montagem dos *synsets* da WN.Pr cujas unidades constitutivas pertencem à categoria dos nomes. Esse teste é ilustrado no Quadro 5 (FELLBAUM, 1998a,b):

Teste		**Sinonímia entre nomes**
Sim	a	Se algo é x, então algo é y
Sim	b	Se algo é y, então algo é x
Condições		x e y são singulares
Exemplo:	a	Se algo é carro, então algo é automóvel
	b	Se algo é automóvel, então algo é carro
Efeito:		*Synset* {carro, automóvel}

Quadro 5: O teste para a construção de *synsets*.

Ressalta-se, neste ponto, que os SLRs não fazem parte dos *synsets*, uma vez que não são tidos como lexicalizações. Os SLRs formam um conjunto próprio de elementos, denominado *phraset* (do inglês, *phrasal synonym set*) (Bentivogli e Pianta, 2003, 2004).

Quanto à montagem dos *synsets* (e/ou *phrasets*), ressalta-se que foram incluídas marcas de coloquialismo e estrangeirismo. Tais informações foram extraídas dos dicionários monolíngües. As unidades lexicais que refletem o uso informal da língua são marcadas com o rótulo *coloq.*; nos casos de empréstimos de outros sistemas lingüísticos, ou seja, da incorporação ao léxico do PB de itens pertencente a outras línguas, a língua de origem é indicada por meio de uma abreviação, por exemplo, para o inglês (britânico ou norte-americano), utiliza-se *ingl.*

6.5.2. O método de identificação e compilação das expressões lingüísticas e os recursos lexicais

A seleção das unidades lexicais (e dos SLRs) do PB que expressam os conceitos lexicalizados no AmE, oriundos da WN.Pr, foi feita por meio de análise manual de informações contidas em dois tipos de recursos: os estruturados e os não-estruturados.

Outro método possível para tal aquisição seria o semi-automático, que se carateriza de um modo geral pela extração automática (isto é, por meio de uma ferramenta computacional) das unidades lexicais de "dicionários legíveis por máquina"[54] (do inglês, *machine-readable dictionaries,* MRDs) bilíngües ou de *corpora* paralelos alinhados[55] (lexicalmente) com subseqüente revisão manual do conhecimento automaticamente extraído. O que se percebe é que inevitavelmente a extração automática de unidades lexicais de MRDs bilíngües ou de *corpora* paralelos alinhados lexicalmente requer a tarefa de desambiguação da unidade lexical polissêmica do AmE. Essa tarefa é comumente feita de modo manual, pois a desambiguação automática é tarefa cara e problemática.

Assim, diante desse cenário, optou-se pelo método manual de identificação das expressões do PB, posto que este é tido como o mais confiável, apesar de mais demorado (RIGAU, 1998). Essa opção é reforçada pelo fato de não se ter MRDs bilíngües disponíveis para o par de línguas AmE-PB que sejam suficientemente robustos para que se possa utilizar o método semi-automático. O mesmo se aplica aos *corpora* paralelos.

A identificação das unidades lexicais (e SLRs) do PB contou com a utilização de um conjunto de recursos estruturados e não-estruturados, que formam o "*corpus* de referência" deste trabalho.

Os recursos estruturados são: dois dicionários bilíngües (AmE-PB), a própria base da WN.Br, dois dicionários de sinônimos e dois dicionários monolíngües do PB. Os dicionários bilíngües são o *'Dicionário Eletrônico Webster´s Inglês-Português/ Português-Inglês*' (HOUAISS, CARDIM, 1982) e o *'Michaelis: moderno dicionário inglês* (inglês-português/ português-inglês)*'* (WEISZFLOG, 2000). O Webster's, em especial, contém mais de 100.000 entradas e fornece, além do grande número de entradas e expressões, ampla sinonímia, o que é relevante para a identificação das lexicalizações no PB e subseqüente montagem dos *synsets*

[54] Os "dicionários legíveis por máquina" são dicionários em formato digital (cd-rom) que foram coligidos por lexicógrafos e concebidos para uso humano. Comumente, há também versões impressas dos MRDs (RIGAU, 1998).

[55] Um *corpus* paralelo é formado por um conjunto de textos originais e sua respectiva tradução em uma ou várias línguas. Um texto e sua tradução podem ser alinhados; nesse caso, os textos apresentam marcas que indicam as correspondências entre o original e sua tradução. Quando esse alinhamento é feito entre unidades lexicais, diz-se que o *corpus* é alinhado lexicalmente (BERBER SARDINHA, 2004).

ou modificação dos *synsets* já existentes na base da WN.Br. Os dicionários de sinônimos são o *'Grande Dicionário de Sinônimos e Antônimos'* (BARBOSA, 2000) e o *'Dicionário de Sinônimos e Antônimos da Língua Portuguesa'* (FERNANDES, 1997). E, por fim, os monolíngües são o *'Dicionário Aurélio Eletrônico'* (FERREIRA, 1999) e o *'Dicionário Eletrônico Houaiss da Língua Portuguesa'* (HOUAISS, VILLAR, 2001), dois dos mais reconhecidos repositórios lexicais do PB.

O conjunto de recursos não-estuturados é, de certa forma, composto por dois *corpora* textuais. Mais especificamente, utilizam-se o *corpus* PLN-BR FULL do projeto PLN-BR[56] e textos disponíveis da *Web*. O PLN-BR FULL, em especial, é um *corpus* do gênero informativo (e subgênero jornalístico) composto por textos do jornal a Folha de São Paulo, mais especificamente, por textos publicados em apenas um mês de cada ano, no intervalo de 1994 a 2005. No total, o PLN-BR FULL contém aproximadamente 29 milhões de palavras e está disponível para consultas na *webpage* do Philologic[57]. Os textos em PB disponíveis na *Web* foram consultados por meio do motor de busca Google[58], lançando-se mão do recurso de restrição das buscas às páginas do Brasil.

A seguir, descrevem-se os processos de identificação e aquisição das lexicalizações e dos SLRs, assim como os recursos utilizados nesses processos.

6.5.3. Etapas da identificação das expressões lingüísticas do PB

Diante de um conceito lexicalizado do AmE, compilado da WN.Pr, a primeira etapa do processo de identificação das lexicalizações ou SLRs do PB consistia na consulta aos dicionários bilíngües.

Consultava-se o verbete relativo à primeira unidade constitutiva do *synset* da WN.Pr, pois essa expressão é a mais freqüentemente empregada, no AmE, para expressar o conceito subjacente ao *synset* do qual faz parte. Em outras palavras, a primeira unidade do *synset* é a que "melhor" representa o conceito subjacente a ele (FELLBAUM, 1998). Por exemplo, dado o *synset* {car; auto; automobile; machine; motorcar}, consultou-se o verbete de *car* nos dois dicionários bilíngües.

---

[56] O projeto PLN-BR teve por objetivo geral a construção de um espaço interinstitucional de intercâmbio de práticas de investigação lingüístico-computacional para a representação e recuperação de informação semântica e pragmático-discursiva. Os subprojetos do PLN-BR compartilham o mesmo ponto de partida: o tratamento da informação mobilizada em um mesmo *corpus* do PB (o PLN-BR FULL). Mais informações podem ser encontradas no endereço: http://www.nilc.icmc.usp.br/plnbr/.

[57] O Philologic é uma ferramenta Web para buscas, recuperação e análise de *corpora* desenvolvida na Universidade de Chicago (UNIVERSITY OF CHICAGO, 2006).

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

Os dicionários bilíngües forneceram, como formas correspondentes, tanto unidades lexicais (simples, compostas e complexas) quanto SLRs do PB.

Nos casos em que as expressões fornecidas eram do tipo unidade lexical, ou seja, unidades armazenadas como entradas ou subentradas em ao menos um dos dicionários monolíngües do "*corpus* de referência", realizou-se a seqüência de passos descrita a seguir:

(iii) montagem de um *synset* preliminar com todas as unidades lexicais extraídas dos dicionários bilíngües;
(iv) consulta à base da WN.Br para verificar se alguma unidade do *synset* preliminar constava da base e se ela codificava o conceito em questão;
(v) reformulação (se necessária) do *synset* preliminar por meio da inclusão de unidades distintas (das identificadas em (i)) que estavam armazenadas na WN.Br;
(vi) consulta aos verbetes dos dicionários de sinônimos relativos a cada uma das unidades do *synset* preliminar (formulado em (i) ou (iii)) para verificar se estes forneciam unidades distintas das identificadas até o momento;
(vii) reformulação (se necessária) do *synset* preliminar (formulado em (i) ou (iii)) por meio da inclusão de unidades novas extraídas dos dicionários de sinônimos;
(viii) consulta aos verbetes dos dicionários monolíngües relativos a cada uma das unidades do *synset* preliminar (formulado em (i), (iii) ou (v)) para verificar quais unidades realmente expressam o conceito em questão;
(ix) reformulação (se necessária) do *synset* preliminar (formulado em (i), (iii) ou (v)) diante da análise dos verbetes dos dicionários monolíngües;
(x) consulta aos *corpora* para verificar a freqüência de uso das unidades do *synset* preliminar (formulado em (i), (iii) ou (v)), pois a freqüência revela a ocorrência observada de tais unidades;
(xi) montagem, com base na noção de sinonímia contextual, do *synset* final com as unidades "validadas" em (viii).

Os passos descritos de (i) a (xi) e os recursos utilizados estão ilustrados na Figura 40.

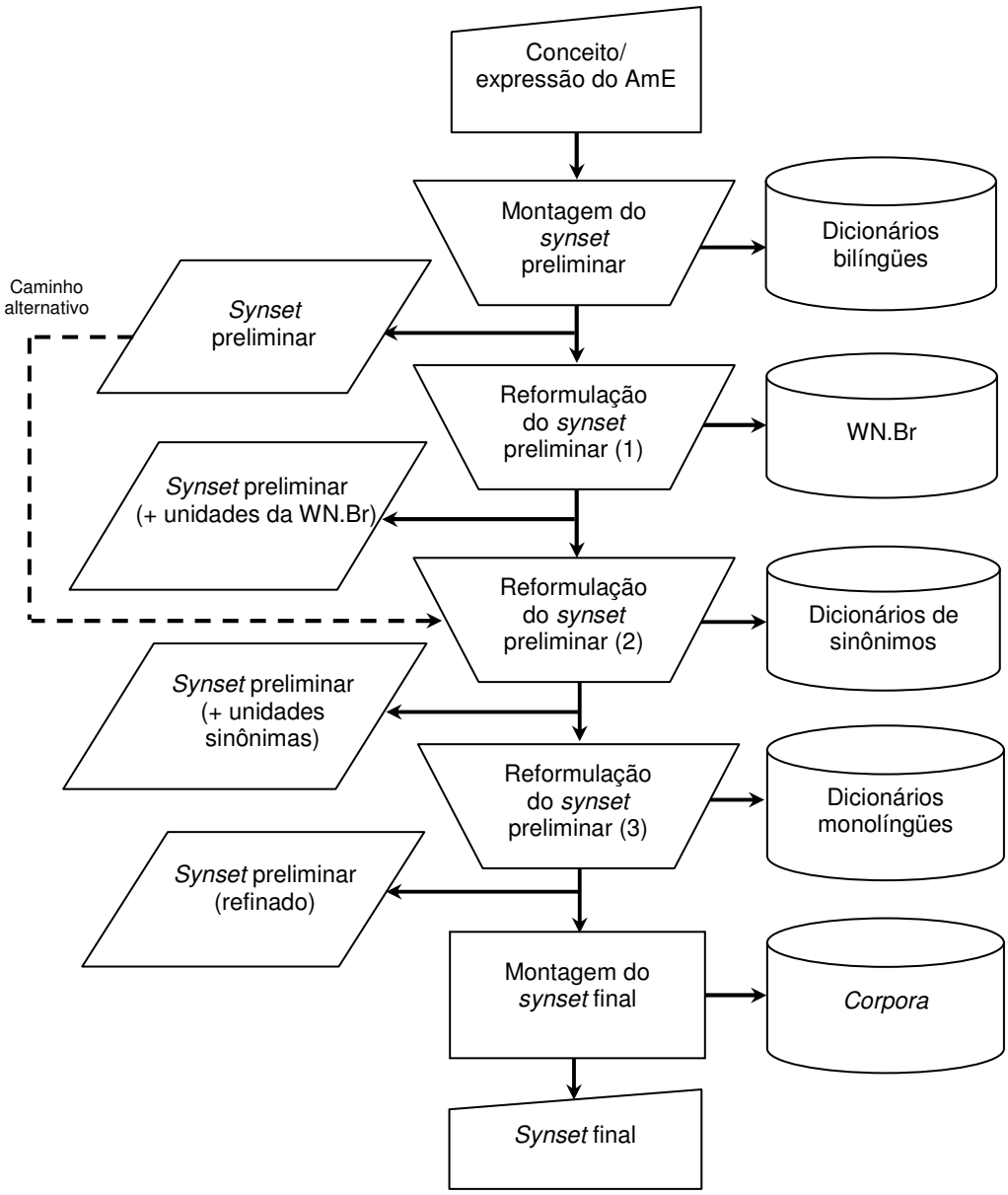


Figura 40: Esquema do processo de identificação e compilação das unidades lexicais do PB.

Vale ressaltar que a análise das ocorrências das unidades lexicais constitutivas dos *synsets* preliminares nos *corpora* teve como principal objetivo verificar a adequação sincrônica das unidades (simples, compostas e complexas) extraídas das obras lexicográficas, posto que tais obras tendem a conter unidades lexicais em desuso. Assim, ao se verificar nos *corpora* que determinada unidade (constitutiva de um *synset* preliminar) é usualmente empregada para expressar determinado conceito, esta passa a ser considerada uma lexicalização do PB propriamente dita. Caso contrário, a unidade é desconsiderada.

Na Figura 40, a linha tracejada indica o processo de identificação das unidades lexicais do PB quando a base da WN.Br não armazena nenhuma das unidades constituintes do *synset*

preliminar. Nesse caso, um *synset* preliminar pode sofrer a primeira reformulação quando os dicionários de sinônimos são consultados.

Nos casos em que as expressões extraídas dos dicionários bilíngües eram do tipo SLR, realizavam-se os passos:

(a) montagem de um conjunto (ou *phraset*) preliminar com os SLRs extraídos dos dicionários;
(b) consulta aos *corpora* para verificar se os SLRs do conjunto preliminar são empregrados para expressar o conceito em questão.

Ressalta-se que os SLRs foram identificados em duas situações distintas. Em uma delas, o SLR era diretamente identificado e extraído dos dicionários bilíngües. Esse foi o caso, por exemplo, de *vagão de mercadorias,* como fica claro no verbete do dicionário *'Michaelis: moderno dicionário inglês*', descrito a seguir.

freight car
n Amer vagão de mercadorias.

Em outra situação, o SLR era fornecido indiretamente. Mais especificamente, o SLR fazia parte da definição dos dicionários bilíngües. Nesse caso, diz-se que os verbetes dos dicionários "escondiam" possíveis SLRs. Essa característica dos dicionários bilíngües já havia sido salientada por Bentivogli e Pianta (2002). Esse foi o caso, por exemplo, da expressão *bibicleta motorizada*, como fica claro no verbete do dicionário *'Michaelis: moderno dicionário inglês'*, descrito na seqüência.

moped
n bicicleta motorizada pequena, motocicleta.

6.5.4. Proposição e preenchimento do *template lexical*

Tanto as lexicalizações do AmE quanto as lexicalizações e SLRs do PB foram inseridas em um *template* lexical, sendo que, para cada *template* conceitual (ou melhor, para cada conceito lexicalizado no AmE), preencheu-se um *template* lexical corrrespondente (Quadro 6). O *template* lexical é formado por dois campos denominados "Expressões lingüísticas" e "Frases-exemplo".

Expressões lingüísticas:	**SynAmE:**	*synset da WN.Pr*
	SynPB: ***Phraset*PB:**	*unidades lexicais ou* *SLRs do PB*
Frases-exemplo:	**AmE:**	*Frases-exemplo extraídas da WN.Pr ou de corpora; para cada unidade lexical do AmE, há uma frase-exemplo correspondente*
	PB:	*Frases-exemplo extraídas de corpora; para cada unidade lexical ou SLRs do PB, há uma frase-exemplo correspondente*

Quadro 6: O *template* lexical.

a) **Especificação do campo Expressões lingüísticas**

O campo Expressões lingüísticas é dividido em SynAmE e SynPB/ *Phraset*PB. No subcampo SynAmE, são inseridas as unidades lexicais provenientes da WN.Pr (mais especificamente, os *synsets*). No subcampo SynPB/ *Phraset*PB, são especificadas: (i) as unidades lexicais do PB, em formato de *synset*, e (ii) as SLRs do PB, em formato de *phraset* (ou seja, um conjunto de SLRs).

b) **Especificação do campo Frases-exemplo**

Nesse campo, especifica-se pelo menos uma frase-exemplo, que fornece o contexto de uso mínimo, para cada lexicalização do AmE e do PB. As frases-exemplo para as unidades do AmE foram extraídas da WN.Pr ou de *corpora*. Para as unidades lexicais (e SLRs) do PB, em especial, as frases-exemplo foram extraídas dos dicionários monolíngües e dos *corpora* textuais que fazem parte do "*corpus* de referência".

Cada uma dessas fontes possui uma sigla identificadora. No caso, as siglas são:

(a) A, para o *'Dicionário Aurélio Eletrônico*';

(b) H, para o *'Dicionário Eletrônico Houaiss da Língua Portuguesa*';

(c) F, para o *corpus* PLN-BR FULL;

(d) I, para as sentenças provenientes da *internet* (ou *Web*).

A sigla é introduzida em uma frase-exemplo como um rótulo associado à unidade lexical (em questão) para indicar a procedência da frase. Para a especificação das frases-exemplo das unidades do AmE, estabeleceu-se, além da sigla I, para as frases provenientes da *Web*, a sigla WN, para aquelas provenientes da WN.Pr.

Após a especificação dos *templates* conceitual e lexical, para cada conceito lexicalizado no AmE, obteve-se, ao final, um modelo de *template* léxico-conceitual, como o ilustrado no

Quadro 7. Esse *template* final nada mais é do que a união do *templates* conceitual e lexical. Assim, todos os conceitos compilados da WN.Pr possuem um *template* léxico-conceitual correspondente.

UniC:	*Identificador da unidade conceitual; símbolo mnemônico*	
Glosa	*Definição lexicográfica informal de uma unidade conceitual*	
Tipo conceitual:	*Tipo conceitual* (*MultiNet*)	
Traços semânticos:	*Traços semânticos*	
Atributos multidimensionais	**REFER:**	*Determinação da referência de uma unidade conceitual*
	ETYPE:	*Caracterização do tipo de extensionalidade de uma unidade conceitual*
	VARIA:	*Caracterização da variabilidade de uma unidade conceitual*
	GENER:	*Caracterização da generalidade de uma unidade conceitual*
	FACT:	*Caracterização da factividade de uma unidade conceitual*
Relações:	**SUB:**	*Relação de subordinação entre unidades conceituais*
	PARS:	*Relação de parte-todo entre unidades conceituais*
	ORIGM:	*Relação de material de origem entre unidades conceituais*
	PROP:	*Relação de propriedade entre unidades conceituais*
	ATTR:	*Relação atributiva entre unidades conceituais*
	PURP	*Relação de propósito entre unidades conceituais*
Expressões lingüísticas:	**SynAmE:**	*synset da WN.Pr*
	SynPB: ***Phraset*PB:**	*unidade lexicais ou* *SLRs do PB*
Frases-exemplo:	**AmE:**	*Frases-exemplo extraídas da WN.Pr ou de corpora; para cada unidade lexical do AmE, há uma frase-exemplo correspondente*
	PB:	*Frases-exemplo extraídas de corpora; para cada unidade lexical ou SLRs do PB, há uma frase-exemplo correspondente*

Quadro 7: O *template* léxico-conceitual.

A seguir, o preenchimento do *template* léxico-conceitual para certos conceitos é exemplificado.

6.6. Exemplo de preenchimento dos *templates* léxico-conceituais

Com base nos procedimentos propostos para a delimitação dos conceitos e identificação das unidades lexicais (e SLRs) do PB, exemplifica-se o preenchimento do *template* léxico-conceitual para os conceitos <motor vehicle>, <car> e <touring car>.

a) **O preenchimento do *template* léxico-conceitual para <motor vehicle>**

UniC:	<motor vehicle>	
Glosa	"veículo com rodas autopropulsado que é dotado de motor e não anda sobre trilhos" (WN.Pr)	
Tipo conceitual:	[mov-art-discrete]	
Traços semânticos:	[ARTIF+] [INSTRU+] [MOVABLE+]	
Atributos multidimensionais	**REFER:**	refer
	ETYPE:	0
	VARIA:	con
	GENER:	ge
	FACT:	real
Relações:	**SUB:**	{amphibian1, amphibious vehicle}, {bloodmobile}, {car, auto, automobile, machine1, motorcar}, {doodlebug1}, {four-wheel drive1, 4WD1}, {go-kart}, {golfcart, golf cart}, {hearse}, {motorcycle, bike2}, {snowplow, snowplough}, {truck, motortruck}
	PARS:	HERDADAS DE <wheeled vehicle> {axle}, {brake}, {splasher}, {wheel} PRÓPRIAS {airbrake}, {bodywork}, {brake system, the brakes}, {cab2}, {car wheel}, {chassis}, {cooling system1, engine cooling system}, {drive line, drive line system}, {electrical system}, {fuel system}, {gearshift, gearstick, shifter, gear lever}, {hand brake, emergency, emergency brake, parking brake}, {internal-combustion engine, ICE1}, {odometer, hodometer, mileometer, milometer}, {pedal, treadle, foot pedal, foot lever}, {power brake}, {second gear, second}, {speedometer, speed indicator}, {suspension, suspension system}, {windshield, windscreen}, {windshield wiper, windscreen wiper, wiper, wiper blade}
	ORIGM:	<Nil>
	PROP:	<Nil>
	ATTR:	<Nil>
	PURP	HERDADA DE <wheeled vehicle>

		{travel; go; move; locomote}
Expressões lingüísticas:	**SynAmE:**	{motor vehicle; automotive vehicle}
	SynPB:	{GAP}
	***Phraset*PB:**	{veículo motorizado, veículo a motor}
Frases-exemplo:	**AmE:**	The Department of Motor Vehicles is immediately suspending the practice that allowed an individual to register a I] motor vehicle in someone else's name I] Automotive vehicle fires can result from post-collision events as well as accidental and intentional means.
	PB:	Não use fones de ouvido quando estiver dirigindo, andando de bicicleta ou operando qualquer I] veículo motorizado. Para conduzir um I] veículo a motor na via pública, é necessário estar legalmente habilitado para o efeito.

Quadro 8: O *template* léxico-conceitual do conceito <motor vehicle>.

Observa-se que, para o *template* léxico-conceitual do conceito <motor vehicle>, descrito no Quadro 8:

- a UniC relativa ao conceito em questão, <motor vehicle>, é a mais representativa do *synset* da WN.Pr;
- a glosa fornecida pela WN.Pr ("self-propelled wheeled vehicle that does not run on rails") foi traduzida para o PB: "veículos com rodas autopropulsado que é dotado de motor e não anda sobre trilhos";
- o tipo conceitual, os traços semânticos e os atributos multidimensionais (e seus valores) são os mesmos do conceito mais geral <wheeled vehicle>;
- o tipo conceitual do conceito é, segundo o MultiNet, assim especificado: [*mov-art-discrete*];
- as relações do tipo SUB foram extraídas da WN.Pr, em um total de 11 conceitos (*synsets*) diretamente subordinados a <motor vehicle>;
- as relações PARS subdividem-se em herdadas e próprias; <motor vehicle> herda as 4 relações PARS de <wheeled vehicle> e possui 21 relações PARS próprias (ou diretas), as quais foram extraídas da WN.Pr;
- a relação PURP foi herdada do conceito mais geral <wheeled vehicle>;
- as relações ORIGM, PROP e ATTR não foram especificadas para o referido conceito; seus respectivos campos foram assinalados com o rótulo <Nil> (zero ou nada);

- as expressões lingüísticas identificadas no PB foram: *veículo a motor,* identificada diretamente em um dos dicionários bilíngües, e *veículo motorizado*, elaborada com base na transformação de *a motor* em *motorizado*.
- as expressões do PB foram validadas pela análise de *corpora*;
- as expressões *veículo motorizado* e *veículo a motor* foram classificadas como SLRs do PB, pois *a motor* e *motorizado* são sinônimos (mais especificamente, *a motor* é a forma preposicionada de *motorizado*) (teste da substituição); dessa forma, o campo relativo à lexicalização do PB foi assinalado com o rótulo {GAP};
- as expressões *veículo motorizado* e *veículo a motor,* por serem SLRs, constituem o *phraset* do PB {veículo motorizado, veículo a motor};
- as frases-exemplo para as unidades do AmE, "The Department of Motor Vehicles is immediately suspending the practice that allowed an individual to register a I] motor vehicle in someone else's name" e "I] Automotive vehicle fires can result from post-collision events as well as accidental and intentional means" foram coletada da *Web*, posto que a WN.Pr não armazena essa informação;
- as frases-exemplo para as expressões do PB foram coletadas da *Web*, são elas: "Não use fones de ouvido quando estiver dirigindo, andando de bicicleta ou operando qualquer I] veículo motorizado" e "Para conduzir um I] veículo a motor na via pública, é necessário estar legalmente habilitado para o efeito".

d) **O preenchimento do *template* léxico-conceitual para <car>**

UniC:	<car>	
Glosa	"veículo com quatro rodas e um motor, que se usa para viajar de um lugar para o outro" (WN.Pr)	
Tipo conceitual:	[mov-art-discrete]	
Traços semânticos:	[ARTIF+] [INSTRU+] [MOVABLE+]	
Atributos multidimensionais	**REFER:**	refer
	ETYPE:	0
	VARIA:	con
	GENER:	ge
	FACT:	real
Relações:	**SUB:**	{ambulance}; {beach wagon, station wagon, wagon, beach waggon, station waggon, wagon}; {bus, jalopy, heap}; {cab, hack, taxi, taxicab}; {compact, compact car}; {convertible}; {coupe}; {cruiser, police cruiser, patrol car, police car, prowl car, squad car}; {electric,

		electric automobile, electric car}; {gas guzzler}; {hardtop}; {hatchback}; {horseless carriage}; {hot rod, hot-rod}; {jeep, landrover}; {limousine, limo}; {loaner}; {minicar}; {Model T}; {pace car}; {racer, race car, racing car}; { roadster, runabout, two-seater}; {sedan}; {sports car, sport car}; {sport utility, sport utility vehicle, S.U.V., SUV}; {Stanley Steamer}; {stock car}; {subcompact, subcompact car}; {touring car, phaeton, tourer}; {used-car, secondhand car}
	PARS:	HERDADAS DE <wheeled vehicle> {axle}, {brake}, {splasher}, {wheel} HERDADAS DE <motor vehicle> {airbrake}, {bodywork}, {brake system, the brakes}, {cab2}, {car wheel}, {chassis}, {cooling system1, engine cooling system}, {drive line, drive line system}, {electrical system}, {fuel system}, {gearshift, gearstick, shifter, gear lever}, {hand brake, emergency, emergency brake, parking brake}, {internal-combustion engine, ICE1}, {odometer, hodometer, mileometer, milometer}, {pedal, treadle, foot pedal, foot lever}, {power brake}, {second gear, second}, {speedometer, speed indicator}, {suspension, suspension system}, {windshield, windscreen}, {windshield wiper, windscreen wiper, wiper, wiper blade} PRÓPRIAS {acelerator, accelerator pedal, gas pedal, gas, throttle, gun}; {air bag}; {auto accessory}; {automobile engine}; {automobile horn, car horn, motor horn, horn, hooter}; {buffer, fender}; { bumper}; {car door}; {car mirror}; {car seat}; {car window}; {fender, wing}; {first gear, first, low gear, low}; {floorboard}; {gasoline engine}; {glove compartment}; {grille, radiator grille}; {high gear, high}; {hood, bonnet, cowl, cowling}; {luggage compartment, automobile trunk, trunk}; {reverseroof}; {running board}; {stabilizer bar, anti-sway bar}; {sunroof, sunshine-roof }; {tail fin, tailfin, fin}; {third gear, third}
	ORIGM:	<Nil>
	PROP:	<Nil>
	ATTR:	<Nil>
	PURP	HERDADA DE <wheeled vehicle> {travel, go, move, locomote}
Expressões lingüísticas:	**SynAmE:**	{car; auto; automobile; machine; motorcar}
	SynPB:	{auto; automóvel; carro}[59]
	***Phraset*PB:**	<Nil>

[59] Os elementos constitutivos dos *synsets* do PB estão organizados em ordem alfabética e separados por ponto-e-vígula.

Frases-exemplo:	**AmE:**	He needs a WN] car to go to work. With the idea of halting the I]auto, he pushed down on the foot brake. The I]automobile was stopped at the Fifth Avenue. The I]machine was a BMW. Any I]motorcar left unattended on the property, overnight or otherwise, is the sole responsibility of its owner.
	PB:	É como o F]carro, que recebe o combustível (gasolina ou álcool) e, em seguida, queima-o, para obter a energia de que precisa para andar. O automóvel se deteve e o animal grunhia, e os grunhidos estremeciam o I]auto. Quando um F]automóvel em movimento freia bruscamente, seus ocupantes tendem a se chocar contra o banco ou o vidro da frente.

Quadro 9: O *template* léxico-conceitual do conceito <car>.

Observa-se que, para o *template* léxico-conceitual do conceito <car> (Quadro 9):

- a glosa da WN.Pr ("a motor vehicle with four wheels; usually propelled by an internal combustion engine") foi substituída pela definição fornecida pelo LDOCE-*Online* ("a vehicle with four wheels and an enginee, that you use to travel from one place to another"), que, considerada adequada, foi traduzida para o PB: "um veículo com quatro rodas e um motor, que se usa para locomover de um lugar para outro";
- as relações do tipo SUB foram extraídas da WN.Pr, em um total de 30 conceitos (*synsets*) diretamente subordinados a <car>;
- as relações PARS subdividem-se em herdadas e próprias; <car> herda as 21 relações PARS de <motor vehicle>, além das 4 relações PARS de <wheeled vehicle>, e possui 28 relações PARS próprias (ou diretas);
- as expressões lingüísticas do referido conceito identificadas no PB foram: *carro, auto* e *automóvel;* ressalta-se que a identificação e aquisição dessas expressões seguiu o processo ilustrado na Figura 40;
- as expressões *carro, auto* e *automóvel* foram classificadas como unidades lexicais (no caso, simples) do PB e, por isso, o campo relativo à lexicalização do PB foi preenchido com o *synset* equivalente {auto; automóvel; carro }; salienta-se que a WN.Br contém o *synset* {carro, automóvel, auto, veículo}, o qual, diante da análise realizada neste trabalho, pode passar a ser {auto; automóvel; carro};

- as frases-exemplo para as unidades do AmE foram coletadas da WN.Pr e da *Web*; são elas: “He needs a WN] car to go to work”, “With the idea of halting the I]auto, he pushed down on the foot brake”, “The I]automobile was stopped at the Fifth Avenue”, “The I]machine was a BMW” e “Any I]motorcar left unattended on the property, overnight or otherwise, is the sole responsibility of its owner”;
- as frases-exemplo para as expressões do PB foram coletadas do *corpus* PLN-BR FULL e da *Web*; são elas: “É como o F]carro, que recebe o combustível (gasolina ou álcool) e, em seguida, queima-o, para obter a energia de que precisa para andar”; “O automóvel se deteve e o animal grunhia, e os grunhidos estremeciam o I]auto” e “Quando um F]automóvel em movimento freia bruscamente, seus ocupantes tendem a chocar-se contra o banco ou o vidro da frente”.

e) **O preenchimento do *template* léxico-conceitual para <touring car>**

UniC:	<touring car>	
Glosa	“carro grande e aberto para quatro pessoas com teto conversível” (WN.PR)	
Tipo conceitual:	[mov-art-discrete]	
Traços semânticos:	[ARTIF+] [INSTRU+] [MOVABLE+]	
Atributos multidimensionais:	**REFER:**	refer
	ETYPE:	0
	VARIA:	con
	GENER:	ge
	FACT:	real
Relações:	**SUB:**	<Nil>
	PARS:	HERDADAS DE <wheeled vehicle> {axle}, {brake}, {splasher}, {wheel} HERDADAS DE <motor vehicle> {airbrake}, {bodywork}, {brake system, the brakes}, {cab2}, {car wheel}, {chassis}, {cooling system1, engine cooling system}, {drive line, drive line system}, {electrical system}, {fuel system}, {gearshift, gearstick, shifter, gear lever}, {hand brake, emergency, emergency brake, parking brake}, {internal-combustion engine, ICE1}, {odometer, hodometer, mileometer, milometer}, {pedal, treadle, foot pedal, foot lever}, {power brake}, {second gear, second}, {speedometer, speed indicator}, {suspension, suspension system}, {windshield, windscreen}, {windshield wiper, windscreen wiper, wiper, wiper blade}

		HERDADAS DE <car> {acelerator, accelerator pedal, gas pedal, gas, throttle, gun}; {air bag}; {auto accessory}; {automobile engine}; {automobile horn, car horn, motor horn, horn, hooter}; {buffer, fender}; { bumper}; {car door}; {car mirror}; {car seat}; {car window}; {fender, wing}; {first gear, first, low gear, low}; {floorboard}; {gasoline engine}; {glove compartment}; {grille, radiator grille}; {high gear, high}; {hood, bonnet, cowl, cowling}; {luggage compartment, automobile trunk, trunk}; {reverseroof}; {running board}; {stabilizer bar, anti-sway bar}; {sunroof, sunshine-roof}; {tail fin, tailfin, fin}; {third gear, third}; {window}
	ORIGM:	<Nil>
	PROP:	<Nil>
	ATTR:	<Nil>
	PURP	HERDADA DE <wheeled vehicle> {travel; go; move; locomote}
Lexicalizações:	**SynAmE:**	{touring car; phaeton; tourer}
	SynPB: ***Phraset*PB:**	{GAP} {carro aberto}
Frases-exemplo:	**AmE:**	A I]touring car with a stock motor and correct gearing will hit 60-65km/h. The car is a I]phaeton, one of two convertible models available in 1937. A I]tourer with relatively flat sides.
	PB:	Parece que podemos comparar a visão de mundo de Faulkner à de um homem sentado num F]carro aberto e que olha para trás.

Quadro 10: O *template* léxico-conceitual do conceito <touring car>.

Observa-se que, para o *template* léxico-conceitual do conceito <touring car>, descrito no Quadro 10:

- o conceito em questão não se relaciona com outros conceitos por meio da relação SUB, ou seja, ele não tem hipônimos;
- as relações PARS foram herdadas; mais especificamente, <touring car> herda as 4 relações PARS de <wheeled vehicle>, as 21 relações PARS de <motor vehicle> e as 28 relações de <car>;
- a expressão lingüística identificada no PB foi: *carro aberto,* identificada em um dos dicionários bilíngües;
- a expressão *carro aberto* foi classificada como SLR; dessa forma, o campo relativo à lexicalização do PB foi assinalado com o rótulo {GAP};

6.7. Os conceitos e as expressões lingüísticas do PB (unidades lexicais e SLRs)

Nesta Subseção, mais especificamente no Quadro 11, apresenta-se o conjunto de todos os conceitos lexicalizados que pertencem ao domínio dos "veículos com rodas" e suas respectivas lexicalizações (unidades lexicais que os representam) no AmE, ambos extraídos da WN.Pr, juntamente com as lexicalizações e SLRs identificados no PB. Nesse Quadro, os conceitos estão descritos por meio da UniC e glosa correspondentes e as lexicalizações ou unidades lexicais, tanto do AmE quanto do PB, estão descritas por meio de *synsets*. Os SLRs do PB estão apresentados por meio de *phrasets*. Os conceitos não-lexicalizados no AmE estão destacados em letras maiúsculas para evidenciar os casos em que a especificação de um *phraset* do PB foi possível, apesar de este não ser o foco do trabalho. Ainda nesta Subseção, são apresentados os dados estatísticos sobre a identificação das expressões lingüísticas dos referidos conceitos no PB.

Lexicalizações do AmE	**Conceitos** (UniC + glosa)	**Lexicalizações do PB** (e/ou *Phrasets*)
{ambulance}	<ambulance> "carro usado para transportar pessoas doentes ou feridas de e para hospitais"	*Synset*={ambulância}
{amphibian; amphibious vehicle}	<amphibian> "veículo motorizado que pode se locomover por terra ou aguá"	{GAP} *Phraset*={veículo anfíbio}
{angledozer}	<angledozer> "buldôzer equipado de lâmina angular ajustável (na altura e na inclinação), usado para nivelar ou limpar terrenos, abrir fossos, nivelar estradas e realizar pequenas escavações"	*Synset*={*ingl.* angledozer; aplanadora; niveladora; patrol; patrola}
{applecart}	<applecart> "carrinho de mão em que maçãs e outras frutas são vendidas na rua"	{GAP}
{armoured car}	<armoured car> "veículo blindado para uso militar, equipado com armas leves"	*Synset*={blindado}
{armoured car}	<armoured car> "veículo blindado com portas e fechaduras resistentes usado no transporte de grandes somas de dinheiro e outros valores"	*Synset*={carro-forte}
{armoured personnel carrier}	<ARMOURED PESONNEL CARRIER> "veículo com rodas blindado usualmente destinado ao transporte de infantaria"	
{armoured vehicle}	<ARMOURED VEHICLE> "veículo com rodas revestido por blindagem"	*Phraset*={veículo blindado}

{assault gun}	<assault gun> "veículo com rodas blindado, constituído por um chassi de tanque (sem torre), equipado com um canhão grande, usado como arma antitanque para dar suporte as tropas"	*Synset*={carro de assalto}
{baby buggy; baby carriage; perambulator; pram; stroller; go-cart; pushchair; pusher}	<baby buggy> "veículo pequeno com quatro rodas em que são transportados bebês ou crianças"	{GAP} *Phraset* ={carrinho de bebê}
{baggage car; luggage van}	<baggage car> "vagão em que são transportadas as bagagens de passageiros"	{GAP} *Phraset*={vagão bagageiro}
{bandwagon}	<bandwagon> "veículo com rodas que precisa ser rebocado ou puxado, usado para transportar bandas de música"	{GAP}
{barouche}	<barouche> "carruagem com dois assentos, puxada por dois cavalos em parelha"	*Synset*={caleche}
{barrow; garden cart; lawn cart; wheelbarrow}	<barrow> "carrinho de mão de uma ou mais rodas dianteiras, com dois varais na parte oposta, usado para transportar pequenas cargas"	{GAP}
{bassinet}	<bassinet> "veículo pequeno com roda que serve para transportar bebês; e que se assemelha a um berço"	{GAP} *Phraset* ={carrinho berço}
{beach wagon; station wagon; wagon}	<beach wagon> "carro que tem corpo longo e portas traseiras; com espaço atrás dos assentos traseiros"	*Synset*={perua; *ingl.* wagon}
{berlin}	<berlin> "limusine com uma divisão de vidro entre os assentos dianteiros e os traseiros"	{GAP}
{bicycle; bike; wheel; cycle}	<bicycle> "veículo com rodas constituído por um quadro, duas rodas alinhadas uma atrás da outra, um selim e manobrado por guidom e pedais"	*Synset*={bicicleta; *ingl.* bike*; coloq.* magrela}
{bicycle-built-for-two; tandem bicycle; tandem}	<bicycle-built-for-two> "bicicleta com dois pares de pedais e dois assentos"	{GAP}
{bloodmobile}	<bloodmobile> "veículo motorizado equipado para coletar doações de sangue"	{GAP}
{boneshaker}	<boneshaker> "veículo com rodas em má condições e desconfortável"	{GAP} *Phraset*={veículo sacolejante}
{bookmobile}	<bookmobile> "caminhão-baú carregado de livros; usado como livraria ou biblioteca móvel"	{GAP} *Phraset*={biblioteca ambulante}

{boxcar}	<boxcar> "vagão de carga fechado, com teto e portas corrediças nas laterais"	{GAP}
{brougham}	<brougham> "carruagem pequena puxada por um cavalo"	*Synset*={berlinda}
{brougham}	<brougham> "sedã que não tem teto sobre os assentos dianteiros"	{GAP}
{buckboard}	<buckboard> "carruagem aberta sem molas"	{GAP}
{buggy; roadster}	<buggy> "carruagem pequena e leve puxada por apenas um cavalo"	*Synset*={charrete}
{bulldozer; dozer}	<bulldozer> "trator grande e pesado equipado com uma lâmina frontal de aço reforçada e perpendicular ao chão, usado. para escavar e empuxar terra e qualquer outro material"	*Synset*={buldôzer; trator de lâmina} *Phraset*={máquina de terraplanagem}
{bus; jalopy}	<bus> "carro velho e que não aparenta segurança"	*Synset*={calhambeque; lata}
{cab; hack; taxi; taxicab}	<cab> "carro provido de taxímetro, dirigido por uma pessoa cujo trabalho é transportar passageiros"	*Synset*={carro de praça; táxi}
{cabin car; caboose}	<cabin car> "vagão para operários e pessoal de trem de carga; usualmente é o último carro de um trem"	{GAP}
{cabriolet; cab}	<cabriolet> "carruagem leve, com duas rodas, com capota móvel e puxada por um cavalo"	*Synset*={cabriolé}
{camper trailer}	<camper trailer> "trailer especialmente equipado para ser usado em campings ou excursões turísticas"	{GAP}
{camper; camping bus; motor home}	<camper> "veículo recreativo, especialmente adaptado para servir de moradia durante viagens"	{GAP}
{car; auto; automobile; machine; motorcar}	<car> "veículo com quatro rodas e um motor, destinado ao transporte de passageiros ou cargas"	*Synset*={auto; automóvel; carro}
{caroche}	<caroche> "carruagem de luxo (Séc. XVII)"	{GAP}
{carriage; equipage}	<carriage> "veículo com quatro rodas puxado por cavalos, dotado de suspensão de molas, usado para transportar pessoas"	*Synset*={carruagem; coche; sege}
{carrier}	<carrier> "veículo com rodas que se locomove com seus próprios meios de propulsão destinado especificamente ao transporte de diversas coisas"	{GAP}

{cart}	<cart> "veículo com duas rodas puxado normalmente por cavalo, burro ou boi, para transporte de cargas"	*Synset*={carroça}
{Caterpillar; cat}	<Caterpillar> "veículo autopropulsionado que se locomove por meio de duas esteiras de aço rolantes ou lagartas; usado principalmente para mover terra em constuções ou fazendas"	*Synset*={caterpílar}
{cattle car}	<cattle car> "vagão de carga fechado usado para transportar gado"	{GAP} *Phraset*={vagão gaiola; vagão de gado}
{chaise}	<chaise> "carruagem leve, de duas ou quatro rodas, com capota móvel e puxada por um cavalo"	{GAP}
{chariot}	<chariot> "carruagem leve com quatro rodas usada em cerimoniais"	*Synset*={carruagem}
{chariot}	<chariot> "veículo com duas rodas puxado por cavalos, usado em guerras e corridas no antigo Egito, Grécia e Roma"	*Synset*={biga}
{chuck wagon}	<chuck wagon> "carroça para transportar provisões e utensílios de cozinha para vaqueiros"	{GAP}
{clarence}	<clarence> "carruagem fechada de quatro lugares"	{GAP}
{club car; lounge car}	<club car> "vagão em trem de passageiros onde servem-se bebidas"	{GAP}
{coach; four-in-hand; coach-and-four}	<coach> "carruagem puxada por quatro cavalos e conduzida por apenas uma pessoa"	{GAP}
{coal car}	<coal car> "vagão de carga com teto e laterais fixas usado para transportar carvão"	{GAP} *Phraset*={vagão carvoeiro}
{coaster wagon}	<coaster wagon> "veículo com rodas que precisa ser rebocado ou puxado, usualmante destinado às crianças"	{GAP}
{compact; compact car}	<compact> "carro pequeno e econômico"	{GAP} *Phraset*={compacto; carro compacto}
{convertible}	<convertible> "carro que tem o teto ou cobertura móvel (capota), ou seja, dobrável ou removível"	{GAP} *Phraset*={conversível; carro conversível}
{coupe}	<coupe> "carro com duas portas, dois assentos dianteiros e um compartimento para bagagem"	*Synset*={cupê}

{covered wagon; Conestoga wagon; Conestoga; prairie wagon; prairie schooner}	<covered wagon> "carroça grande, com coberta de proteção, usada para transporte de pessoas"	*Synset*={carroção}
{cruiser; police cruiser; patrol car; police car; prowl car; squad car}	<cruiser> "carro utilizado pela polícia para vigiar as ruas; é equipado com comunicação via rádio"	*Synset*={radiopatrulha} *Phraset*={carro de polícia}
{delivery truck; delivery van; panel truck}	<delivery truck> "caminhão-baú usado no transporte de pequenas cargas"	{GAP} *Phraset*={caminhão de entregas}
{diesel locomotive}	<DIESEL LOCOMOTIVE> "locomotiva movida por um motor a diesel"	*Phraset*={locomotiva a diesel}
{diesel-electric locomotive}	<DIESEL-ELECTRIC LOCOMOTIVE> "locomotiva movida por corrente elétrica gerada por um motor a diesel"	
{diesel-hydraulic locomotive}	<DIESEL-HYDRAULIC LOCOMOTIVE> "locomotiva movida por transmissão hidráulica gerada por um motor a diesel"	
{dining car; diner; dining compartment; buffet car}	<dining car> "vagão de passageiros onde funciona serviço de restaurante em um trem"	*Synset*={vagão-restaurante}
{dinky; dinkey}	<dinky> "locomotiva pequena"	{GAP}
{dogcart}	<dogcart> "carroça pequena puxada por um cachorro"	{GAP}
{doodlebug}	<doodlebug> "veículo motorizado pequeno"	{GAP}
{dray}	<dray> "carroça baixa e de grande resistência, empregada no transporte de toras"	*Synset*={carretão}
{droshky}	<droshky> "carruagem leve e aberta, com quatro rodas, usada na Rússia e na Polônia"	{GAP}
{dump truck; dumper; tipper truck; tipper lorry; tip truck; tipper}	<dump truck> "caminhão cuja carga pode ser colocada na caçamba sem manipulação; a caçamba funciona com um movimento de básculo"	{GAP} *Phraset*={caminhão basculante}
{dumpcart}	<dumpcart> "carroça que se pode inclinar para despejo"	{GAP}
{dune buggy; beach buggy}	<dune buggy> "veículo recreativo aberto que trafega em qualquer tipo de terreno, com motor atrás, carroceria simplificada e pneus muito largos, usado especialmente em prais ou dunas"	*Synset*={*ingl.* buggy}
{electric car; electric automobile}	<electric car> "carro movido a energia elétrica"	*Synset*={automóvel elétrico; carro elétrico}
{electric locomotive}	<ELECTRIC LOCOMOTIVE> "locomotiva movida por um motor elétrico"	*Phraset*={locomotiva elétrica}

{fire engine; fire truck}	<fire engine> "caminhão destinado ao transporte de bombeiros e seus equipamentos contra incêndios"	{GAP} *Phraset*={caminhão de bombeiro; carro de bombeiro}
{flatcar; flatbed; flat}	<flatcar> "vagão de carga sem paredes laterais nem teto"	*Synset*={vagão-plataforma}
{forklift}	<forklift> "veículo com rodas que se locomove com seus próprios meios de propulsão, equipado com uma plataforma aforquilhada na parte da frente que pode ser inserida sob cargas para levantá-las e movê-las"	*Synset*={empilhadeira}
{four-wheel drive; 4WD}	<four-wheel drive> "veículo motorizado com tração nas quatro rodas"	*Synset*={todo-terreno} *Phraset*={veículo com tração nas quatro rodas; 4x4}
{four-wheeler}	<four-wheeler> "carruagem de quatro rodas, especialmente para aluguel"	{GAP}
{freight car}	<freight car> "vagão em que são transportadas mercadorias ou carga"	{GAP} *Phraset*={vagão de carga; vagão de mercadorias}
{funny wagon}	<funny wagon> "ambulância usada para transportar pacientes de e para hospitais psiquiátricos"	{GAP}
{garbage truck; dustcart}	<garbage truck> "caminhão usado na coleta de lixo"	{GAP} *Phraset*={caminhão de lixo}
{gas guzzler}	<gas guzzler> "carro pouco econômico, ou seja, que gasta muito combustível"	{GAP} *Phraset*={carro beberrão; carro gastão}
{gharry}	<gharry> "carruagem de aluguel na Índia"	{GAP}
{gig}	<gig> "carruagem pequena, com duas rodas, dois assentos e sem capota"	*Synset*={trole}
{go-kart}	<go-kart> "veículo motorizado de pequenas dimensões destinado às corridas"	*Synset*={*ingl.* kart}
{golfcart}	<golfcart> "veículo morotizado pequeno em que se transporta o equipamento de golfe pelo campo"	{GAP} *Phraset*={carrinho de golfe}
{gondola car; gondola}	<gondola car> "vagão de carga com laterais fixas e sem teto"	*Synset*={gôndola}
{guard's van}	<guard´s van> "vagão destinado ao transporte de policiais"	{GAP}

{gypsy cab}	<gypsy cab> "táxi que é licencionado somente para pegar passageiros que o solicitam por telefone, mas que freqüentemente procura passageiros nas ruas ilegalmente"	{GAP}
{hackney}	<hackney> "carruagem para aluguel"	{GAP}
{half track}	<half track> "veículo autopropulsionado que possui rodas comuns na parte dianteira e esteiras de aço rolantes ou lagartas na parte traseira"	{GAP}
{hand truck}	<hand truck> "carrinho de mão com duas rodas usado para transportar cargas pesadas; constituído por uma estrutura vertical que possui alças na parte superior e duas pás inferiores que são inseridas sob a carga"	{GAP}
{handcar}	<handcar> "vagão de carga pequeno e aberto (plataforma sobre rodas) que é impelido manualmente"	*Synset*={vagoneta}
{handcart; pushcart; cart; go-cart}	<handcart> "veículo com uma ou mais rodas, que pode ser empurrado por uma pessoa, usado para transportar várias mercadorias"	*Synset*={carrinho de mão; carriola; carro de mão}
{hansom; hansom cab}	<hansom> "carruagem com duas rodas, coberta, em que o condutor senta acima e atrás dos passgeiros"	*Synset*={fiacre}
{hardtop}	<hardtop> "carro que se parece com um conversível, mas o teto é fixo"	{GAP}
{hatchback}	<hatchback> "carro que possui duas grandes partes: a frente e o resto do carro"	{GAP}
{hearse}	<hearse> "veículo motorizado usado para transportar defuntos"	*Synset*={carro fúnebre}
{horse cart; horse-cart}	<horse cart> "veículo com duas rodas puxado normalmente por cavalo e usado para transpote em fazenda"	{GAP}
{horsecar}	<horsecar> "bonde movido por meio de tração aninal"	{GAP}
{horse-drawn vehicle}	<HORSE-DRAW VEHICLE> "veículo com rodas puxado por cavalos"	
{hot rod}	<hot rod> "carro velho modificado para obter melhor desempenho"	{GAP} *Phraset*={carro envenenado}
{ice wagon; ice-wagon}	<ice wagon> "carroça antiga, puxada a cavalo, que entregava gelo de porta em porta"	{GAP}

{jaunting car}	<jaunting car> "carroça irlandesa sem capota e com dois bancos longitudinais"	{GAP}
{jeep; landrover}	<jeep> "carro construído para ser usado em terrenos acidentados"	*Synset*={jipe}
{jinrikisha; ricksha; rickshaw}	<jinrikisha> "carroça pequena e leve, geralmente para um só passageiro, eventualmente para carga, puxado por um homem a pé; originário do Japão"	*Synset*={jinriquixá}
{ladder truck; aerial ladder truck}	<ladder truck> "carro de bombeiros equipado com escada"	{GAP}
{landau}	<landau> "carruagem com quatro rodas e dois bancos de passageiros que se defrontam, coberta com capota em fole dividida em duas seções que se podem arriar, levantar ou remover independentemente uma da outra"	*Synset*={fr. landau; landô}
{laundry cart}	<laundry cart> "carrinho de mão usado destinado ao transpote de roupa para lavar"	{GAP}
{laundry truck}	<laundry truck> "caminhão-baú usado por lavanderias para o transporte de roupas"	{GAP}
{limber}	<limber> "veículo com duas rodas puxado por cavalos, usado para rebocar peças de artilharia"	*Synset*={armão}
{limousine; limo}	<limousine> "carro grande e luxuoso, usualmente dirigido por um chofer"	*Synset*={limusine}
{loaner}	<loaner> "carro emprestado como substituto a um que está sendo reparado"	{GAP} *Phraset*={carro reserva}
{locomotive; railway locomotive}	<locomotive> "veículo com rodas que se locomove com seus próprios meios de propulsão, destinado a rebocar trens nas estradas de ferro"	*Synset*={locomotiva; locomotora}
{lorry; camion}	<lorry> "caminhão grande destinado ao transporte de cargas pesadas; usualmente sem laterais"	{GAP} *Phraset*={caminhão de carga}
{lorry}	<lorry> "carroça grande e baixa sem laterais"	{GAP}
{mail car}	<mail car> "vagão de carga destinado ao transporte de correspondências"	{GAP}
{milk float}	<milk float> "caminhão-baú com uma lateral aberta, usado na entrega de leite em domicílio"	{GAP}
{milk wagon; milkwagon}	<milk wagon> "carroça usada para entregar leite"	{GAP} *Phraset*={carroça de leite}

{minibike; motorbike}	<minibike> "motocicleta pequena e leve"	{GAP}
{minicab}	<minicab> "carro pequeno ou mini usado como táxi"	{GAP}
{minicar}	<minicar> "carro menor que um subcompacto"	{GAP}
{minivan}	<minivan> "van pequena, normalmente com três fileiras de assentos, sendo a última removível, destinada ao uso familiar"	{GAP}
{mobile home; manufactured home}	<mobile home> "trailer grande que pode ser conectado a utilitários e estacionado em lugares fixos para servir de moradia permanente"	{GAP} *Phraset*={casa sobre rodas}
{Model T}	<Model T> "o primeiro carro comercial"	{GAP}
{moped}	<moped> "motocicleta pequena e leve; pode ser movida por pedais ou por um motor de baixa potência"	*Synset*={mobilete} *Phraset*={bicicleta motorizada}
{motor scooter; scooter}	<motor scooter> "veículo com rodas pequenas e motor a gasolina de pouca potência engrenado à roda traseita"	*Synset*={lambreta; motoneta; *ingl.* scooter; vespa}
{motor vehicle; automotive vehicle}	<motor vehicle> "veículo com rodas autopropulsado que é dotado de motor e não anda sobre trilhos"	{GAP} *Phraset*={veículo motorizado; veículo a motor}
{motorcycle; bike}	<motorcycle> "veículo motorizado com duas rodas que pode transportar uma ou duas pessoas"	*Synset*={moto; *coloq.* motoca; motocicleta; motociclo}
{mountain bike; all-terrain bike; off-roader}	<mountain bike> "bicicleta de quadro leve e resistente projetada para prática esportiva em trilhas e terrenos acidentados"	*Synset*={*ingl.* mountain bike}
{moving van}	<moving van> "caminhão-baú usado em mudanças"	{GAP} *Phraset*={caminhão de mudança}
{nonsmoking car}	<NONSMOKING CAR> "vagão de passageiros destinado a pessoas que não fumam"	
{ordinary; ordinary bicycle}	<ordinary> "bicicleta antiga com uma roda dianteira grande e uma traseira pequena"	{GAP}
{oxcart}	<oxcart> "carroça, geralmente tosca, puxada por bois"	*Synset*={carro de boi}
{pace car}	<pace car> "carro que lidera uma corrida durante algumas voltas e depois deixa a pista para que a corrida continue"	*Synset*={carro-madrinha}

{panda car}	<panda car> "carro utilizado pela polícia para pecorrer as ruas; usualmente preto com uma faixa branca sobre a carroceria"	{GAP}
{panzer}	<panzer> "veículo com rodas blindado de combate ou tanque"	*Synset*={*ingl.* panzer}
{parlor car; parlour car; drawing room car; palace car; chair car}	<parlor car> "vagão de passageiros luxuoso e privativo destinado à viagens durante o dia e a noite"	*Synset*={carro-salão}
{passenger car; coach; carriage}	<passenger car> "vagão destinado ao transporte de passageiros"	{GAP} *Phraset*={carro de passageiros; vagão de passageiros}
{passenger van}	<passenger van> "veículo parecido com um caminhão-baú pequeno, usado no transporte de passageiros"	*Synset*={*ingl.* van; furgão}
{pastry cart}	<pastry cart> "carrinho de mão usado para servir sobremesas"	{GAP} *Phraset*={carrinho de sobremesas}
{pedicab}	<pedicab> "triciclo coberto, movido a pedal, para transportar passageiros"	{GAP}
{pickup; pickup truck}	<pickup> "caminhão pequeno, geralmente aberto e com laterais baixas, usado para transporte de mercadorias"	*Synset*={caminhonete; picape; *ingl.* pickup}
{pilot engine}	<pilot engine> "locomotiva que precede o trem para checar os trilhos"	{GAP}
{police van; police wagon; paddy wagon; patrol wagon; wagon; black Maria}	<police van> "van usada pela polícia"	*Synset*={camburão} *Phraset*={carro de presos}
{pony cart; ponycart; donkey cart; tub-cart}	<pony cart> "carroça que se assemelha a uma tina de lavar roupa"	{GAP}
{post chaise}	<post chaise> "carruagem com quatro rodas, fechada, usada para o transporte rápido de passageiros e correspondência no séc. XVIII e início do XIX"	{GAP} *Phraset*={diligência postal}
{Pullman; Pullman car}	<Pullman> "vagão de passageiros luxuoso destinado à viagens durante o dia e a noite"	*Synset*={*ingl.* pullman}
{race car; racing car}	<race car> "carro que compete em corridas de velocidade"	*Synset*={carro de corrida}
{railcar; car; railway car; railroad car}	<railcar> "veículo com rodas adaptado para se locomover sobre trilhos"	*Synset*={carro; vagão}

{reconnaissance vehicle; scout car}	<RECONNAISSANCE VEHICLE> "veículo autopropulsionado, usualmente blindado, rápido e sem teto, usado pelos militares para fazer o reconhecimento de terrenos"	*Phraset*={veículo de reconhecimento}
{recreational vehicle; RV; R.V.}	<recreational vehicle> "veículo autopropulsionado, geralmente grande, em que se pode dormir e cozinhar, usado para acampamento ou outra atividade recreativa"	{GAP} *Phraset*={veículo recreativo}
{refrigerator car}	<refrigerator car> "vagão de carga equipado com sistema de refrigeração"	*Synset*={vagão-frigorífico}
{remise}	<remise> "carruagem com quatro rodas, especialmente para aluguel; luxuosa"	{GAP}
{roadster; runabout; two-seater}	<roadster> "carro aberto para duas pessoas, com um assento na parte da frente e outro suplementar"	*Synset*={baratinha; *ingl.* roadster} *Phraset*={dois lugares; dois-lugares}
{safety bicycle; safety bike}	<safety bicycle> "bicicleta cujas rodas são do mesmo tamanho e a transmissão é feita por corrente"	{GAP}
{scooter}	<scooter> "veículo com duas rodas para criança, constituído por uma tábua sobre duas rodas no sentido longitudinal, onde se apóia um dos pés, enquanto se dá impulso com o outro"	*Synset*={patinete}
{sedan}	<sedan> "carro que possui três grandes partes: a frente, o meio (em que ficam os passageiros) e o porta-malas"	*Synset*={sedã}
{self-propelled vehicle}	<SELF-PROPELLED VEHICLE> "veículo com rodas que se locomove com seus próprios meios de propulsão"	*Phraset*={veículo autopropulsado; veículo autopropulsionado}
{serving cart}	<serving cart> "carrinho de mão usado para servir à mesa	{GAP}
{shopping cart}	<shopping cart> "carrinho de mão usado para carregar compras"	{GAP} *Phraset*={carrinho de compras}
{shunter}	<shunter> "locomotiva usada para mover vagões, mas não durante viagens"	{GAP}
{skateboard}	<skateboard> "veículo com rodas, constituído basicamente por uma pequena prancha estreita, não muito longa, sobre a qual se equilibra e desloca impelido por um dos pés"	*Synset*={esqueite; *ingl.* skate; *ingl.* skateboard}

{skidder}	<skidder> "trator grande e pesado, equpado com guincho, usado principalmente para rebocar toras em terrenos florestais"	{GAP} *Phraset*={trator florestal}
{sleeping car; wagon-lit}	<sleeping car> "vagão de passageiros equipado com camas ou beliches, para viagens noturnas"	*Synset*={carro-leito; vagão-dormitório; vagão-leito}
{slip coach}	<slip coach> "vagão engatado ao final do trem; pode ser desengatado na estação com o trem em movimento"	{GAP}
{smoking car; smoking carriage; smoking compartment}	<smoking car> "vagão de passageiros destinado a fumantes"	{GAP} *Phraset*={vagão para fumantes}
{sno-cat}	<sno-cat> "veículo autopropulsionado que se locomove por meio de esteiras de aço rolantes ou lagartas, especialmente destinado à remoção de neve"	{GAP}
{snowmobile}	<snowmobile> "veículo autopropulsionado que se locomove por meio de esteiras de aço rolantes ou lagartas, especialmente destinado à remoção de neve"	{GAP}
{snowplow; snowplough}	<snowplow> "veículo motorizado usado para limpar a neve das ruas e estradas"	{GAP}
{sound truck}	<sound truck> "caminhão equipado com alto-falantes, usado para fazer propaganda"	{GAP} *Phraset*={caminhão de som}
{sport utility; sport utility vehicle; SUV}	<sport utility> "carro resistente, empregado geralmente no transporte de mercadorias"	{GAP} *Phraset*={utilitário esportivo; utilitário-esportivo; veículo utilitário esportivo}
{sports car; sport car}	<sports car> "carro com duas portas, de formato aerodinâmico e munido de motor potente para atingir grandes velocidades"	*Synset*={carro esporte} *Phraset*={carro esportivo}
{stagecoach; stage}	<stagecoach> "carruagem puxada por quatro cavalos e conduzida por apenas uma pessoa, usada para o transporte de passageiros e de correspondências em rotas regulares entre cidades"	*Synset*={diligência}
{stanhope}	<stanhope> "carruagem leve e aberta, com duas ou quatro rodas"	{GAP}
{Stanley Steamer}	<STANLEY STEAMER> "carro antigo movido por motor a vapor"	*Phraset*={Stanley Steamer}
{steam locomotive}	<steam locomotive> "locomotiva movida por uma máquina a vapor"	*Synset*={maria-fumaça} *Phraset*={locomotiva a vapor}

{stock car}	<stock car> "carro de corrida que tem o chassi de um carro comum"	*Synset*={*ingl.* stock car}
{stockcar}	<stockcar> "vagão de carga fechado usado para transportar animais"	{GAP}
{streetcar; tram; tramcar; trolley; trolley car}	<streetcar> "veículo autopropulsionado movido por meio de energia elétrica e que se locomove sobre trilhos, empregado no transporte urbano"	*Synset*={bonde; trâmuei; *ingl.* tramway}
{subcompact; subcompact car}	<subcompact> "carro menor que o compacto"	{GAP} *Phraset*={subcompacto; carro subcompacto}
{sulky}	<sulky> "veículo com duas rodas puxado por apenas um cavalo, destinado ao transporte de uma pessoa"	{GAP}
{surrey}	<surrey> "carruagem leve com quatro rodas e dois lugares frente a frente"	{GAP}
{swamp buggy; marsh buggy}	<swamp buggy> "veículo anfíbio que tipicamente possui quatro rodas e um corpo grande"	{GAP}
{switch engine; donkey engine}	<switch engine> "locomotiva para manobrar material rodante em estradas de ferro"	{GAP} *Phraset*={locomotiva de manobras}
{tandem trailer}	<tandem trailer> "caminhão grande com um cavalo mecânico e dois reboques (ou carretas) um atrás do outro"	{GAP}
{tank car; tank}	<tank car> "vagão de carga equipado com sistema de refrigeração usado para transportar líquidos, especialmente combustíveis"	*Synset*={vagão-tanque}
{tank destroyer}	<tank destroyer> "veículo com rodas blindado, equipado com canhão antitanque e capaz de atingir altas velocidades"	{GAP} *Phraset*={peça antitanque}
{tank engine; tank locomotive}	<tank engine> "locomotiva que carrega seu próprio combustível e água; não precisa de um tênder"	{GAP} *Phraset*={locomotiva tênder}
{tank; army tank}	<tank> "veículo com rodas militar fechado, equipado com metralhadoras ou canhão (antiaéreo), próprio para locomover-se em terrenos acidentados, graças a um sistema especial de tração, o de lagartas"	*Synset*={carro-de-combate; tanque} *Phraset*={tanque de guerra}
{tea cart; teacart; tea trolley; tea wagon}	<tea cart> "carrinho de mão usado para sevir chá ou refrescos"	{GAP} *Phraset*={carrinho de chá}

{technical}	<technical> "caminhonete equipada com uma arma na caçamba; normalmente, uma metralhadora"	{GAP}
{tender}	<tender> "vagão de carga engatado à locomotiva que transporta o suprimento de água e combustível para abastecer a máquina"	*Synset*={tênder}
{touring car; phaeton; tourer}	<touring car> "carro grande e aberto para quatro pessoas com teto conversível"	{GAP} *Phraset*={carro aberto}
{tow truck; tow car; wrecker}	<tow truck> "caminhão equipado com um pequeno guindaste, usado em geral para rebocar carros enguiçados"	*Synset*={carro-guincho; carro-socorro; guincho; reboque} *Phraset*={carro de socorro}
{tracked vehicle}	<tracked vehicle> "veículo autopropulsionado que se locomove por meio de esteiras de aço rolantes (lagartas)"	{GAP}
{traction engine}	<traction engine> "locomotiva movida por uma máquina a vapor, usada especialmente no campo para rebocar carga ou fardo pesado"	{GAP}
{tractor}	<tractor> "veículo autopropulsionado equipado com roda grandes, destinado ao reboque de cargas ou de operar equipamentos agrícolas, de terraplenagem, etc."	*Synset*={trator}
{trail bike; dirt bike}	<trial bike> "motocicleta leve, equipada com pneus e suspensões resistentes, projetada para terrenos acidentados"	{GAP}
{trailer truck; tractor trailer; trucking rig; rig; articulated lorry; semi}	<trailer truck> "caminhão grande constituído por duas partes, o cavalo mecânico e o reboque (ou carreta)"	*Synset*={carreta; jamanta}
{trailer; house trailer}	<trailer> "veículo com rodas fechado que serve de moradia e que pode ser rebocado por outro veículo"	*Synset*={*ingl.* trailer}
{tramcar; tram}	<tramcar> "carroça com quatro rodas que se locomove sobre trilhos (em minas) para transporte de minérios"	*Synset*={vagoneta}
{transporter; car transporter}	<transporter> "caminhão com carroceria grande especial para transporte de carros de passeio"	*Synset*={caminhão-cegonha; cegonha} *Phraset*={caminhão cegonheiro}

{trap}	<trap> "carruagem leve com duas rodas"	*Synset*={aranha}
{tricycle; trike}	<tricycle> "veículo com três rodas, que se move por meio de dois pedais ligados à roda dianteira"	*Synset*={triciclo; velocípede}
{troika}	<troika> "carruagem russa puxada por três cavalos emparelhados"	*Synset*={tróica}
{truck; motortruck}	<truck> "veículo motorizado, com quatro ou mais rodas, para transporte de carga"	*Synset*={caminhão}
{tumbrel; tumbril}	<trumbel> "carroça que se pode inclinar para despejo, destinada ao carregamento de esterco; esse tipo de carroça foi usado para transportar prisioneiros durante a Revolução Francesa"	{GAP}
{unicycle; monocycle}	<unicycle> "veículo com uma única roda à qual os pedais que a impulsionam estão ligados"	*Synset*={monociclo}
{used-car; secondhand car}	<used-car> "carro que já teve um dono; que não é novo"	{GAP} *Phraset*={carro usado; carro de segunda mão; carro seminovo}
{van; caravan}	<van> "veículo recreativo, adaptado para servir de moradia durante viagens, especialmente equipado com ambiente para dormir"	{GAP}
{van}	<van> "caminhão cuja carroceria é fechada"	*Synset*={caminhão-baú}
{velocipede}	<velocipede> "bicicleta antiga que se move impelida por dois pedais ligados à roda dianteira"	*Synset*={velocípede}
{wagon; waggon}	<wagon> "veículo com rodas que precisa ser rebocado ou puxado"	{GAP}
{wain}	<wain> "carroça grande e aberta, usada em trabalhos rurais"	{GAP}
{water cart}	<water cart> "carroça equipada com um tanque para carregar água"	{GAP}
{watering cart}	<watering cart> "carroça equipado com um tanque para carregar água e com uma espécie de esguicho para molhar as estradas"	{GAP}
{weapons carrier}	<weapons carrier> "veículo autopropulsado que se parece com um caminhão pequeno, cuja carroceria é equipada com bancos laterais, o que permite o transporte de armas e de tropas pelos militares"	{GAP}

{welcome wagon}	<welcome wagon> "veículo com rodas que precisa ser rebocado ou puxado, usualmante destinado a fornecer informações e presentes de comerciantes locais para os novos moradores da área"	{GAP}
{wheeled vehicle}	<WHEELED VEHICLE> "veículo que se move sobre rodas e que possui compartimento para transportar pessoas/ coisas"	*Phraset*={veículo com roda}

Quadro 11: Os conceitos lexicalizados no AmE do domínio dos "veículos com rodas" e sua expressão no PB.

Com base nos dados do Quadro 11, apresentam-se, a seguir, as estatísticas sobre as expressões lingüísticas do PB (Quadro 12).

6.7.1. Dados estatísticos

Interpretando-se o Quadro 12, observa-se que, dos 205 conceitos lexicalizados no AmE que pertencem ao domínio dos "veículos com rodas", apenas 84 estão lexicalizados no PB, o que equivale aproximadamente a 40,9% do total de conceitos analisados. Dessa forma, nota-se que no domínio conceitual dos "veículos com rodas", menos da metade dos conceitos analisados lexicalizam-se no PB. Para os demais 121 conceitos (ou 59,1%), o PB apresenta lacunas lexicais, ou seja, o PB não possui unidades lexicais para expressar tais conceitos.

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

Quadro 12: As estatísticas das lexicalizações estudadas no PB.

Dentre os 84 conceitos lexicalizados no PB e codificados em termos de *synsets*, 11 deles possuem um *phraset* sinômino como informação adicional, o que equivale a 13% do total de conceitos que o PB lexicaliza. Os demais 73 (ou 87%) não possuem *phraset* sinônimo.

Dentre as 121 lacunas, observa-se que, em 40 casos, foi possível identificar um *phraset* que expressa no PB o conceito que é expresso por unidades lexicais no AmE. Em outras palavras, pode-se dizer que, para 33% das lacunas, foi possível montar um conjunto de SLRs.

Para as demais 81 lacunas, não foi possível identificar SLRs correspondentes, o que equivale a 67% do total de lacunas lexicais identificadas no PB.

6.7.2. Observações sob a perspectiva semasiológica

Também com base nos dados do Quadro 11, é possível tecer algumas observações sob a perspectiva semasiológica, ou seja, aquela em que se parte das expressões da língua em direção à identificação dos conceitos.

Sob essa perspectiva, vê-se, em especial, vários casos em que uma mesma unidade lexical expressa conceitos distintos, ou seja, casos de unidades **polissêmicas**. Com base no critério semântico, diz-se que a polissemia ocorre quando, na oposição significativa das formas, há intersecção de pelo menos um traço componencial (LYONS, 1981, ALMEIDA, 1990, VILELA, SILVA, 2004). Assim, as seguintes unidades, por exemplo, configuram-se como polissêmicas: *carro, carruagem* e *vagoneta.*

A unidade *carro*, por exemplo, expressa dois conceitos distintos: <veículo com rodas adaptado para se locomover sobre trilhos>, que também é lexicalizado por *vagão*, e <veículo com quatro rodas e um motor, destinado ao transporte de passageiros ou cargas>, que também é lexicalizado por *auto* e *automóvel*. Nesse caso, pode-se observar a intersecção de pelo menos um traço componencional, que pode ser assim descrito: [andar sobre rodas]. A unidade *carruagem* expressa dois conceitos distintos, porém relacionados, são eles: <veículo com quatro rodas puxado por cavalos, dotado de suspensão por molas, usado para transportar pessoas>, que também é lexicalizado por *coche*, e <carruagem leve com quatro rodas usada em cerimoniais>. Aqui, percebe-se o traço comum: [ser puxado por cavalos]. A unidade *vagoneta*, por fim, expressa o conceito <vagão de carga pequeno e aberto, ou plataforma sobre rodas, que é impelido manualmente> e <carroça com quatro rodas que se locomove sobre trilhos (em minas) para transporte de minérios>. Nesse caso, percebe-se o traço [andar sobre trilhos].

Assumindo-se os pressupostos da WN.Pr, uma unidade lexical, quando polissêmica, pertence a mais de um *synset*, posto que cada *synset* codifica um conceito específico. No âmbito da WN.Pr, a polissemia é descrita ou entendida em função do que se denominou **matriz lexical** (FELLBAUM, 1998a,b).

O Quadro 13, em que são consideradas apenas as unidades *carro, vagão, automóvel, auto, sege, coche* e *carruagem*, ilustra a noção de matriz lexical. Em tal matriz, as unidades lexicais (no caso, do PB) (ou formas lingüísticas) estão listadas como cabeçalhos das colunas e os conceitos (codificados em *synsets*), como cabeçalhos das linhas. Como o mapeamento entre

formas e conceitos é muitos:muitos, quando há duas células preenchidas em uma mesma coluna, diz-se que a unidade é polissêmica.

Assim, a unidade *carro,* ou forma F1, por expressar os conceitos codificados pelos *synsets* S1 e S2, é considerada polissêmica e, por isso, é elemento constitutivo desses dois *synsets*. A noção de matriz lexical também explicita a sinonímia: quando há duas células preenchidas em uma mesma linha, diz-se que as unidades são sinônimas. O Quadro 13, por exemplo, mostra que as unidades *carro* e *vagão*; *auto, automóvel* e *carro*; *carruagem, coche* e *sege*, formam *synsets*.

Conceitos Lexicalizados (*synsets*)	**Formas/ Unidades lexicais**						
	F1 *carro*	**F2** *vagão*	**F3** *automóvel*	**F4** *auto*	**F5** *sege*	**F6** *coche*	**F7** *carruagem*
S1={carro; vagão}	S1<F1	S1<F2					
S2={ auto; automóvel; carro}	S2<F1		S2<F3	S2<F4			
S3={carruagem; coche; sege}					S3<F5	S3<F6	S3<F7
S4={carruagem}							S4<F2

Quadro 13: A matriz lexical, polissemia e a sinonímia.

6.8. Síntese da Seção VI

Nesta Seção, parte dos "objetos concretos discretos" e sua representação no PB foram identificadas.

A delimitação dos conceitos foi, do ponto de vista prático, guiada pela proposição e prenchimento de *templates* conceituais. O *template* conceitual foi elaborado com base nos recursos teóricos fornecidos pelo MultiNet (tipo conceitual, traços semânticos, atributos multidimensionais e relações conceituais) e em certas informações adicionais (UniC e glosa).

A identificação das expressões lingüísticas também foi guiada pelo preenchimento de um *template*, no caso, lexical. O *template* lexical constitui-se por dois campos principais: expressões lingüísticas e frases-exemplo. A identificação das expressões lingüísticas do PB (unidades lexicais e SLRs) foi realizada manualmente, utilizando-se, como fonte, recursos de natureza lingüístico-computacional, lexicográfica e textual e seguindo determinada definição de unidade lexical. Além das unidades lexicais, realizou-se a identificação dos SLRs, que constituem informação relevante para o PLN.

Assim, ao final dessas análises, propõe-se, para cada conceito analisado, um *template* léxico-conceitual correspondente, que é a união dos *templates* conceitual e lexical.

Quanto às expressões lingüísticas do PB, salienta-se que, dos 205 conceitos analisados, 84 são expressos por unidades lexicais no PB; no restante dos casos, lacunas lexicais foram detectadas.

Na próxima Seção, descrevem-se a implementação de parte do alinhamento entre as bases da WN.Pr e WN.Br e a criação de uma base lexical.

Seção VII

Construção da base léxico-conceitual bilíngüe e o alinhamento das WN.Pr e WN.Br

Nesta Seção, discutem-se:

a) o editor, ou seja, a ferramenta computacional que auxilia a construção da base bilíngüe em que os conceitos lexicalizados pelo AmE e PB são alinhados por meio da interlíngua estruturada e formal; de um modo geral, nesse editor, é possível representarem-se os conceitos da interlíngua no formato *multinet* e suas respectivas lexicalizações no AmE e no PB;
b) as estratégias de inserção, no editor, das informações especificadas nos *templates* léxico-conceituais e a subseqüente construção da base léxico-conceitual bilíngüe;
c) a proposição de novas funcionalidades para o editor da WN.Br que permitem o alinhamento das bases da WN.Pr e da WN.Br; por meio da extensão do editor, o alinhamento das bases é realizado com a estratégia "assistida por computador" (do inglês, *computer-aided*), o que auxilia o trabalho do lingüista.

Antes, porém, dessas discussões, duas ressalvas são importantes.

A primeira diz respeito ao fato de que o editor denominado MWR (do alemão, *Multinet WissensRepräsentation*, isto é, Representação do conhecimento MultiNet) (HELBIG, 2006), que tem como base teórica o próprio paradigma MultiNet, não está disponível. Dessa forma, foi preciso buscar formas alternativas para a construção da base bilíngüe. E é aqui que a segunda ressalva se encaixa. Ao retomar o pressuposto segundo o qual determinado conhecimento, quando representado por um modelo de RC, pode ser definido como uma "ontologia" ("especificação formal de uma conceitualização compartilhada"), conclui-se que a representação do mesmo pode ser feita com o auxílio dos recursos desenvolvidos na área denominada Engenharia de Ontologias[60] (do inglês, *Ontology Engineering*).

[60] A Engenharia de Ontologias é considerada a sucessora da Engenharia do Conhecimento (MIZOGUCHI, 2004).

Trata-se, mais especificamente, dos recursos classificados sob o rótulo "editores de ontologia", ou seja, ferramentas computacionais que auxiliam as tarefas de criação e manutenção de ontologias.

Atualmente, diante das inúmeras possibilidades de aplicação das ontologias, como em sistemas de recuperação de informação, sistemas de extração de informação, sistemas de sumarização, desambiguação lexical de sentido (VOSSEN, 2004), etc., há um grande número de editores (MIZOGUCHI, 2004). Dentre eles, citam-se: o OntoEdit[61] (SURE et al, 2002), o WebODE[62] (CORCHO et al., 2002), o Protégé-Frames[63](GENNARI et al., 2003), o Protégé-OWL (HORRIDGE et al., 2004, KNUBLAUCH et al., 2004) e o Hozo (KOZAKI, 2002)[64].

Não foi possível identificar um editor que tratasse especificamente de redes semânticas. Assim, a investigação passou a ser guiada primordialmente pela necessidade de seleção de um editor para representar, da forma mais apropriada, as informações contidas no *template* léxico-conceitual. Em outras palavras, investigou-se, em primeiro lugar, a possibilidade de os referidos editores representarem: (i) uma rede semântica; (b) os tipos conceituais, (c) os atributos multidimensionais; (d) o mecanismo de herança entre os conceitos; (e) as glosas; (f) as lexicalizações do AmE e do PB (e também os SLRs do PB), as quais podem ser vistas como as parcelas dos léxicos dessas línguas responsáveis por recobrir o domínio dos "veículos com rodas"; e (g) as frases-exemplo para cada lexicalização ou unidade lexical (e SLR).

Diante desses requisitos, optou-se pelo editor Protégé-OWL (versão 3.3.1), apresentado na na Seção 7.1, em que se descrevem especificamente as principais características desse editor.

7.1. O editor Protégé

O Protégé é um editor de ontologias desenvolvido pelo Centro de Informática Biomédica (do inglês, *Stanford Center for Biomedical Informatics Research*) da Escola de Medicina da Universidadde Stanford (do inglês, *Stanford University School of Medicine*) no final da década de 1990.

Do ponto de vista computacional, o Protégé é um ***software* livre** e de **código aberto**[65], que auxilia o desenvolvimento de ontologias e sistemas baseados em conhecimento[66]. Além disso,

[61] http://www.ontoknowledge.org/tools/ontoedit.shtml
[62] http://webode.dia.fi.upm.es/WebODEWeb/index.html
[63]http://protege.stanford.edu
[64] http://www.hozo.jp
[65] Ou seja, um programa de computador que pode ser usado, copiado, estudado, modificado e redistribuído sem nenhuma restrição.

esse editor é uma aplicação *stand-alone*[67] (ou ainda *desktop*). Essa ferramenta foi elaborada com base em três premissas principais: interoperabilidade[68], usabilidade[69] e extensibilidade[70], que também constribuíram para a escolha desse editor.

7.1.1. O editor Protégé-OWL: noções preliminares

Atualmente, a plataforma do Protégé inclui duas metalinguagens formais, as quais dão origem às versões Protégé-Frames e Protégé-OWL[71]. O Protégé-Frame, como o próprio nome indica, baseia-se na tecnologia ou metalinguagem formal dos *frames*. O Protégé-OWL, por sua vez, é a extensão do Protégé que modela uma ontologia em função da linguagem formal denominada *Web Ontology Language* (OWL)[72].

Diante da adoção da premissa da interoperabilidade, o editor Protégé-OWL, além de suportar a linguagem OWL, também suporta a importação e exportação de bases de dados em outros formatos, como N_3 (do inglês, *Notation 3*), TURTLE (do inglês, *Terse RDF Triple Language*), HTML (do inglês, *HyperText Markup Language*) e RDF/XML (do inglês, *Resource Description Framework* e *Extensible Markup Language*), todos eles destinados à estruturação das informações na Web (SMITH et al, 2004).

Com a adoção da premissa da expansão, o Protégé-OWL pode ser expandido por vários *plug-ins*[73]. Dentre eles, destaca-se o *plug-in* denominado TGVizTab (do inglês, *TouchGraph Visualisation Tab*) (ALANI, 2003), que permite a visualização de ontologias por meio da tecnologia TouchGraph[74]. Mais especificamente, o TGVizTab é responsável por gerar uma "representação gráfica (da rede) interativa" das ontologias, em que os conceitos e relações

---

[66] Na *webpage* do Protégé, é possível encontrar um número relativamente grande de ontologias, para várias áreas do conhecimento, que foram construídas com o auxílio dessa ferramenta computacional (http://protege.stanford.edu/download/ontologies.html).

[67] Ou seja, um programa completamente auto-suficiente que para ser instalado em um computador pessoal não requer um software auxiliar (MICROSOFT PRESS, 1998). O termo stand-alone é comumente utilizado em oposição à "aplicação Web", que designa sistemas computacionais projetados para utilização através de um navegador, seja na internet ou em "redes privadas" (no inglês, intranets) (MICROSOFT PRESS, 1998).

[68] Premissa segundo a qual se busca consentir a compatibilidade com outros sistemas de representação do conhecimento.

[69] Premissa segundo a qual se busca garantir a facilidade de uso da ferramenta.

[70] Premissa segundo a qual se busca garantir o crescimento do programa pela adição de novos componentes.

[71] O Protégé-OWL engloba três "sublinguagens OWL": OWL-Lite, OWL-DL e OWL-Full. A OWL-Full é considerada a mais expressiva das três e, por isso, foi a escolhida neste trabalho.

[72] A OWL é a mais recente linguagem desenvolvida pelo *World Wide Web Consortium* (W3C) (http://www.w3.org/) para promover a Web Semântica, uma proposta feita por Berners-Lee (2001) para a estruturação dos documentos da Web. Nesse cenário, a OWL foi projetada como anotação-padrão para o conteúdo semântico a ser disponibilizado na Web.

[73] 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

[74] http://www.touchgraph.com/

entre eles estão dispostos no espaço. Esse *plug-in* teve papel importate, como descrito na Subseção 7.4.3.2 (pág. 188), no processo de identificação das relações interlinguais segundo o projeto EuroWordNet.

Na Seção 7.2, serão apresentadas as estratégias de inserção dos dados no referido editor.

7.2. Adaptações do editor para o alinhamento e para a criação da base bilíngüe

7.2.1. Inserção das informações no editor e a criação da base bilíngüe

A linguagem OWL possiu três construtos centrais: as **classes**, as **propriedades** e os **indivíduos**. Como conseqüência, o editor Protégé-OWL também está pautado nesses três elementos. Ao longo desta Subseção, apresentam-se as adaptações feitas no editor para que as informações contidas no *template* léxico-conceitual pudessem ser inseridas no Protégé-OWL. Para tanto, fez-se uma adaptação dos construtos do editor para que as informações do *template* léxico-conceitual pudessem ser editadas.

7.2.1.1. Os conceitos como classes

Os conceitos da rede semântica foram inseridos no Protégé-OWL como `classes`. As `classes`, representações concretas dos conceitos, organizam-se hierarquicamente, contemplando, assim, a relação SUB. Nessa hierarquia, as subclasses especializam (ou seja, "são subordinadas a") sua superclasse. Por exemplo, considerando-se os conceitos WheeledVehicle[75] e Skateboard, representações concretas do conceitos <wheeled vehicle> e <skateboard> no editor, a relação entre eles se estabelece da seguinte maneira: Skateboard é um conceito mais específico que WheeledVehicle e este, conseqüentemente, é um conceito mais geral que Skateboard. Em outras palavras, WheeledVehicle relaciona-se a Skateboard por meio da relação SUB. Assim, considerando os casos prototípicos, diz-se que "os skatebords são veículos com rodas", isto é, "os membros da classe Skateboard são membros da classe WheeledVehicle".

A construção da hierarquia conceitual, assim como o próprio preenchimento dos *templates* léxico-conceituais, foi feita com base no método denominado **top-down**, ou seja, partiu-se da

[75] A nomeação das classes segue Knublauch et al. (2004), os quais recomendam que as classes de nomes complexos sejam nomeadas com as iniciais de seus elementos constitutivos em maiúscula e sem espaço entre eles, p.ex.: MotorVehicle. Além disso, devido a uma restrição do editor, as classes de mesmo nome receberam um indicador numérico para diferenciá-las, como Lorry1 (<caminhão grande destinado ao transporte de cargas pesadas; usualmente sem laterais>) e Lorry2 (<carroça grande e baixa sem laterais>).

especificação do conceito mais genérico para a subseqüente especificação dos conceitos mais específicos (USCHOLD, GRUNINGER 1996).

A Figura 41 mostra a interface (ou GUI) principal do editor, que é composta por diferentes abas. Dentre elas, a Figura destaca a aba `OWL Classes`, responsável pela manipulação dos conceitos.

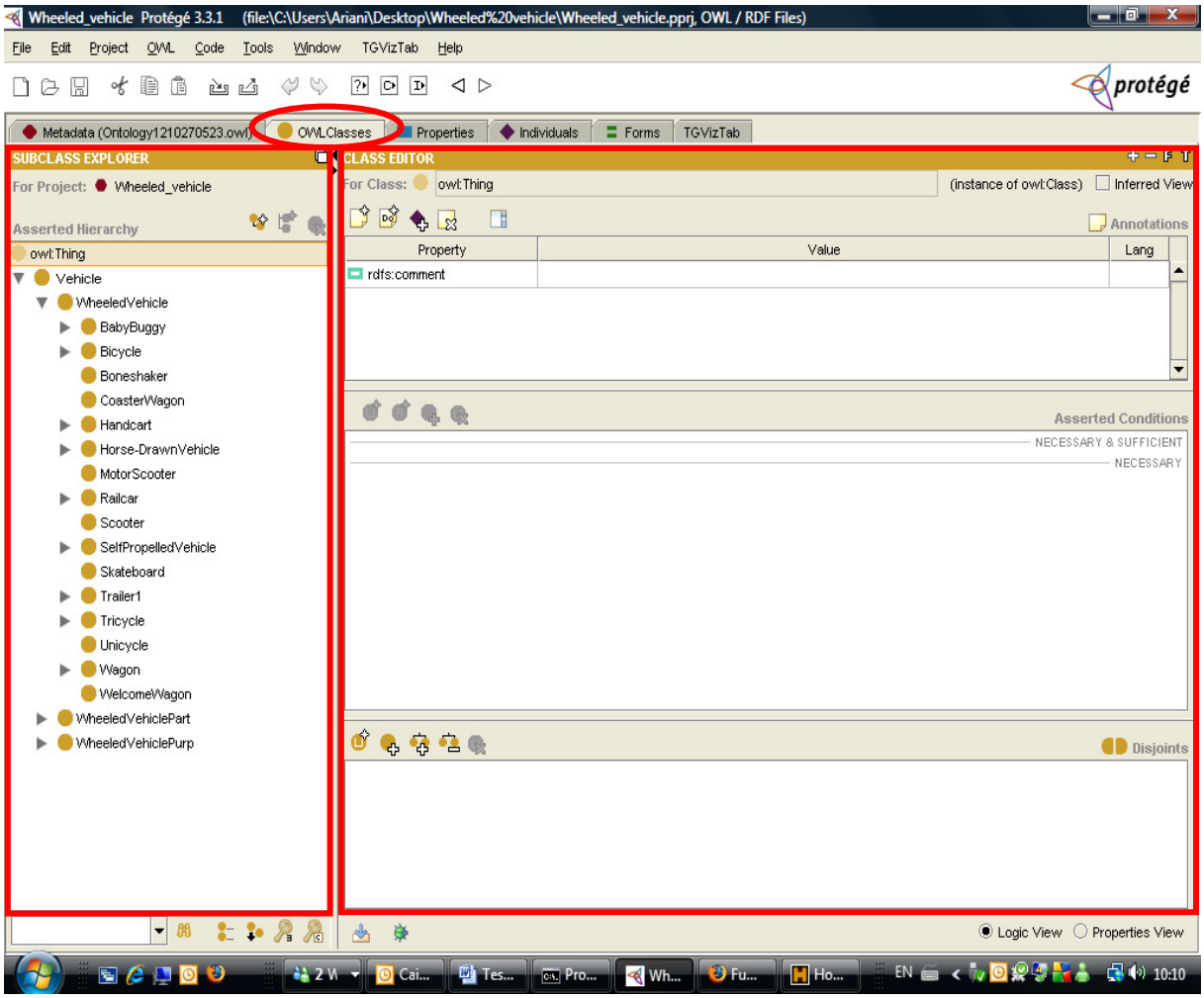


Figura 41: Os conceitos como classes.

A aba selecionada na Figura e destacada pelo círculo vermelho é responsável pela abertura de uma janela[76] composta por dois módulos. O módulo em que a hierarquia de conceitos pode ser visualizada no formato de árvore (do inglês, *folder-tree*) localiza-se à esquerda e é denominado `Subclass Explorer`. O módulo à direita, denominado `Class Editor`, constitui a janela em que os conceitos são gerenciados, ou seja, criados e/ou manipulados.

[76] Na Informática, uma janela é uma área visual, geralmente em formato retangular, usada para exibir informação especial, que pode ser selecionada e examinada a qualquer momento e que se sobrepõe à informação que já estiver na tela (MICROSOFT PRESS, 1998, p.768).

Na Figura 41, os conceitos que representam as partes dos veículos com rodas e que, conseqüemente, estão relacionados aos conceitos que expressam os vários veículos com rodas por meio da relação PARS, e o conceito que expressa a função e que, por conseguinte, está relacionado aos conceitos que expressam os vários veículos com rodas por meio da relação PURP, foram agrupados em termos dos conceitos genéricos e não-lexicalizados WheeledVehiclePart e WheeledVehiclePurp, respectivamente. Isso ocorreu devido a uma restrição do editor para que as relações (PARS e PURP) entre os conceitos pudessem ser estabelecidas. Tanto os conceitos organizados sob o conceito geral WheeledVehicle quanto aqueles subordinados aos conceitos WheeledVehiclePart e WheeledVehiclePurp constituem o domínio conceitual em questão. Tal domínio está articulado em termos do conceito Vehicle, que, por sua vez, está subordinado à classe owl:Thing[77].

Uma vez adicionados os superconceitos WheeledVehiclePart, WheeledVehiclePurp e WheeledVehicle, o editor permite especificar que tais conceitos são **disjuntos**, como indicado pelo retângulo vermelho inferior da Figura 42. Isso quer dizer que um "indivíduo" só pode ser instância de apenas um desses três superconceitos.

Ainda quanto à edição dos conceitos no editor, ressalta-se que a glosa, responsável por definir informalmente o conceito, foi introduzida como o valor da propriedade `rdfs:comment`, a qual é destinada exatamente para a inserção de informações adicionais sobre os conceitos. Além da glosa, essa propriedade também permite a especificação da língua em que o comentário foi descrito. No caso, às glosas, foi associada a sigla PB (cf. Figura 42).

[77] A classe `owl:Thing` é parte do vocabulário da própria linguagem OWL. Essa classe equivale ao tipo **entidade** [*ent*] da hierarquia de tipos do MultiNet.

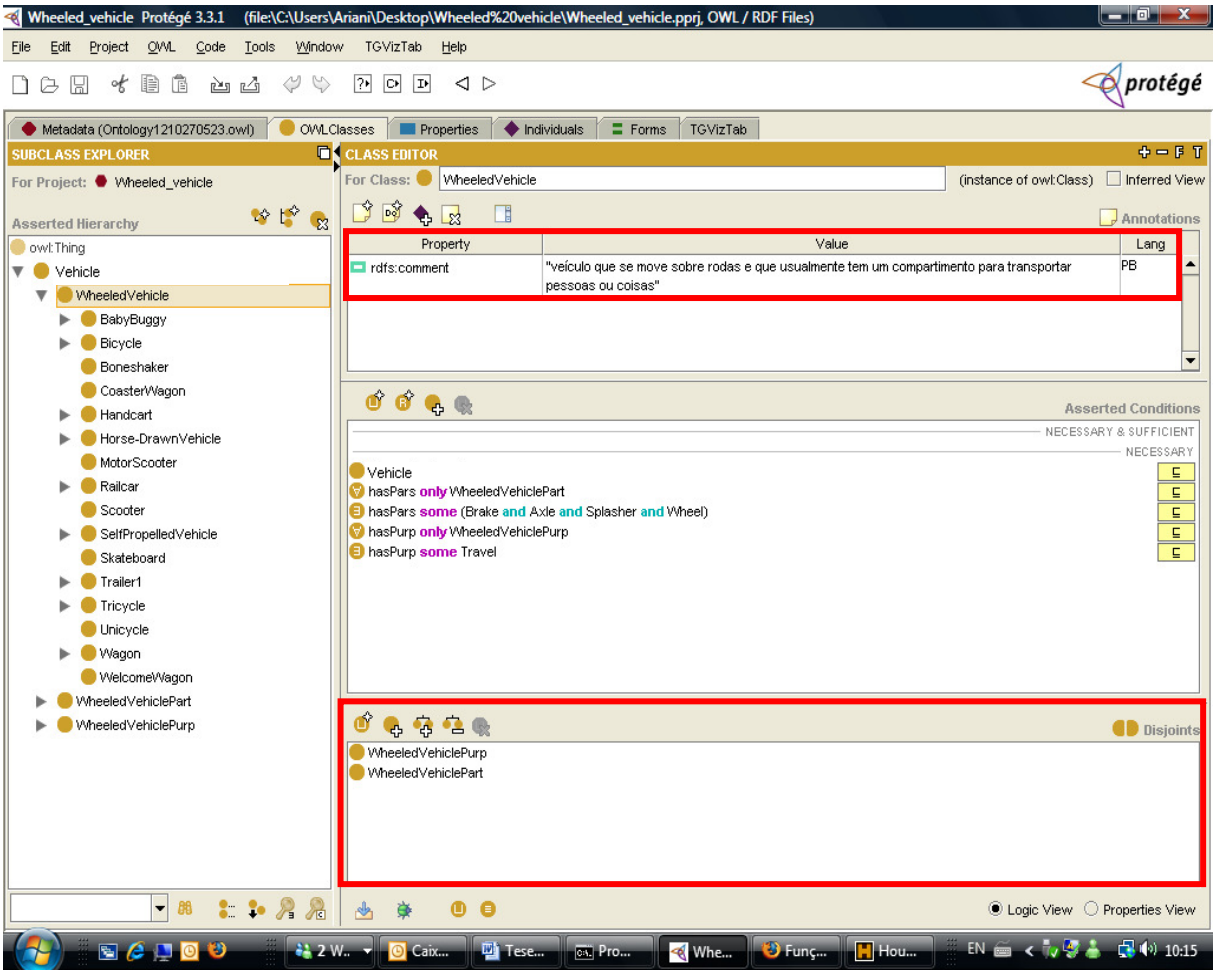


Figura 42: A inserção das glosas como comentários de classes.

7.2.1.2. As relações como propriedades

a) **Os tipos de propriedades**

Sabe-se que as `classes` pemitem a especificação de informação conceitual robusta. Seguindo os pressupostos do MultiNet, é necessário especifica as relações estabelecidas entre os conceitos, seus atributos multidimensionais, seu tipo conceitual e os traços semânticos para que se tenha uma descrição robusta dos conceitos.

O Protégé-OWL fornece dois tipos de recursos, denominados **propriedades,** para a descrição das informações associadas a uma classe. A linguagem OWL possui dois tipos principais de propriedades, `Object Properties` e `Datatype Properties`. O primeiro deles pode ser empregado para representar relações entre classes ou entre indivíduos. No geral, tanto as relações intraobjetivas quanto as interobjetivas podem ser descritas como `Object Properties`. O segundo tipo de propriedade, `Datatype Properties`, relaciona, de um modo

geral, objetos a vários tipos de dados, os quais não são representados pelas `Object Properties` (ANTONIOU, HARMELEN, 2004).

Neste trabalho, as relações PURP e PARS foram inseridas como `Object Properties`. Embora não haja uma convenção para a nomeação das propriedades, recomenda-se que estas sejam grafadas da seguinte forma: hasPars e hasPurp. As demais informações, ou seja, o tipo conceitual, os traços semânticos e os atributos multidimensionais, foram inseridas no editor Protégé-OWL como `Datatype Properties` e nomeadas, respectivamente, como: Tipo, Traços, AttrMult_GENER, AttrMult_FACT, AttrMult_VARIA, AttrMult_REFER e AttrMult_ETYPE.

As relações inseridas como `Object Properties` e as demais informações inseridas como `Datatype Properties` podem ser visualizadas na Figura 43, em que consta a interface do Protégé-OWL está no modo de visualização das propriedades (do inglês, *Properties View*) (círculo vermelho).

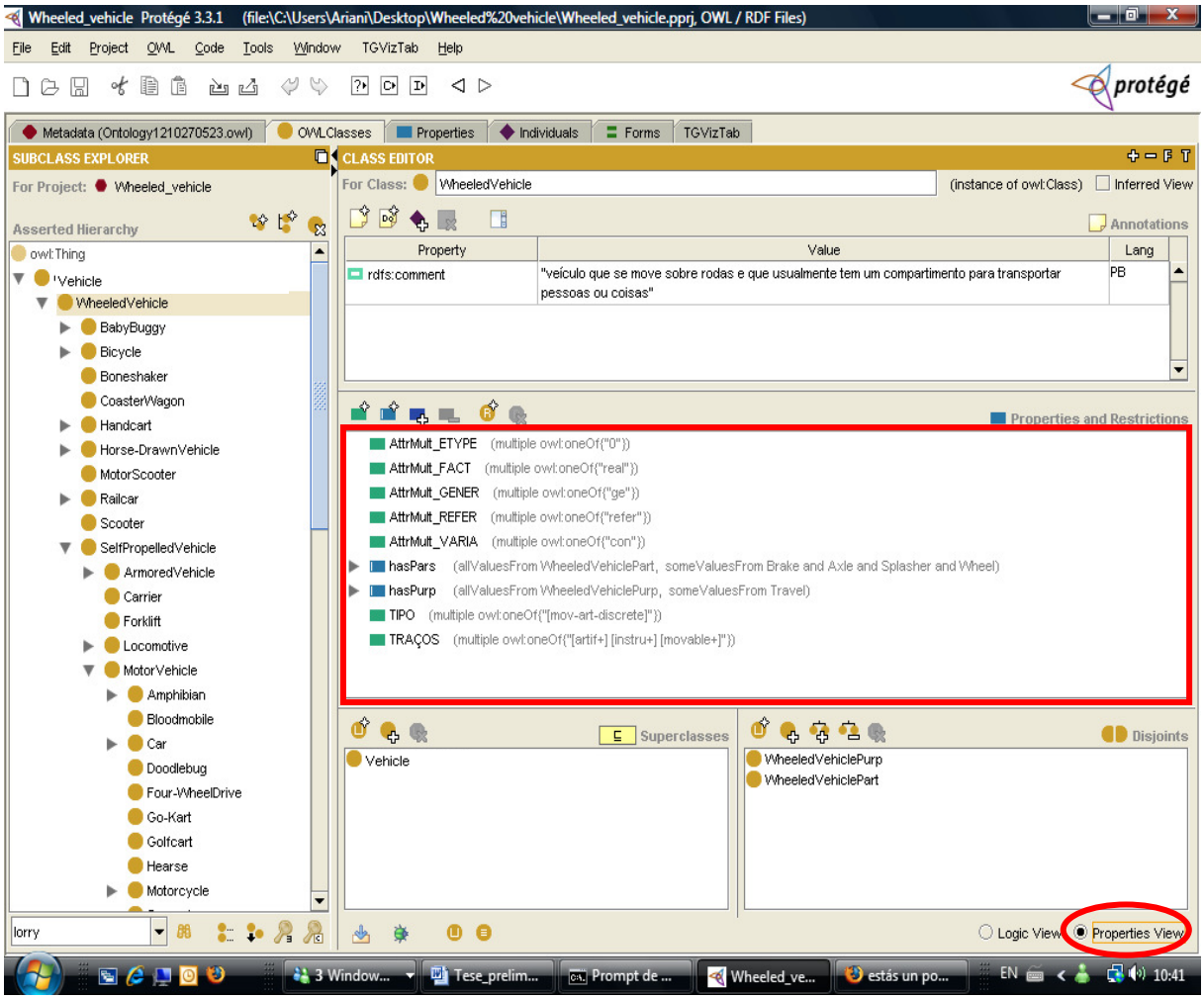


Figura 43: A inserção das relações como **`Object Properties`** e do tipo conceitual, traços semânticos e atributos multidimensionais como **`Datatype Properties`**.

b) **As restrições das propriedades**

Como as relações PURP e PARS foram inseridas como propriedades, os conceitos relacionados aos conceitos do tipo WheeledVehicle por meio dessas relações foram inseridos como **restrições** das propriedades (do inglês, *property restrictions*) hasPars e hasPurp, que são as representações concretas no editor das relações PARS e PURP do MultiNet. Dois tipos específicos de restrições foram utilizados: `allValuesFrom` (ou ∀) e `someValueFrom` (ou ∃). A restrição `allValuesFrom` é usada para especificar a classe de possíveis valores que uma `Object Property` pode ter. Em outras palavras, isso quer dizer que todos os valores de uma dada `Object Property` provêm de uma determinada classe. No caso, a relação hasPars, por exemplo, tem como valores apenas conceitos provenientes do superconceito WheeledVehiclePart, assim como a relação hasPurp tem como valores apenas conceitos do tipo WheeledVehiclePurp (Figura 44).

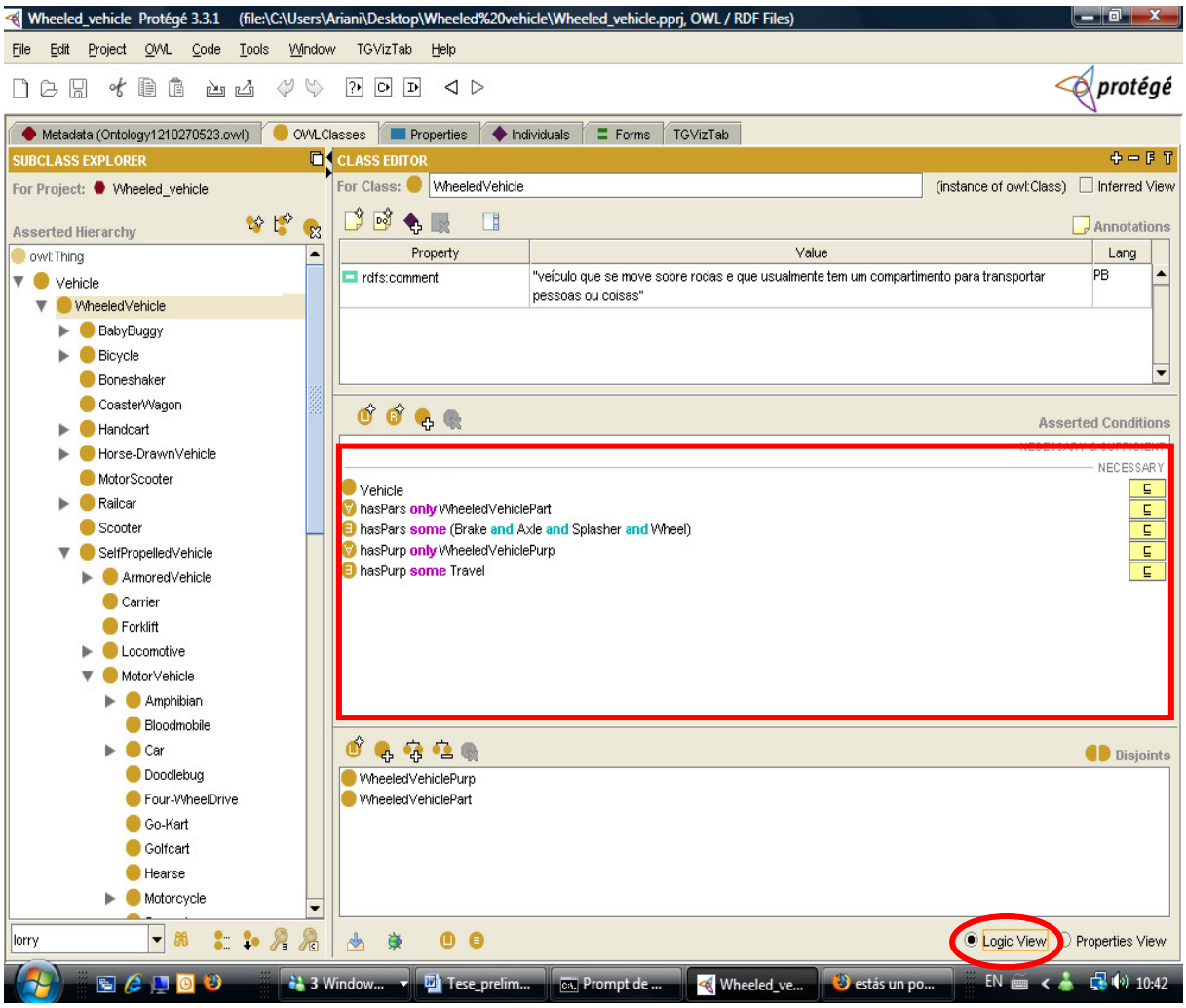


Figura 44: As relações e os conceitos por elas relacionados.

O outro tipo de restrição, `someValuesFrom`, foi utilizado para indicar, por exemplo, que ao menos uma das relações hasPars de WheeledVehicle deve apontar para os conceitos Brake, Axle, Splasher e Wheel. Os dois tipos de restrições são ilustrados na Figura 44, que mostra a interface do Protégé-OWL no módulo de visualização lógica (do inglês, *Logical View*) (círculo vermelho).

Em outras palavras, tais restrições podem ser entendidas da seguinte forma, considerando-se o conceito WheeledVehicle e a relação hasPars: (a) `allValuesFrom`: o conceito WheeledVehicle somente se relaciona por meio de hasPars com conceitos que sejam subconceitos de WheeledVehiclePart; (b) `someValuesFrom`: o conceito WheeledVehicle possui pelo menos uma relação hasPars com o conceito Brake, uma com o conceito Axle, uma outra com o conceito Splasher e uma com o conceito Wheel.

c) **A herança das informações no editor**

Vale ressaltar que as informações registradas em termos das propriedades do tipo `Datatype Properties` foram especificadas apenas uma vez e herdadas por todos os conceitos da hierarquia. As relações conceituais inseridas como propriedades do tipo `Object Properties`, quando especificadas para um superconceito, também foram herdadas por todos os conceitos mais específicos da hierarquia. Esse mecanismo de herança pode ser observado na Figura 45, em que o conceito representado pelo conceito BabyBuggy herda as informações do conceito mais geral representado pelo superconceito WheeledVehicle.

Nessa Figura, a herança das informações é assinalada de duas formas: pelo tom mais claro da cor da fonte e principalmente pela mensagem explícita de herança (no inglês, *Inherited)* localizada no lado direito do retângulo vermelho.

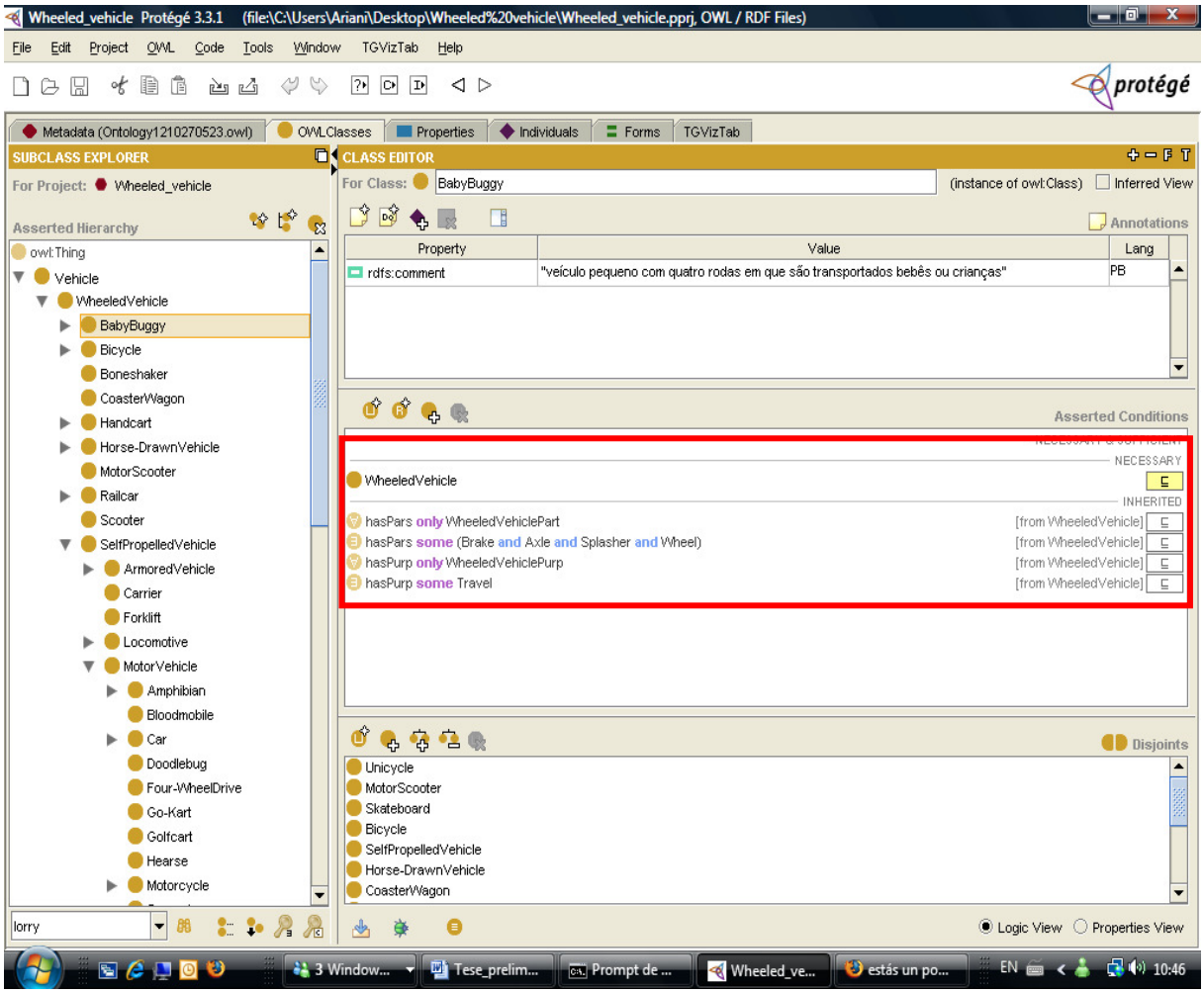


Figura 45: A herança de informações.

d) **A especificação do domínio e contradomínio**

Ao especificar uma relação *x*, é possível determinar o seu o domínio e contradomínio. No caso, o superconceito WheeledVehicle é o domínio de ambas as relações, hasPars e hasPurp. Tais relações diferem, no entanto, quanto ao contradomínio. Enquanto o contradomínio de hasPars é o conceito WheeledVehiclePart, o contradomínio de hasPurp é WheeledVehiclePurp.

Como exemplificação, a relação hasPars e seus respectivos domínio e contradomínio estão assinalados com retângulos vermelhos na Figura 46.

A Figura 46 mostra a GUI do editor em que as relações, inseridas como `propriedades`, são gerenciadas.

Mais especificamente, a aba selecionada, denominada `Properties` e destacada pelo círculo vermelho, é responsável pela abertura de uma janela composta por dois módulos. No módulo à esquerda, observa-se o elenco de relações especificado, do qual hasPars (em

destaque) faz parte. No módulo à direita, denominado `Property Editor`, as relações podem ser propriamente editadas. Nesse módulo, o retângulo vermelho superior salienta a definição estabelecida para cada propriedade, seja ela `Object Property` ou `Datatype Property`. Essa definição foi introduzida como o valor da propriedade `rdfs:comment` vinculada a cada propriedade.

Na Figura 46, a definição em destaque é a da relação hasPars: "relação conceitual entre *x* e *y*, tal que *x* tem como parte *y*". Além da própria definição, especificou-se a língua na qual a definição foi elaborada (no caso, PB); essa especificação localiza-se na seqüência da definição. Ainda no módulo `Property Editor`, os dois retângulos mais abaixo indicam, por sua vez, o domínio e contradomínio da relação que está selecionada (hasPars).

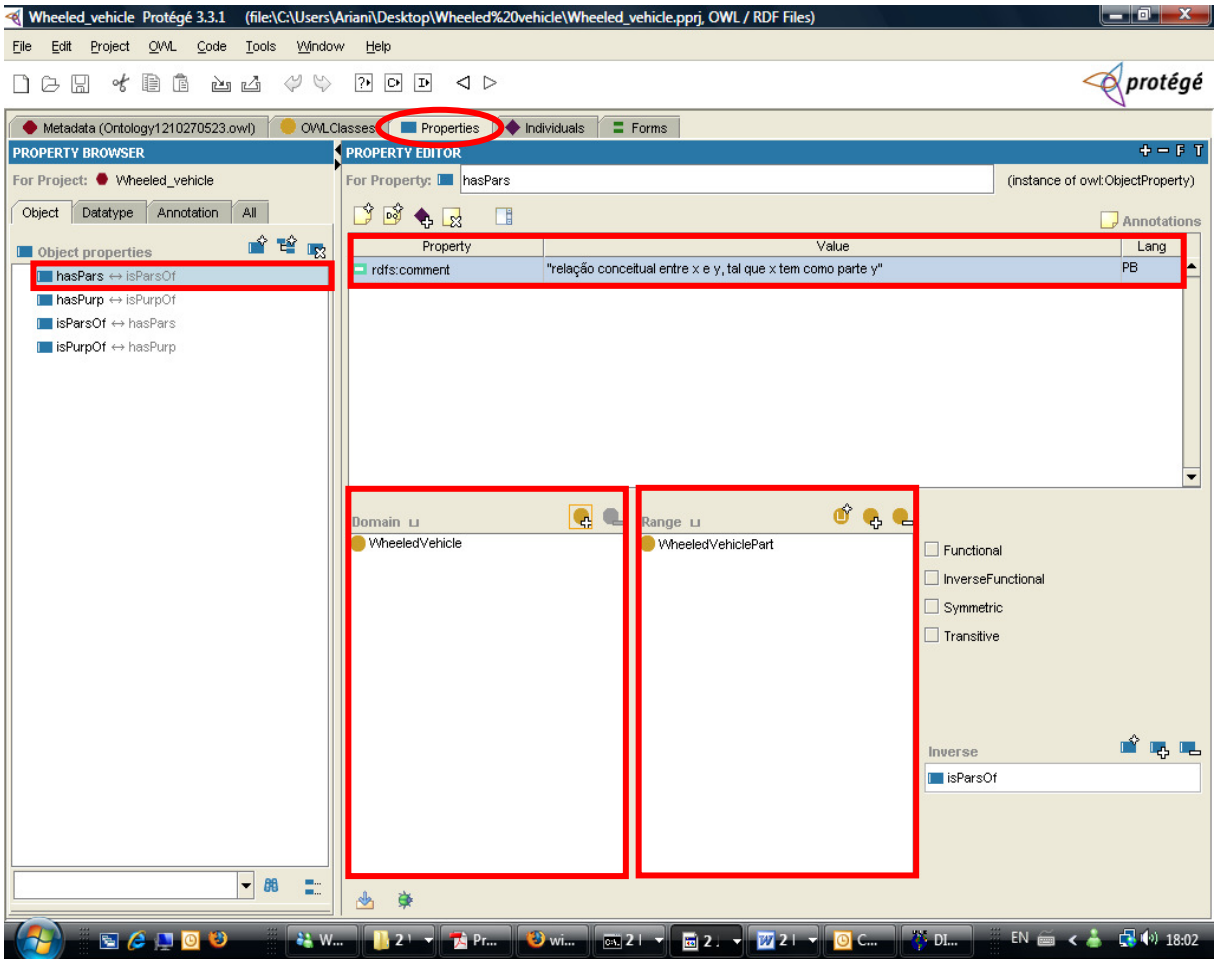


Figura 46: Gerenciamento das relações como propriedades.

e) **A especificação das relações inversas**

Apesar de não constarem nos *templates* léxico-conceituais, as relações denominadas **inversas** (do inglês, *inverse properties*) também foram especificadas, posto que o editor Protégé-OWL

oferece esse recurso (ANTONIOU, HARMELEN, 2004). Assim, dada uma propriedade, esta pode ter uma propriedade inversamente correspondente. Isso quer dizer que, se uma relação relaciona um conceito *x* a um conceito *y*, então sua relação inversa relaciona *y* a *x* (HORRIDGE et al., 2004).

A Figura 47 mostra que, neste trabalho, as relações hasPars e hasPurp relacionam-se às suas relações inversas, as quais foram representadas, respectivamente, pelas propriedades: isParsOf e isPurpOf. A relação representada por isParsOf é responsável por relacionar os conceitos do tipo WheeledVehiclePart aos conceitos do tipo WheeledVehicle (Figura 47). A relação inversa isPurpOf, por sua vez, é responsável por relacionar os conceitos do tipo WheeledVehiclePurp aos conceitos do tipo WheeledVehicle.

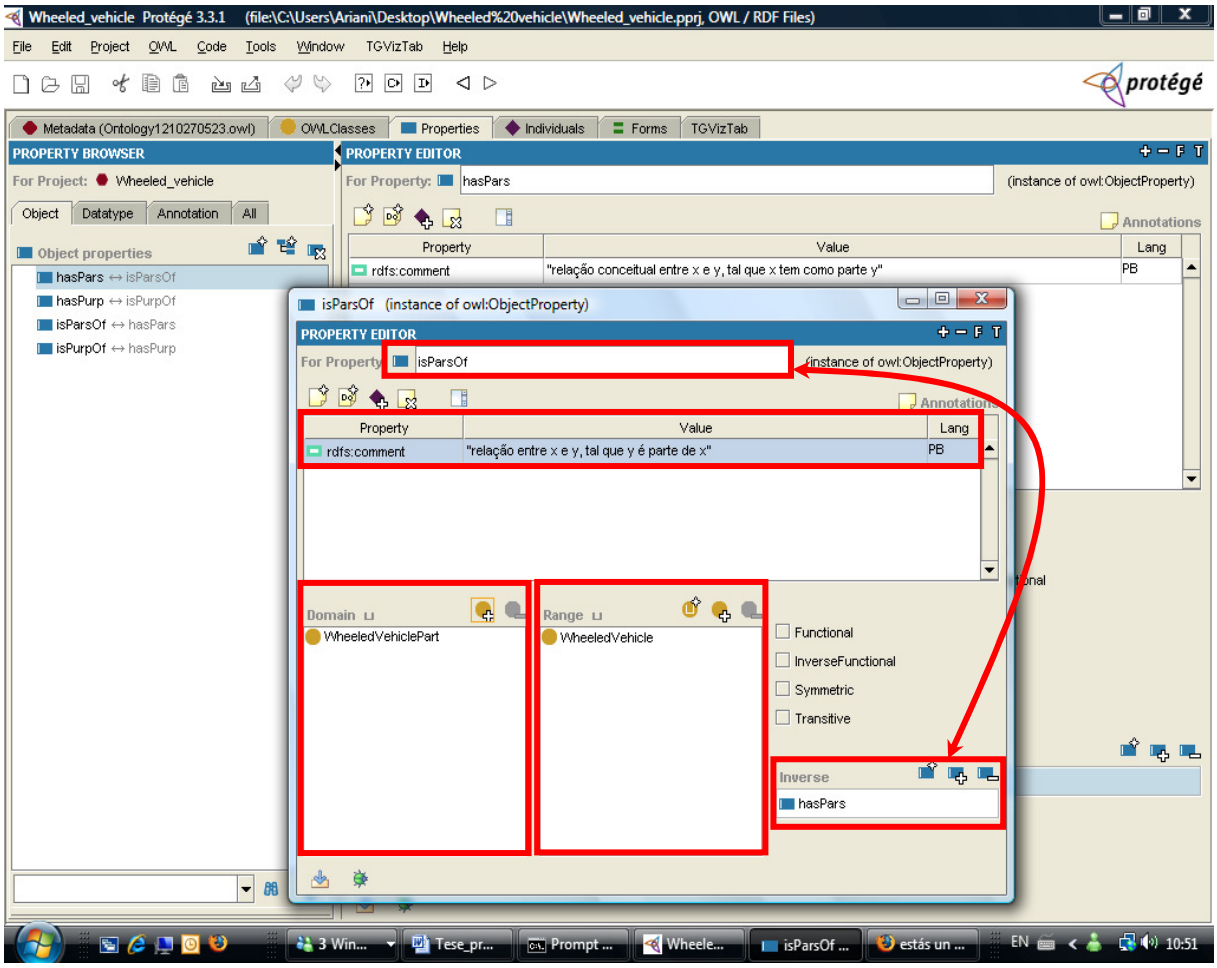


Figura 47: A especificação das relações inversas.

Dessa forma, o domínio e contradomínio de uma relação inversa são inversos ao domínio e contradomínio da relação de origem. A relação isParsOf, por exemplo, tem como domínio WheeledVehiclePart e como contradomínio WheeledVehicle, enquanto que a relação hasPars tem WheeledVehicle como domínio e WheeledVehiclePart como contradomínio.

A Figura 47 mostra a GUI (mais especificamente, o *menu pop-up*[78]) em que as relações inversas são criadas. Nessa Figura, ilustra-se, especificamente, a criação da relação isParsOf, inversa a hasPars, e a especificação do domínio e contradomínio da relação inversa. Além disso, é possível observar que a relação inversa também está associada a uma definição; no caso, isParsOf é definida como: "relação entre x e y, tal que y é parte de x".

7.2.1.3. As expressões lingüísticas como instâncias

Como mencionado na Subseção 7.2.1.1 (pág. 159), a hierarquia conceitual no Protégé-OWL é representada por classes, as quais são definidas como conjuntos de **indivíduos** (do inglês, *individuals*). Por exemplo, em uma ontologia para o domínio dos vinhos, MountandamPinotNoir seria uma instância da classe PinotNoir (HORRIDGE et al. 2004; SMITH et al, 2004). Neste trabalho, as instâncias foram adaptadas para representar as expressões lingüísticas dos conceitos. Assim, pode-se dizer que uma unidade lexical (ou um SLR) é uma instância de um determinado conceito. A Figura 48 mostra a GUI do Protégé-OWL responsável pela manipulação das expressões lingüísticas (unidades lexicais e SLRs).

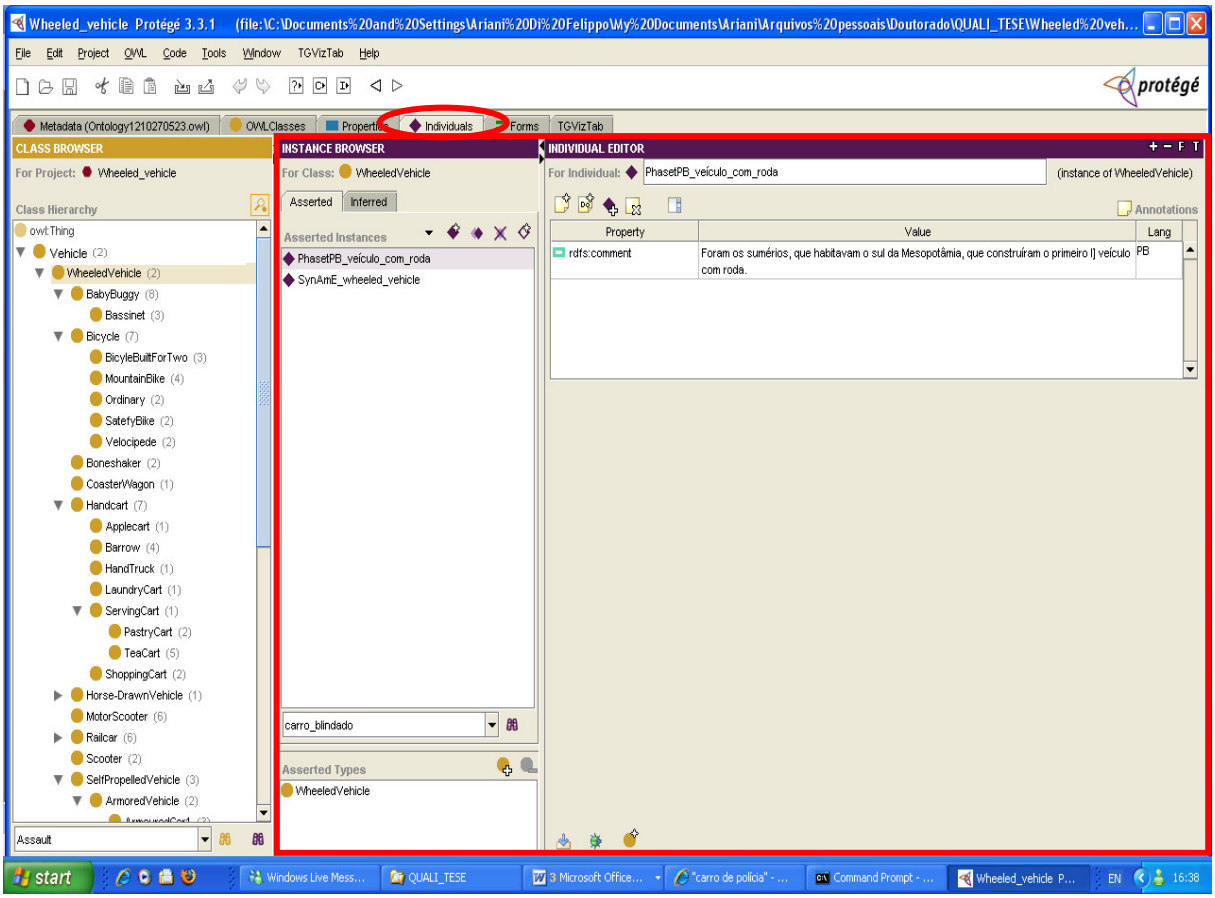


Figura 48: A inserção das expressões lingüísticas como instâncias.

[78] O *pop-up* é uma janela extra que se abre mediante certo comando.

Mais especificamente, essa interface é aberta quando a aba `Individuals` é selecionada (círculo vermelho). Essa interface é uma janela composta por dois módulos, os quais estão indicados pelo retângulo vermelho. No módulo denominado `Instance Browser` (à esquerda do retângulo vermelho), visualiza-se o elenco de instâncias vinculado a um determinado conceito. No módulo mais à direita, denominado `Individual Editor`, pode-se criar e/ou manipular as instâncias que aparecem elencadas no módulo `Instance Browser`. No exemplo da Figura 48, o conceito em destaque é WheeledVehicle. Esse conceito possui duas expressões lingüísticas: *Phraset*`PB_veículo_com_ rodas` e `SynAmE_wheeled vehicle`.

Além dos dois módulos diretamente relacionados à criação das instâncias, a Figura 48 mostra que a interface apresenta outro módulo, à esquerda, denominado `Class Browser`. Esse módulo exibe a hierarquia de conceitos. Vale ressaltar que, quando a aba `Individuals` é selecionada, os conceitos automaticamente aparecem associados a um número entre parênteses. Esse número indica a quantidade de expressões lingüísticas associadas ao conceito. Por exemplo, o conceito representado pela classe WheeledVehicle associa-se a duas expressões lingüísticas, ou melhor, há duas expressões lingüísticas que expressam esse conceito (*veículo com rodas* e *wheeled vehicle*) (cf. Figura 48). O módulo `Individual Editor`, em destaque na Figura 49, é composto por dois campos, indicados por meio dos retângulos vermelhos.

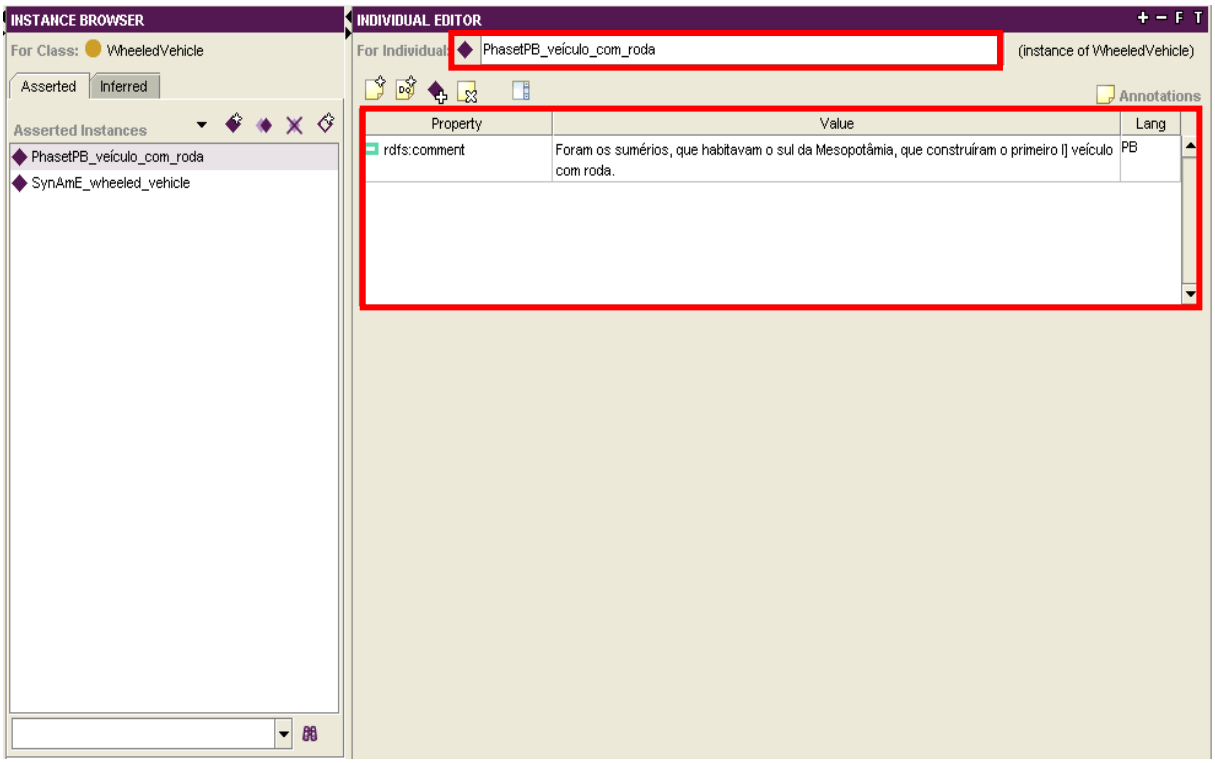


Figura 49: A especificação das expressões lingüísticas nos campos do módulo `Individual Editor`.

No campo superior, especificam-se as expressões lingüísticas do AmE e do PB por meio do emprego de letras minúsculas. Já para as unidades lexicais complexas e para os SLRs, a separação dos elementos constitutivos das unidades e dos sintagmas foi marcada pela inserção de sublinhados (do inglês, *underscores*), p.ex.: *veículo_com_rodas*. A utilização do sublinhado é uma imposição do editor.

Além disso, devido a certas restrições do editor, foi necessário inserir as expressões lingüísticas separadamente e não sob a forma de conjuntos de sinônimos (*synsets* ou *phrasets*). Para indicar a pertença de uma unidade lexical ou SLR a um conjunto de sinônimos, foram criadas quatro etiquetas, SynAmE, SynPB, *Phraset*AmE e *Phraset*PB. Essas etiquetas indicam, respectivamente: (i) que a unidade lexical em questão é elemento constitutivo de um *synset* do AmE; (ii) que a unidade lexical em questão é elemento constitutivo de um *synset* do PB; (iii) que o SLR em questão é elemento constitutivo de um *phraset* do AmE, e (iv) que o SLR em questão é elemento constitutivo de um *phraset* do PB. No exemplo da Figura 49, a expressão lingüística especificada foi nomeada como *PhrasetPB_veículo_com_rodas*.

Além dessas etiquetas, expressões lingüísticas homógrafas, assim como feito para os conceitos, receberam um indicador numérico para diferenciá-las, como *SynAmE_lorry1* (<caminhão grande destinado ao transporte de cargas pesadas; usualmente sem laterais>) e *SynAmE_lorry2* (<carroça grande e baixa sem laterais>).

No campo localizado imediatamente abaixo do superior, especifica-se a frase-exemplo relativa à expressão lingüística em questão. As frases-exemplo foram introduzidas como valores da propriedade `rdfs:comment` vinculada a cada expressão lingüística. Além da frase-exemplo, essa propriedade permite especificar também a língua do contexto de ocorrência da expressão sob análise. Tal especificação foi feita pelas siglas PB e AmE. No exemplo da Figura 49, a frase-exemplo relacionada a *veículo com rodas* é "Foram os sumérios, que habitavam o sul da Mesopotâmia, que construíram o primeiro I]veículo com rodas".

7.2.2. A visualização gráfica da interlíngua e das expressões lingüísticas de seus conceitos constitutivos

Dentre os vários *plug-ins* que podem ser acoplados ao editor Protégé-OWL, destaca-se, aqui, o 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.

Mais especificamente, o TGVizTab baseia-se na tecnologia denominada **TouchGraph**, que oferece vários recursos de visualização de uma rede conceitual, como alto grau de interação, rápida renderização[79], visão panorâmica e zum, entre outros[80].

A seguir, apresentam-se os principais recursos do TGVizTab aplicados na visualização da interlíngua estruturada. A Figura 50 mostra a GUI do Protégé-OWL acrescida da aba (círculo vermelho) que aciona o *plug-in* TGVizTab.

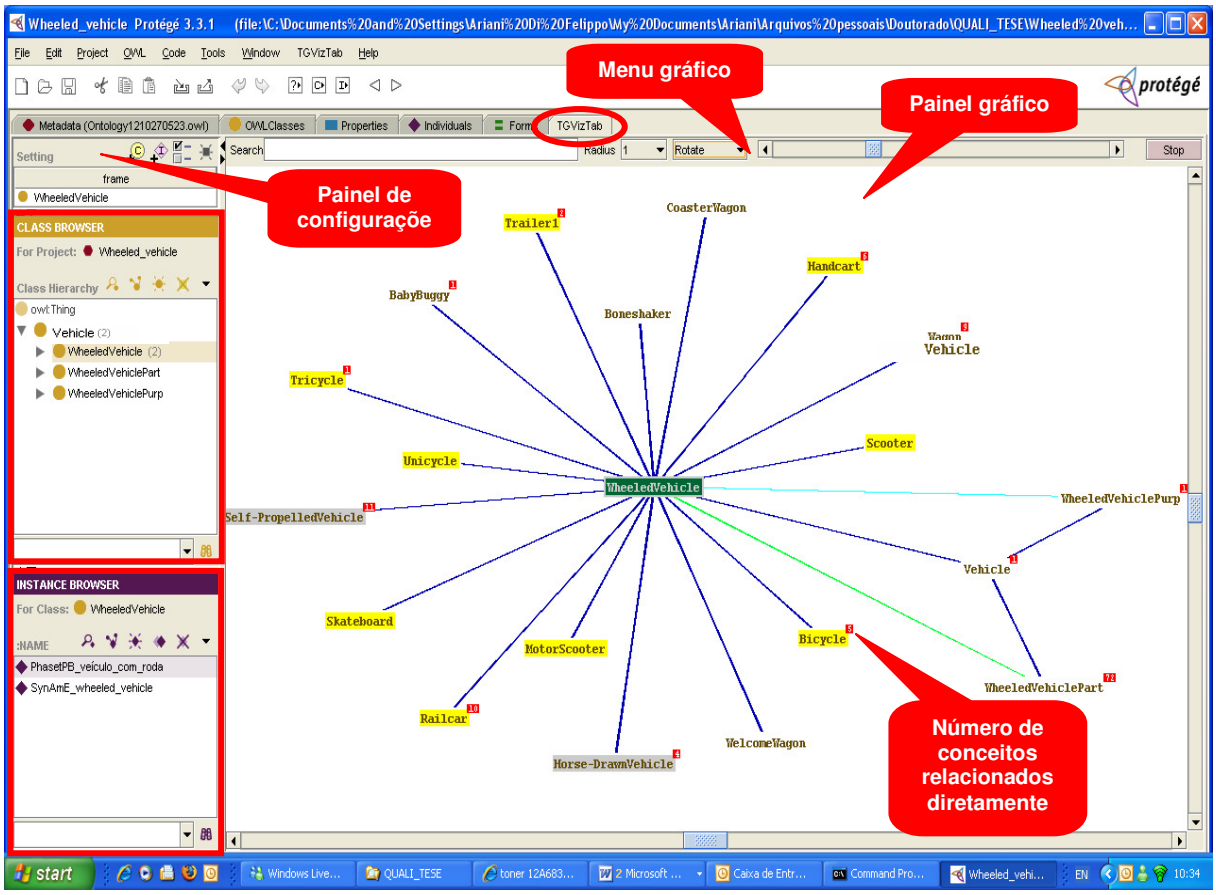


Figura 50: A *interface* do *plug-in* TGVizTab, exibindo o conceito WheeledVehicle no centro do grafo.

A hierarquia conceitual é exibida à esquerda, no campo denominado `Class Browser`, e em formato arbóreo (retângulo vermelho superior da Fig. 50). Quando um conceito é selecionado, a lista das expressões lingüísticas associadas a ele é mostrada no campo denominado

[79] 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).

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

`Instance Browser` (retângulo vermelho inferior da Fig. 50). Para gerar o grafo, um conceito ou uma expressão lingüística precisa ser selecionado para atuar como o nó central do grafo. Um conceito é selecionado no campo `Class Browser` e uma expressão lingüística é selecionada no campo `Instance Browser`, sendo que a adição de ambas é feita por botões especiais, localizados no Painel de Configurações (do inglês, *Settings Panel*). Um conceito é inserido em um grafo pelo botão e uma expressão lingüística pelo botão . A geração do grafo é feita pela seleção do botão , também localizado no Painel de Configurações. Quando esse botão é clicado, o grafo correspondente é gerado e exibido na área denominada Painel Gráfico (do inglês, *Graph Panel*). Na Figura 50, selecionou-se o conceito representado pela classe WheeledVehicle, que é colocado como o nó central do grafo exibido no Painel Gráfico.

O Painel de Configurações contém ainda um *menu*, indicado pelo botão , para selecionar cores, fontes, entre outras informações a serem exibidas no grafo. Ao se clicar no referido botão, uma janela de configuração, composta por várias abas, é aberta. A Figura 51 mostra essa janela quando as abas `Slots` e `Classes` são selecionadas. A aba `Slots` fornece recursos, por exemplo, para especificar quais relações devem ser exibidas e com quais cores. No caso, especificou-se que as relações hasPars e hasPurp devem ser exibidas no grafo, sendo que os seus respectivos arcos devem ter as cores verde e azul. A aba `Classes`, por sua vez, fornece o recurso de especificar qual a cor de fundo dos nós de um grafo. A Figura 51 exibe a especificação da cor amarela para certos nós.

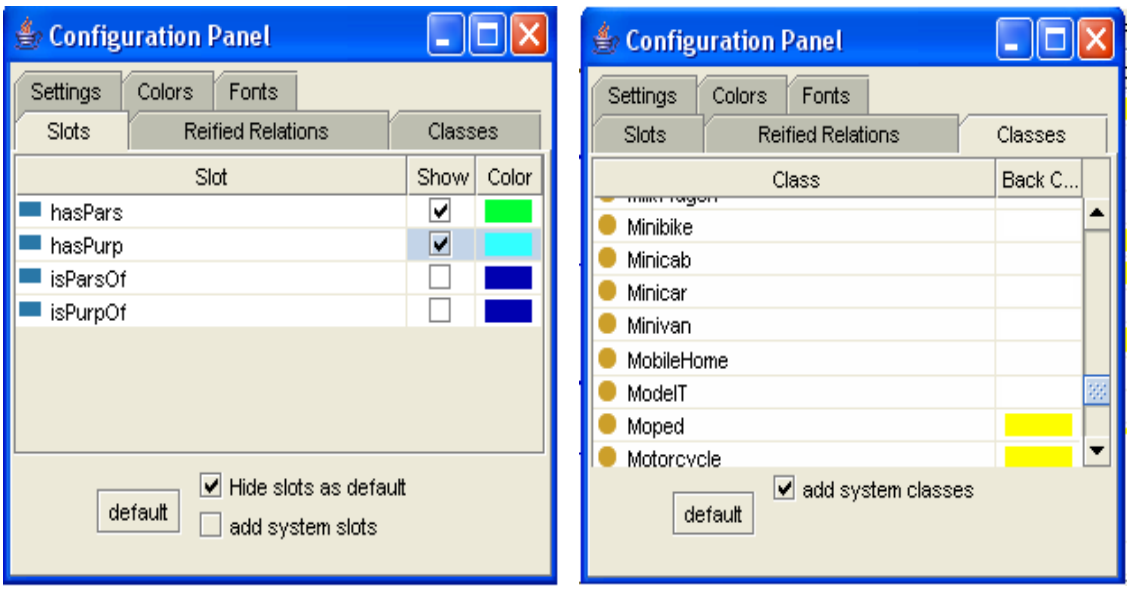


Figura 51: Painel de Configurações do TGVizTab: abas `Slots` e `Classes`.

Vinculado ao Painel Gráfico, encontra-se o *Menu* Gráfico, que oferece vários recursos de controle, como zum, busca por nós, níveis de relacionamento e controle de localidade.

Para a geração do grafo da Figura 52, especificou-se no *Menu* Gráfico que apenas os conceitos diretamente relacionados a WheeledVehicle devem ser exibidos, ou seja, apenas os conceitos de nível 1. Devido à seleção desse nível, o número total dos conceitos, pertencentes ao próximo nível, fica assinalado ao lado dos nós por meio de um índice numérico. No grafo da Figura 52, o nó Handcart está associado ao índice numérico 6, indicando que Handcart está diretamente relacionado a 6 conceitos. Se o nível 2 fosse especificado ao gerar o grafo em que WheeledVehicle é o nó central, esses 6 conceitos também seriam exibidos.

Quanto aos grafos gerados pelo TGVizTab, salienta-se que os conceitos e as expressões lingüísticas são exibidas como nós, os quais podem ser especificados por diferentes cores. As relações, por sua vez, são exibidas como arcos rotulados entre os nós, sendo que os arcos também podem ter diferentes cores. A cor dos nós e dos arcos, como mencionado anteriormente, é definida no Painel de Configurações (cf. Figura 51). Os nós e arcos estão destacados na Figura 52.

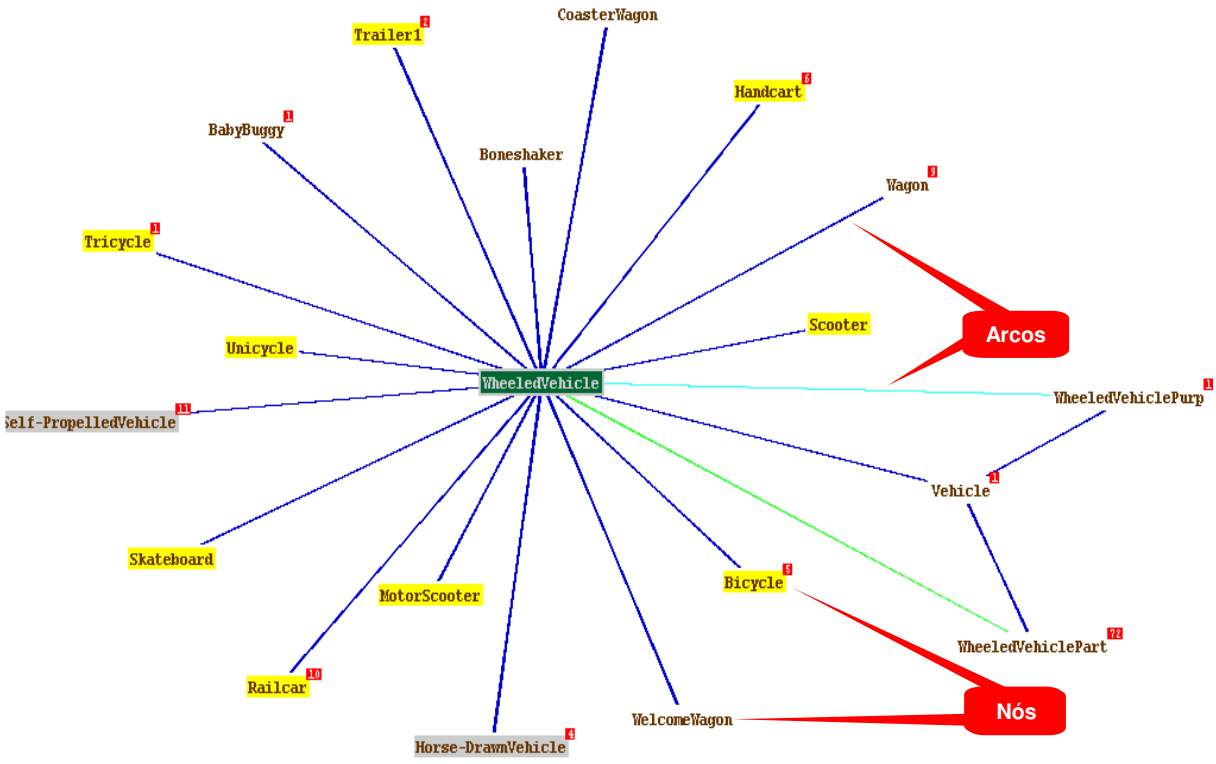


Figura 52: Exemplo do grafo gerado pelo TGVizTab.

Quanto aos rótulos dos arcos, vale ressaltar que, para evitar confusão ou desordem visual, eles são exibidos apenas quando o cursor é posicionado sobre um arco ou sobre o nó de

origem do arco, como ilustrado pela Figura 53. Para gerar a imagem contida na Figura 53, o cursor deve ser posicionado sobre o nó WheeledVehicle.

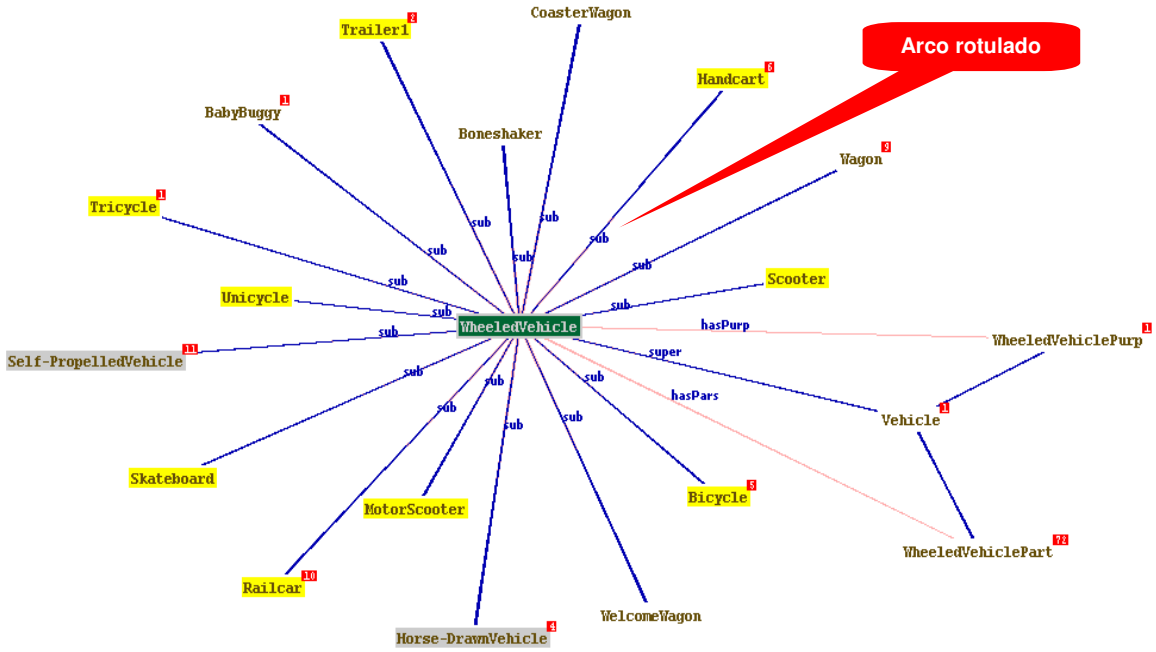


Figura 53: A exibição dos rótulos dos arcos do grafo.

Por fim, ressalta-se que o clique com o botão direito do *mouse* sobre um nó abre o *Menu* do Nó (do inglês, *Node Menu*). Esse *menu* fornece opções para esconder, expandir, reduzir os nós e ver a descrição completa do conceito representado no nó (Figura 54).

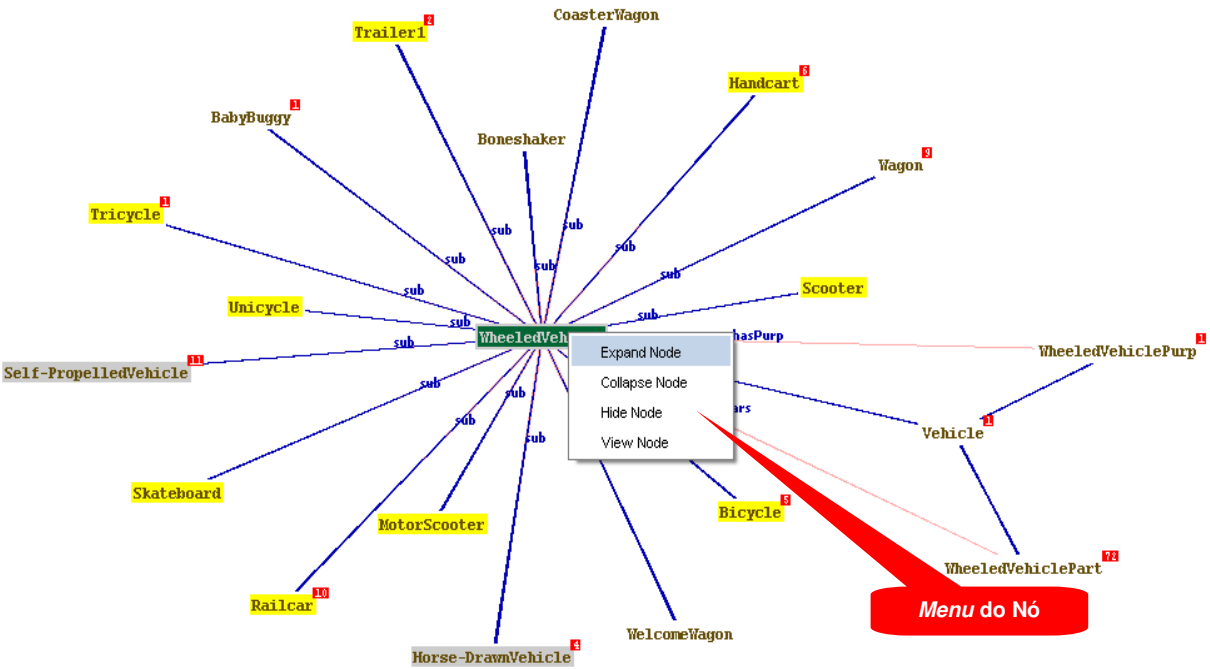


Figura 54: *Menu* do TGVizTab que fornece opções de controle para os nós do grafo.

Dessa forma, observa-se que o TGVizTab fornece recursos (Painel de Configuração) para que o usuário possa controlar a cor e também a visibilidade de cada nó e arco. Tais recursos permitem identificar claramente diferentes tipos de informação no grafo, o que auxilia a tarefa de análise visual da ontologia.

No caso deste trabalho, cujo tema é a lexicalização de conceitos, os recursos do TGVizTag foram especificamente utilizados para evidenciar as diferenças de lexicalização entre o AmE e o PB. Mais especificamente, utilizou-se o recurso de especificar a cor de fundo dos nós para destacar os que graficamente representam conceitos lexicalizados no PB. Para os conceitos lexicalizados no PB, em especial, foi utilizada a cor de fundo amarela. Os nós com cor de fundo cinza são aqueles que representam os conceitos não-lexicalizados no AmE. Assim, por exclusão, os nós sem cor são os conceitos não-lexicalizados no PB. Essa distinção foi fundamental para a tarefa de especificação das relações interlinguais descrita em 7.4.3.2. Com base nessa distinção, é possível obter uma visão panorâmica (geral) da expressão lexical dos conceitos pertencentes ao domínio dos "veículos com rodas" no AmE e no PB. Essa visão geral fica evidente quando o grafo é gerado com todos os nós expandidos, ou seja, considerando-se todos os níveis das relações. A Figura 55 ilustra parte dessa visão panorâmica. Diz-se "parte" porque a visualização completa do grafo só é possível com o auxílio das barras de rolagens (cf. Figura 55).

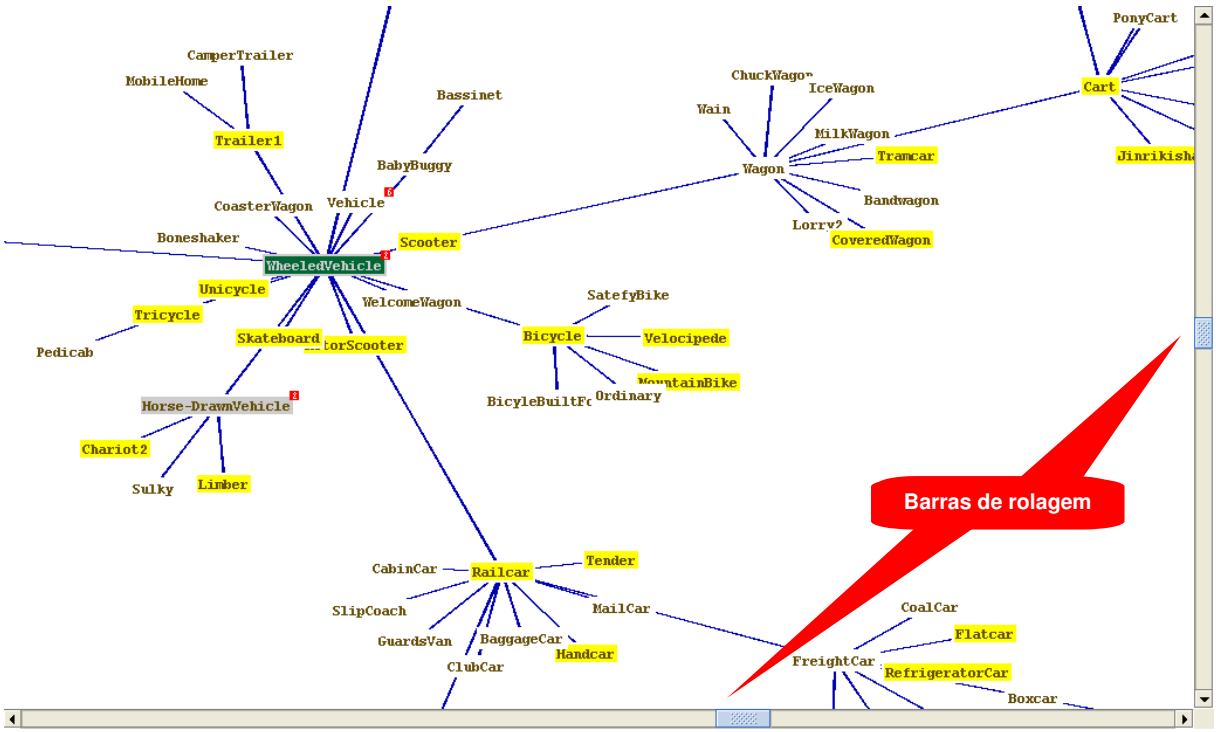


Figura 55: Distinção gráfica que representa os conceitos lexicalizados e os não-lexicalizados no PB.

7.3. A base léxico-conceitual bilíngüe REBECA

Por meio do editor de ontologias Protégé-OWL, foi possível construir uma base léxico-conceitual bilíngüe para o par de línguas AmE-PB. Essa base, que passa a ser aqui denominada REBECA[81], é caracterizada principalmente por: (i) incluir os conceitos pertencentes ao domínio dos "veículos com rodas"; (ii) utilizar uma interlíngua conceitual estruturada e formal; (iii) armazenar conceitos lexicalizados e (vi) ter como base teórica o paradigma de RC MultiNet.

Vale ressaltar aqui que, nessa base, os conceitos extraídos da WN.Pr, cujas lexicalizações no PB foram investigadas, são ao mesmo tempo: (i) os conceitos lexicalizados no AmE pertencentes ao domínio em questão, ou seja, a parcela do léxico do AmE que recobre o domínio dos "veículos com rodas", e (ii) os elementos constitutivos da interlíngua semântica.

Desempenhando papel central nessa base, está a interlíngua, que funciona como elo conceitual entre as referidas línguas. Essa interlíngua é composta por 217 conceitos expressos no AmE, os quais foram extraídos da WN.Pr. Desses 217, 205 são lexicalizados e 12 não são lexicalizados. Por incluir os conceitos expressos no AmE, o conhecimento léxico-conceitual nela armazenado é, assim, lingüisticamente motivado por essa língua. Os conceitos da interlíngua estão representados em função do paradigma de RC MultiNet, que se fundamenta em uma concepção cognitiva de significado e na metalinguagem formal das redes semânticas. Dessa forma, essa interlíngua é entendida como uma ontologia, já que é "uma especificação formal de uma conceitualização". Como os conceitos dessa ontologia estão organizados em função de certas relações que se estabelecem entre eles, principalmente a de hiperonímia/ hiponímia (ou SUB), diz-se que a interlíngua é, portanto, do tipo estruturado.

Do ponto de vista estrutural, a ontologia da base bilíngüe REBECA distingue, de certa forma, dois "níveis" conceituais distintos, denominados aqui de "tipos conceituais" (ou metaconceitos) e "conceitos básicos" (cf. Figura 56).

Os tipos conceituais, especificados entre os símbolos [ ] na Figura 56, foram fornecidos diretamente pelo MultiNet e são os tipos conceituais mais abstratos. Esses tipos não aparecem diretamente na hierarquia conceitual inserida no Protégé-OWL, mas sim como o valor complexo [mov-art-discrete] da `DatatypeProperty` denominada TIPO, a qual está associada a cada um dos conceitos lexicalizados.

[81] REBECA, do hebraico *Ribqah*, significa "aquela que une, liga". No Antigo Testamento, a personagem Rebeca era filha de Betuel, irmã de Labão, mulher de Isaque e mãe de Jacó e Esaú. REBECA é também o nome da base de dados léxico-conceitual bilíngüe resultante deste trabalho. 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 AmE a uma do PB.

Os conceitos básicos englobam os conceitos lexicalizados analisados neste trabalho e que constam na hierarquia conceitual inserida no editor (cf. Figura 41, pág. 60), juntamente com o conceito <vehicle>, inserido para facilitar a organização conceitual. Na Figura 56, os conceitos básicos estão marcados com os símbolos < >.

Vale ressaltar que os conceitos da interlíngua estão relacionados a outros por meio das relações PARS e PURP. Estes, no entanto, não fazem parte diretamente da interlíngua, mas sim participam da especificação dos conceitos constitutivos da mesma.

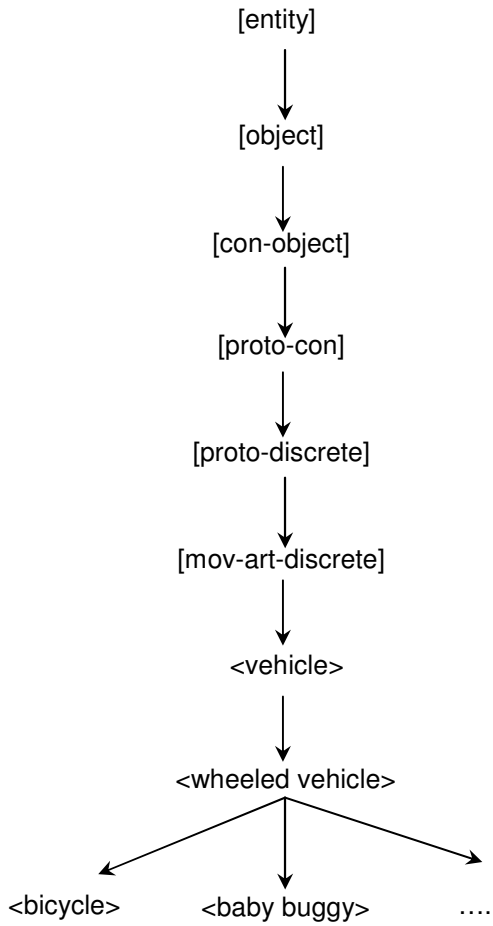


Figura 56: Estrutura ontológica da base REBECA.

Na base REBECA, os *synsets* do AmE e os do PB que lexicalizam o mesmo conceito estão vinculados a um mesmo elemento da interlíngua (ou conceito básico). Em alguns casos de lacuna lexical, pode-se ter um *phraset* do PB vinculado ao conceito básico da ontologia. Quando um conceito *x*, não lexicalizado no PB, não possui nem mesmo um SLR correspondente, é possível percorrer a estrutura da ontologia e identificar, por exemplo, qual o conceito mais geral em relação a *x* que fora lexicalizado no PB. O fornecimento de SLRs e lexicalizações de conceitos mais genéricos é estratégia relevante para se lidar com as lacunas

lexicais no âmbito do processamento automático das línguas naturais. A identificação de expressões alternativas para conceitos não-lexicalizados é ilustrada na Figura 57. Nessa Figura, ilustra-se que, para os conceitos não-lexicalizados <dogcart> e <dumpcart>, é possível identificar que o conceito mais geral <cart> é lexicalizado no PB ao percorrer a interlíngua ou ontologia, o que fornece a unidade lexical *carroça* como expressão alternativa para as lacunas do PB.

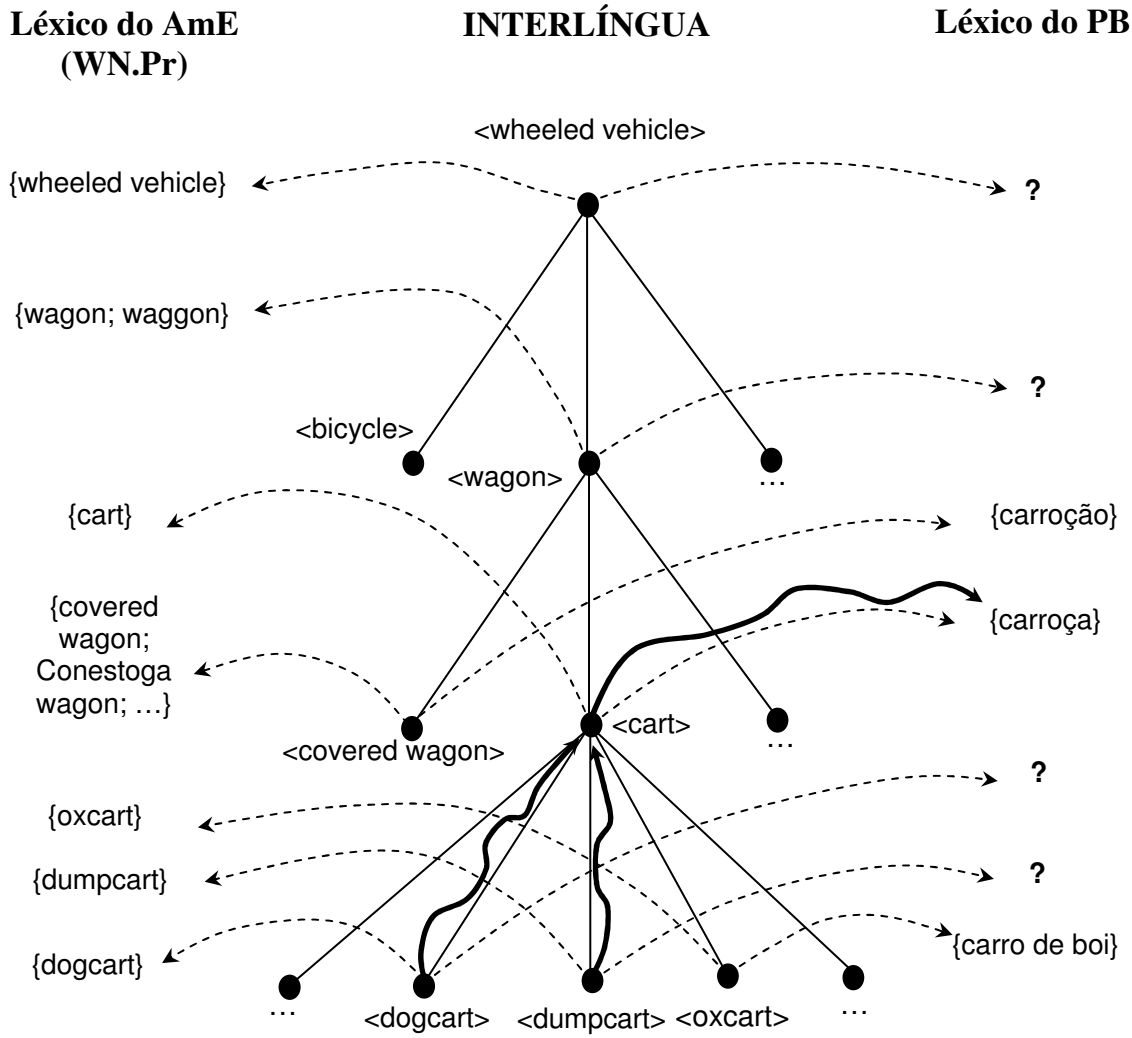


Figura 57: Um exemplo de identificação de expressões alternativas para conceitos não-lexicalizados.

Por fim, salienta-se que o editor Protégé-OWL pode exportar a base REBECA em vários formatos, como o próprio OWL, Turtle, etc. Como exemplificação, apresenta-se a seguir um pequeno fragmento da base REBECA, na linguagem OWL. Nesse fragmento, estão presentes os dois componentes iniciais de um arquivo em OWL, o conjunto de XML *namespace* e o cabeçalho, além da codificação do conceito <dogcart> como uma subclasse de <cart>

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns="http://www.owl-ontologies.com/Ontology1210270523.owl#"
  xml:base="http://www.owl-ontologies.com/Ontology1210270523.owl">
  <owl:Ontology rdf:about="">
    <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >version 1</owl:versionInfo>
    <rdfs:comment xml:lang="AmE">An ontology of lexicalized concepts from the wheeled
vehicles domain and their linguistic expressions in American English and Brazilian
Portuguese.</rdfs:comment>
  </owl:Ontology>
</owl:Class>
  <owl:Class rdf:ID="Dogcart">
    <rdfs:comment xml:lang="PB">"carroça pequena puxada por um
cachorro"</rdfs:comment>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Cart"/>
    </rdfs:subClassOf>
</owl:Class>
```

Figura 58: A base REBECA no formato OWL.

Nesse fragmento, é possível observar que o arquivo no formato OWL é iniciado por uma série de declarações iniciadas pela etiqueta `xml`, denominada XML *namespace*. Tais declarações nada mais são do que endereços eletrônicos dos desenvolvedores do formato OWL, o grupo W3C, e indicam o formato específico em que os dados se encontram codificados. Dentre essas declarações, há uma responsável por especificar que um arquivo em formato OWL é também um arquivo RDF, linguagem formal que deu origem à OWL. Na seqüência, vê-se uma espécie de cabeçalho, responsável por fornecer informações gerais sobre o ontologia em questão. Na Figura 58, especificam-se no cabeçalho a versão (1.0) e um comentário sobre a ontologia (no caso, uma descrição informal em inglês). Por fim, o conceito <dogcart> é especificado como um subtipo conceitual de <cart>. Para tanto, a linguagem OWL dispõe da etiqueta Class para indicar que determinado elemento é um conceito e a etiqueta `subClassOf` para codificar a relação hierárquica ou hiponímica entre <dogcart> e <cart>. Além disso, vê-se a glosa de <dogcart> (“carroça pequena puxada por um cachorro") como o comentário relacionado à classe Dogcart.

7.4. Contribuições para o desenvolvimento da WordNet.Br

7.4.1. A WN.Br

Diante da relevância lingüística e tecnológica das BDLs em formato *wordnet*, iniciou-se definitivamente em 2003, o empreendimento de construção da base[82] da *wordnet* para o português do Brasil, a WN.Br (DIAS-DA-SILVA, et al. 2002). Esse empreendimento, com base em pressupostos da Semântica Lexical, pura e computacional, e da Lexicografia Computacional, visa ao desenvolvimento de uma BDL no formato *wordnet* que poderá gerar léxicos especiais para a aplicação em sistemas de processamento do PB.

A base da WN.Br originou-se de um outro recurso, o Thesaurus Eletrônico para o Português do Brasil (TeP), ou seja, um inventário de sinônimos e antônimos armazenado na memória do computador que pode ser acoplado a um processador de textos. (DIAS-DA-SILVA, et al., 2000, 2002). A construção da base lexical do TeP, em especial, foi feita com base em certos pressupostos da WN.Pr, mais precisamente, com base na (i) noção de *synset* como unidade de representação do significado e (ii) no princípio de que cada *synset* aponta para um conceito lexicalizado (DIAS-DA-SILVA et al., 2006). Dessa forma, o TeP pôde ser "transformado" na base da WN.Br sem grandes obstáculos.

Dessa forma, a estruturação da base da WN.Br começou a partir de um inventário de 44.678 unidades lexicais do PB (17.388 substantivos, 15.073 adjetivos, 11.078 verbos e 1.139 advérbios), distribuídas em aproximadamente 19.872 *synsets*.

Os *synsets* da base da WN.Br foram construídos a partir de um conjunto de cinco dicionários do PB. Dentre eles, dois são dicionários gerais do PB contemporâneo (WEISZFLOG, 1998; FERREIRA, 1999), que na maioria das vezes definem a acepção de uma entrada por meio de unidades lexicais que expressam sentidos similares, ao invés de empregarem exclusivamente a definição aristotélica baseada em gênero próximo e diferença específica. Os outros dicionários empregados como fontes de extração de informações lexicais para a montagem de *synsets* foram: (a) um dicionário específico de verbos (BORBA, 1990), montado a partir de *corpus*; (b) dois dicionários de sinônimos e antônimos (BARBOSA, 2000; FERNANDES, 1997).

No que diz respeito às relações semântico-conceituais entre os *synsets*, responsáveis pela estruturação das redes no formato *wordnet*, a base da WN.Br conta inicialmente apenas com a interligação de 21,55% de seus *synsets* via antonímia.

---

[82] Diz-se "base da WN.Br" porque a *wordnet* para o PB está em pleno desenvolvimento e, por isso, não contém todas as características que definem uma BDL em formato *wordnet.*

No seu desenvolvimento, estão previstas as tarefas de (DIAS-DA-SILVA et al, 2006): (i) especificação das demais relações (hiponímia, meronímia, acarretamento e causa) e informações periféricas (glosas e frases-exemplo) previstas pela WN.Pr e (ii) indexação, nos moldes da EuroWordNet, da base da WN.Br à base da WN.Pr.

Atualmente, o trabalho de construção da WN.Br focaliza a associação ou o alinhamento dos *synsets* de verbos aos *synsets* correspondentes da WN.Pr (versão 2.1), assim como a inserção de glosas associadas aos conceitos verbais e de frases-exemplo associadas às formas verbais.

7.4.2. Contribuições para o refinamento e extensão dos *synsets* da WN.Br

Tendo em vista o fato de que a base da WN.Br está em plena construção, este trabalho contribui não somente para o alinhamento das bases norte-americana e brasileira, mas para o próprio desenvolvimento da WN.Br, por meio especificamente de: (i) montagem de *synsets* novos, ou seja, inclusão de conceitos lexicalizados ainda não armazenados; (ii) refinamento de *synsets* já armazenados, (iii) confirmação da boa-formação de *synsets* armazenados; (iii) elaboração de glosas, tanto para os *synsets* novos como para os refinados, e (iv) identificação de frases-exemplos, tanto para as unidades constitutivas dos *synsets* novos como para as dos *synsets* modificados.

Com base na metodologia adotada neste trabalho, foi possível identificar que, dos 205 conceitos lexicalizados no AmE, 84 deles são expressos no PB por meio de unidades lexicais. Tendo em vista que as redes *wordnets* armazenam apenas conceitos lexicalizados, a base da WN.Br pode, então, englobar tais conceitos.

Confrontando os dados gerados neste trabalho com a base atual da WN.Br, conclui-se que, dos 84 conceitos lexicalizados:

(a) 69 deles não constam na WN.Br e, por isso, são considerados "*synsets* novos";
(b) 13 deles constam na WN.Br, mas apresentam alguma diferença em relação aos originais da base; tais *synsets* são considerados como armazenados, porém, modificados;
(c) 2 deles são exatamente iguais aos armazenados na WN.Br, o que confirma a boa-formação dos *synsets* originais.

Os Quadros 14, 15 e 16 sintetizam essas informações, respectivamente[83].

[83] Os Quadros 14, 15 e 16 exibem apenas os *synsets*. As glosas, que também podem ser inseridas na WN.Br, estão especificadas no Quadro 11. As frases-exemplo, por sua vez, podem ser verificadas nos Apêndices 1 e 2.

Synsets novos
{aranha}
{armão}
{automóvel elétrico; carro elétrico}
{baratinha; roadster}
{berlinda}
{bicicleta; bike*;* magrela}
{biga}
{blindado}
{buggy}
{buldôzer; trator de lâmina}
{cabriolé}
{camburão}
{caminhão}
{caminhão-baú}
{caminhão-cegonha; cegonha}
{carreta; jamanta}
{carretão}
{carro de assalto}
{carro de boi}
{carro de corrida}
{carro de praça; táxi}
{carro esporte}
{carro fúnebre}
{carro; vagão}
{carroça}
{carroção}
{carro-forte}
{carro-madrinha}
{carro-salão}
{carruagem}
{caterpílar}
{charrete}
{cupê}
{empilhadeira}
{esqueite; skate; skateboard}
{fiacre}
{gôndola}
{jinriquixá}
{jipe}
{kart}
{landau; landô}
{limusine}
{locomotiva; locomotora}
{maria-fumaça}
{mobilete}
{monociclo}
{mountain bike}
{panzer}
{patinete}
{perua; wagon}
{pullman}
{radiopatrulha}

{sedã}
{stock car}
{tênder}
{todo-terreno}
{trailer}
{trator}
{triciclo; velocípede}
{tróica}
{trole}
{vagão-frigorífico}
{vagão-plataforma}
{vagão-restaurante}
{vagão-tanque}
{vagoneta}
{vagoneta}
{van; furgão}
{velocípede}

Quadro 14: Extensões da WN.Br: *synsets* (e glosas) novos.

***Synsets* modificados**	
***Synsets* originais da WN.Br**	***Synsets* modificados**
BR1065[84]={ambulância; assistência}	{ambulância}
BR1461={auto2; automóvel; carro; veículo}	{auto2; automóvel; carro}
BR1766={carruagem; caleça; caleche}	{caleche}
BR1820={carro-guincho; guincho; reboque}	{carro-guincho; carro-socorro; guincho; reboque}
BR1823={carruagem; diligência}	{diligência}
BR1883={carrinho de mão; carriola}	{carrinho de mão; carriola; carro de mão}
BR4372={bonde; trâmuei}	{bonde; trâmuei; tramway}
BR5390={carruagem; coche}	{carruagem; coche; sege}
BR6261={lambreta; motoneta; vespa}	{lambreta; motoneta; scooter; vespa}
BR6981={calhambeque; maxambomba}	{calhambeque; lata}
BR7324={moto; motocicleta; motociclo}	{moto; motoca; motocicleta; motociclo}
BR7477={aplanadora; niveladora; patrol}	{angledozer; aplanadora; niveladora; patrol; patrola}
BR7935={picape; pick up}	{caminhonete; picape; pickup}

Quadro 15: Refinamento da WN.Br: *synsets* modificados.

***Synsets* confirmados**
BR1819={carro-de-combate; tanque}
BR1821={carro-leito; vagão-dormitório; vagão-leito}

Quadro 16: Refinamento da WN.Br: *synsets* confirmados.

[84] Número de registro do *synset* na base da WN.Br

7.4.3. Contribuições para o alinhamento das bases WN.Br e WN.Pr

7.4.3.1. A tarefa de alinhamento léxico-conceitual

Como consta na Seção I deste trabalho, um dos objetivos propostos era o de contribuir diretamente para o alinhamento das bases da WN.Pr e WN.Br. O alinhamento dessas bases segue o método utilizado no desenvolvimento da base multilíngüe EuroWordNet. Esse método baseia-se na utilização de um ILI e de um conjunto de relações interlinguais. No processo de alinhamento das bases WN.Br e WN.Br, a indexação, em especial, (i) parte-se dos *synsets* já existentes na WN.Br e (ii) consideram-se como índice não-estruturado o conjunto dos conceitos lexicalizados na WN.Pr (versão 2.1), aos quais os *synsets* da base brasileira devem ser alinhados. Ou seja, nesse processo, a língua-fonte é o PB e a língua-alvo é o AmE.

A equivalência entre os *synsets* é especificada pelo mesmo conjunto de relações interlinguais originalmente previsto na EuroWordNet. Tais relações são representadas por meio dos seguintes rótulos:

- EQ-SYNONYM: indica "relação de equivalência sinonímica" entre um *synset* da língua-fonte e um *synset* da língua-alvo;
- EQ_NEAR_SYNONYM: indica "relação de equivalência sinonímica aproximada" entre um *synset* da língua-alvo e vários *synsets* da língua-fonte;
- EQ_HAS_HYPERONYM: rotula relações de equivalência em que a língua-fonte lexicaliza um conceito mais específico que a língua-alvo;
- EQ_HAS_HYPONYM: rotula relações de equivalência em que o conceito lexicalizado na língua-fonte é mais genérico que o conceito aproximado lexicalizado na língua-alvo.

Com base nesse elenco de relações, vê-se que, quando um mesmo conceito é lexicalizado, no caso, no AmE e no PB, os *synsets* que o codificam podem ser alinhados pela relação EQ_SYNONYM. Em outras palavras, entende-se que as lexicalizações em línguas distintas de um mesmo conceito devem ser alinhadas pela relação interlingual EQ_SYNONYM. Dessa forma, com base nos dados do Quadro 11 (pág. 152), foi possível identificar claramente os alinhamentos do tipo EQ_SYNONYM, os quais estão registrados no Quadro 17.

***Synsets* PB**	**Alinhamentos** (Rótulo da relação)	***Synsets* WN.Pr**
{ambulância}	EQ_SYNONYM	{ambulance}
{angledozer; aplanadora; niveladora; patrol; patrola}	EQ_SYNONYM	{angledozer}
{aranha}	EQ_SYNONYM	{trap}
{armão}	EQ_SYNONYM	{limber}
{auto; automóvel; carro}	EQ_SYNONYM	{car; auto; automobile; machine; motorcar}
{automóvel elétrico; carro elétrico}	EQ_SYNONYM	{electric car; electric automobile}
{baratinha; roadster}	EQ_SYNONYM	{roadster; runabout; two-seater}
{berlinda}	EQ_SYNONYM	{brougham}
{bicicleta; bike*;* magrela}	EQ_SYNONYM	{bicycle; bike; wheel; cycle}
{biga}	EQ_SYNONYM	{chariot}
{blindado}	EQ_SYNONYM	{armoured car}
{bonde; trâmuei; tramway}	EQ_SYNONYM	{streetcar; tram; tramcar; trolley; trolley car}
{buggy}	EQ_SYNONYM	{dune buggy; beach buggy}
{buldôzer; trator de lâmina}	EQ_SYNONYM	{bulldozer; dozer}
{cabriolé}	EQ_SYNONYM	{cabriolet; cab}
{caleche}	EQ_SYNONYM	{barouche}
{calhambeque; lata}	EQ_SYNONYM	{bus; jalopy}
{camburão}	EQ_SYNONYM	{police van; police wagon; paddy wagon; patrol wagon; wagon; black Maria}
{caminhão}	EQ_SYNONYM	{truck; motortruck}
{caminhão-baú}	EQ_SYNONYM	{van}
{caminhão-cegonha; cegonha}	EQ_SYNONYM	{transporter; car transporter}
{caminhonete; picape; pickup}	EQ_SYNONYM	{pickup; pickup truck}
{carreta; jamanta}	EQ_SYNONYM	{trailer truck; tractor trailer; trucking rig; rig; articulated lorry; semi}
{carretão}	EQ_SYNONYM	{dray}
{carrinho de mão; carriola; carro de mão}	EQ_SYNONYM	{handcart; pushcart; cart; go-cart}
{carro de boi}	EQ_SYNONYM	{oxcart}
{carro de corrida}	EQ_SYNONYM	{race car; racing car}
{carro de praça; táxi}	EQ_SYNONYM	{cab; hack; taxi; taxicab}
{carro esporte}	EQ_SYNONYM	{sports car; sport car}
{carro fúnebre}	EQ_SYNONYM	{hearse}
{carro; vagão}	EQ_SYNONYM	{railcar; car; railway car; railroad car}
{carroça}	EQ_SYNONYM	{cart}
{carroção}	EQ_SYNONYM	{covered wagon; Conestoga wagon; Conestoga; prairie wagon; prairie schooner}
{carro de assalto}	EQ_SYNONYM	{assault gun}
{carro-de-combate; tanque}	EQ_SYNONYM	{tank; army tank}
{carro-forte}	EQ_SYNONYM	{armoured car}
{carro-guincho; carro-socorro; guincho; reboque}	EQ_SYNONYM	{tow truck; tow car; wrecker}
{carro-leito; vagão-dormitório;	EQ_SYNONYM	{sleeping car; wagon-lit}

vagão-leito}		
{carro-madrina}	EQ_SYNONYM	{pace-car}
{carro-salão}	EQ_SYNONYM	{parlor car; parlour car; drawing room car; palace car; chair car}
{carruagem; coche; sege}	EQ_SYNONYM	{carriage; equipage}
{carruagem}	EQ_SYNONYM	{chariot}
{caterpílar}	EQ_SYNONYM	{Caterpillar; cat}
{charrete}	EQ_SYNONYM	{buggy; roadster}
{cupê}	EQ_SYNONYM	{coupe}
{diligência}	EQ_SYNONYM	{stagecoach; stage}
{empilhadeira}	EQ_SYNONYM	{forklift}
{esqueite; skate; skateboard}	EQ_SYNONYM	{skateboard}
{fiacre}	EQ_SYNONYM	{hansom; hansom cab}
{gôndola}	EQ_SYNONYM	{gondola car; gondola}
{jinriquixá}	EQ_SYNONYM	{jinrikisha; ricksha; rickshaw}
{jipe}	EQ_SYNONYM	{jeep; landrover}
{kart}	EQ_SYNONYM	{go-kart}
{lambreta; motoneta; scooter; vespa}	EQ_SYNONYM	{motor scooter; scooter}
{landau; landô}	EQ_SYNONYM	{landau}
{limusine}	EQ_SYNONYM	{limousine; limo}
{locomotiva; locomotora}	EQ_SYNONYM	{locomotive; railway locomotive}
{maria-fumaça}	EQ_SYNONYM	{steam locomotive}
{mobilete}	EQ_SYNONYM	{moped}
{monociclo}	EQ_SYNONYM	{unicycle; monocycle}
{moto; motoca; motocicleta; motociclo}	EQ_SYNONYM	{motorcycle; bike}
{mountain bike}	EQ_SYNONYM	{mountain bike; all-terrain bike; off-roader}
{panzer}	EQ_SYNONYM	{panzer}
{patinete}	EQ_SYNONYM	{scooter}
{perua; wagon}	EQ_SYNONYM	{beach wagon; station wagon; wagon}
{pullman}	EQ_SYNONYM	{Pullman}
{radiopatrulha}	EQ_SYNONYM	{cruiser; police cruiser; patrol car; police car; prowl car; squad car}
{sedã}	EQ_SYNONYM	{sedan}
{stock car}	EQ_SYNONYM	{stock car}
{tênder}	EQ_SYNONYM	{tender}
{todo-terreno}	EQ_SYNONYM	{four-wheel drive; 4WD}
{trailer}	EQ_SYNONYM	{trailer; house trailer}
{trator}	EQ_SYNONYM	{tractor}
{triciclo; velocípede}	EQ_SYNONYM	{tricycle; trike}
{tróica}	EQ_SYNONYM	{troika}
{trole}	EQ_SYNONYM	{gig}
{vagão-frigorífico}	EQ_SYNONYM	{refrigerator car}
{vagão-plataforma}	EQ_SYNONYM	{flatcar; flatbed; flat}
{vagão-restaurante}	EQ_SYNONYM	{dining car; diner; dining compartment; buffet car}
{vagão-tanque}	EQ_SYNONYM	{tank car; tank}
{vagoneta}	EQ_SYNONYM	{tramcar; tram}

{vagoneta}	EQ_SYNONYM	{handcar}
{van; furgão}	EQ_SYNONYM	{passenger van}
{velocípede}	EQ_SYNONYM	{velocipede}

Quadro 17: Os alinhamentos por meio da relação EQ_SYNONYM.

Para os casos em que há lacuna lexical, o projeto EuroWordNet estabeleceu as demais relações: EQ_NEAR_SYNONYM, EQ_HAS_HYPERONYM e EQ_HAS_HYPONYM, denominadas "relações de equivalência complexas" (do inglês, *complex-equivalence relations*) (PETERS et al., 1998). A identificação dos alinhamentos que se baseiam nessas relações, no entanto, não é tarefa tão clara ou direta quanto os alinhamentos descritos no Quadro 17. Conseqüentemente, neste trabalho, os alinhamentos que envolviam lacunas lexicais no PB foram identificados com o auxílio do *plug-in* de visualização TGVizTab. O procedimento de utilização do TGVizTab nessa tarefa é relatado a seguir.

7.4.3.2. A identificação das relações interlinguais por meio do plug-in TGVizTab

Como mencionado na Subseção anterior, o alinhamento das bases WN.Br e WN.Pr tem como ponto de partida os conceitos lexicalizados (os *synsets*) no PB. Conseqüentemente, o mesmo procedimento foi adotado neste trabalho.

Mais especificamente, parte-se de um conceito *x* para identificar todas as relações interlinguais de tipo complexo que esse conceito estabelece com os conceitos não lexicalizados no PB imediatamente a ele relacionados. Em outras palavras, pode-se descrever esse procedimento da seguinte forma: parte-se de um conceito *x* lexicalizado no PB e, com base nele, identificam-se as relações complexas estabelecidas com os conceitos relacionados a ele em nível 1.

Como a identificação das relações interlinguais de tipo complexo foi realizada com o auxílio do TGVizTab, nada mais ilustrativo que descrever mais detalhadamente esse procedimento com base no grafo gerado pelo referido *plug-in* de visualização. Para tanto, considera-se o grafo da Figura 59, em que se pode visualizar parte dos 217 conceitos que formam a interlíngua da base REBECA.

Vale salientar que, nesse grafo, todos os nós estão expandidos, pois, dessa forma, todas as relações interlinguais complexas estabelecidas entre um conceito lexicalizado no PB e os não-lexicalizados podem ser visualmente identificadas.

Na Figura 59, em especial, parte-se do nó que graficamente representa o conceito <railcar> ("veículo com rodas adaptado para se locomover sobre trilhos"), lexicalizado no PB por *carro*

e *vagão*. As lexicalizações de <railcar> podem ser verificadas no campo `Instance Browser`, destacado pelo retângulo vermelho.

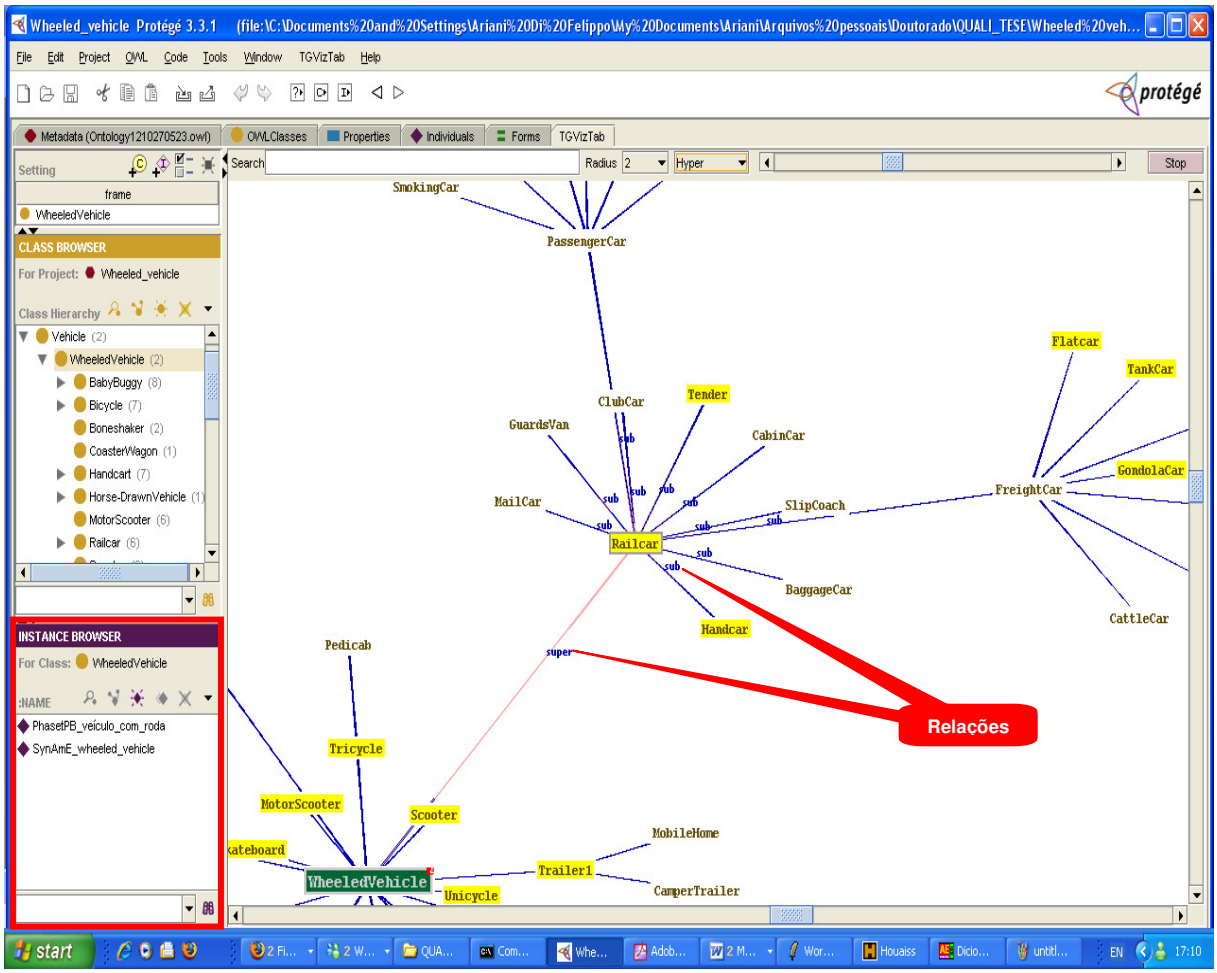


Figura 59: As relações de hierarquia estabelecidas por <railcar>.

Na Figura 59, o nó `Railcar`, em destaque, relaciona-se diretamente a oito nós que não possuem cor de fundo, ou seja, a oito conceitos que não são lexicalizados no PB: <passenger car>, <baggage car>, <cabin car>, <slip coach>, <guards van>, <freight car>, <mailcar> e <club car>. Além desses, Railcar estabelece relação com o conceito não lexicalizado no AmE WheeledVehicle.

Especificamente pelos rótulos dos arcos, vê-se que o conceito `Railcar` relaciona-se a esses conceitos por meio de relações conceituais distintas. Interpretando-se os rótulos dos arcos, verifica-se que Railcar é (i) hiperônimo de <passenger car>, <baggage car>, <cabin car>, <slip coach>, <guards van>, <freight car>, <mailcar> e <club car>, já que o arco é rotulado **sub**, e (ii) hipônimo de <wheeled vehicle>, já que o arco é rotulado **super**.

Em outras palavras, pode-se dizer que, no nível lexical, o *synset* do PB que expressa o conceito <railcar>, ou seja, o *synset* {carro; vagão} é:

(a) mais genérico que os *synsets* que expressam no AmE os conceitos <passenger car>, <baggage car, <cabin car>, <slip carriage>, <guards van>, <freight car>, <mailcar> e <club car>, ou seja, {passenger car; coach; carriage}, {baggage car; luggage van}, {cabin car; caboose}, {slip carriage; slip coach}, {guard's van}, {freight car} e {mailcar}, respectivamente;

(b) mais específico que o *synset* que codifica no AmE o conceito <wheeled vehicle>, ou seja, {wheeled vehicle};

Dessa forma, foi possível identificar as relações interlinguais complexas ilustradas na Figura 60 e sintetizadas no Quadro 18.

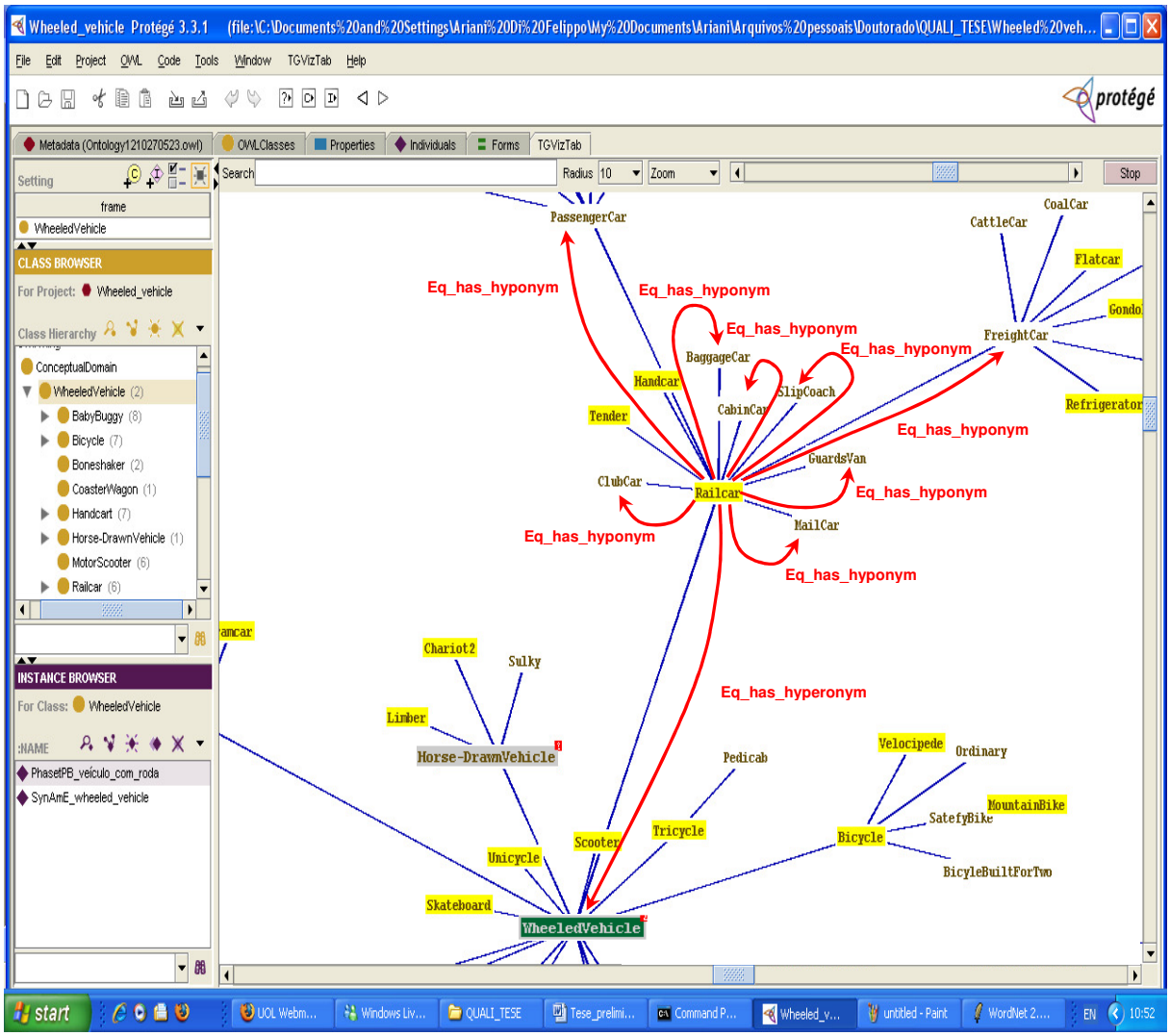


Figura 60: Identificação das relações interlinguais complexas com o TGVizTab.

***Synset* PB**	**Alinhamentos** (Rótulos das relações complexas)	***Synsets* WN.Pr**
{carro; vagão}	EQ_HAS_HYPONYM	{passenger car; coach; carriage} {baggage car; luggage van} {cabin car; caboose} {slip carriage; slip coach} {guard's van} {freight car} {mailcar} {club car; lounge car}
	EQ_HAS_HYPERONYM	{wheeled vehicle}

Quadro 18: As relações interlinguais complexas estabelecidas pelo *synset* {carro; vagão}.

Com o auxílio do TGVizTab, todos os nós que representam graficamente os 84 conceitos lexicalizados no PB foram "percorridos" e, devido à estruturação da interlíngua, todas as relações interlinguais complexas estabelecidas pelos *synsets* que codificam esses conceitos foram identificadas. Nesse alinhamento, também foram considerados os conceitos não-lexicalizados no AmE originários da WN.Pr, já que esses alinhamentos têm o objetivo principal de auxiliar na indexação das bases *wordnets* americana e brasileira.

O conjunto total das relações interlinguais está descrito no Quadro 19, em que os conceitos não-lexicalizados no AmE estão descritos em letras maiúsculas.

***Synsets* PB**	**Alinhamentos** (Rótulos das relações complexs)	***Synsets* WN.Pr**
{ambulância}	EQ_HAS_HYPONYM	{funny wagon}
{armão}	EQ_HAS_HYPERONYM	{HORSE-DRAWN VEHICLE}
{auto; automóvel; carro}	EQ_HAS_HYPERONYM	{motor vehicle}
	EQ_HAS_HYPONYM	{hatchback} {Model T} {minicar} {hot rod} {gas guzzler} {convertible} {compact, compact car} {hardtop} {sport utility, sport utility vehicle; SUV} {touring car, phaeton, tourer} {used-car, secondhand car} {subcompact, subcompact car} {loaner}
{bicicleta; bike*;* magrela}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{bicycle-built-for-two, tandem

		bicycle, tandem} {safety bicycle, safety bike} {ordinary, ordinary bicycle}
{blindado}	EQ_HAS_HYPERONYM	{ARMOURED VEHICLE}
{bonde; trâmuei; tramway}	EQ_HAS_HYPERONYM	{SELF-PROPELLED VEHICLE}
	EQ_HAS_HYPONYM	{horsecar}
{buggy}	EQ_HAS_HYPERONYM	{recreational vehicle}
{caminhão}	EQ_HAS_HYPERONYM	{motor vehicle}
	EQ_HAS_HYPONYM	{lorry, camion} {sound truck} {dump truck, dumper, tipper truck, tipper lorry, tip truck, tipper} {trailer truck, tractor trailer, trucking rig, rig, articulated lorry, semi}
{caminhão-baú}	EQ_HAS_HYPONYM	{bookmobile} {milk float} {laundry truck} {moving van} {delivery truck, delivery van, panel truck}
{caminhonete; picape; pickup}	EQ_HAS_HYPONYM	{technical}
{carreta; jamanta}	EQ_HAS_HYPONYM	{tandem trailer}
{carretão}	EQ_HAS_HYPERONYM	{horse cart, horse-cart}
{carrinho de mão; carriola; carro de mão}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{applecart} {barrow, garden cart, lawn cart, wheelbarrow} {hand truck, truck} {laundry cart} {serving cart} {shopping cart}
{carro de praça; táxi}	EQ_HAS_HYPONYM	{gypsy cab} {minicab}
{carro fúnebre}	EQ_HAS_HYPERONYM	{motor vehicle}
{carro; vagão}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{passenger car; coach; carriage} {baggage car; luggage van} {cabin car; caboose} {slip carriage; slip coach} {guard's van} {freight car} {mailcar} {club car; lounge car}
{carroça}	EQ_HAS_HYPONYM	{dogcart} {dumpcart} {horse cart, horse-cart}, {jaunting car, jaunty car} {pony cart, ponycart, donkey

		cart, tub-cart} {water cart}
{carroção}	EQ_HAS_HYPERONYM	{wagon; waggon}
{carro de assalto}	EQ_HAS_HYPERONYM	{ARMOURED VEHICLE}
{carro-de-combate}	EQ_HAS_HYPERONYM	{ARMOURED VEHICLE}
{carro-forte}	EQ_HAS_HYPERONYM	{ARMOURED VEHICLE}
{carro-leito; vagão-dormitório; vagão-leito}	EQ_HAS_HYPERONYM	{passenger car, coach, carriage}
{carro-salão}	EQ_HAS_HYPERONYM	{passenger car, coach, carriage}
{carruagem; coche; sege}	EQ_HAS_HYPERONYM	{HORSE-DRAWN VEHICLE}
	EQ_HAS_HYPONYM	{buckboard} {caroche} {chaise, shay} {chariot} {clarence} {coach, four-in-hand, coach-and-four} {droshky, drosky} {gharry} {post chaise} {stanhope} {surrey}
{caterpílar}	EQ_HAS_HYPERONYM	{tracked vehicle}
{diligência}	EQ_HAS_HYPERONYM	{coach, four-in-hand, coach-and-four}
{empilhadeira}	EQ_HAS_HYPERONYM	{SELF-PROPELLED VEHICLE}
{esqueite; skate; skateboard}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
{gôndola}	EQ_HAS_HYPERONYM	{freight car}
{kart}	EQ_HAS_HYPERONYM	{motor vehicle}
{lambreta; motoneta; scooter; vespa}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
{limusine}	EQ_HAS_HYPONYM	{berlin}
{locomotiva; locomotora}	EQ_HAS_HYPERONYM	{SELF-PROPELLED VEHICLE}
	EQ_HAS_HYPONYM	{diesel locomotive} {dinky} {electric locomotive} {pilot engine} {shunter} {switch engine} {traction engine} {tank engine; tank locomotive} {steam locomotive}
{mobilete}	EQ_HAS_HYPERONYM	{minibike; motorbike}
{monociclo}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
{moto; motoca; motocicleta; motociclo}	EQ_HAS_HYPERONYM	{motor vehicle}
	EQ_HAS_HYPONYM	{trail bike, dirty bike} {minibike; motorbike}
{patinete}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
{pullman}	EQ_HAS_HYPERONYM	{passenger car, coach, carriage}
{radiopatrulha}	EQ_HAS_HYPONYM	{panda car}

{sedã}	EQ_HAS_HYPONYM	{brougham}
{todo-terreno}	EQ_HAS_HYPERONYM	{motor vehicle}
{trailer}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{mobile home} {camper trailer}
{trator}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{skidder}
{triciclo; velocípede}	EQ_HAS_HYPERONYM	{WHEELED VEHICLE}
	EQ_HAS_HYPONYM	{pedicab}
{vagão-frigorífico}	EQ_HAS_HYPERONYM	{freight car}
{vagão-plataforma}	EQ_HAS_HYPERONYM	{freight car}
{vagão-restaurante}	EQ_HAS_HYPERONYM	{passenger car, coach, carriage}
{vagão-tanque}	EQ_HAS_HYPERONYM	{freight car}
{vagoneta}	EQ_HAS_HYPERONYM	{wagon; waggon}

Quadro 19: Os alinhamentos por meio das relações interlinguais complexas.

Os alinhamentos descritos no Quadro 19 são informações que podem ser diretamente inseridas na base da WN.Br. No entanto, além de fornecer as próprias relações interlinguais, este trabalho tem auxiliado o desenvolvimento da extensão do editor atual da WN.Br responsável por tornar o alinhamento das bases WN.Br e WN.Pr um processo auxiliado ou assistido por computador (do inglês, *the computer-aided alignment*) (DI FELIPPO, DIAS-DA-SILVA, 2007). Tal extensão é descrita a seguir.

7.4.4. O editor da WN.Br e a tarefa de alinhamento léxico-conceitual

No estágio atual do projeto de desenvolvimento da base WN.Br, os lingüistas responsáveis por "povoar" a base de dados lexicais são auxiliados por um editor, ou seja, uma ferramenta computacional dotada de uma interface gráfica.

Esse editor possui várias funcionalidades, as quais permitem: (i) inserir unidades lexicais; (ii) montar *synsets*; (iii) acrescentar uma ou mais glosas a cada um dos *synset* e (iv) acrescentar uma ou mais frases-exemplo a cada uma das unidades lexicais armazenadas na base.

Para dar continuidade ao desenvolvimento da WN.Br como um processo auxiliado por computador, propôs-se uma extensão desse editor, a qual auxilia a tarefa específica de alinhamento das bases *wordnets* brasileira e norte-americana, nos moldes do projeto EuroWordNet.

De certa forma, a extensão proposta para esse editor pauta-se nas etapas ou atividades, relativas ao domínio lingüístico, que foram realizadas neste trabalho. Atualmente, a extensão

do editor concretiza-se em uma janela composta por três módulos interconectados, destacadas pelos retângulos vermelhos na Figura 61.

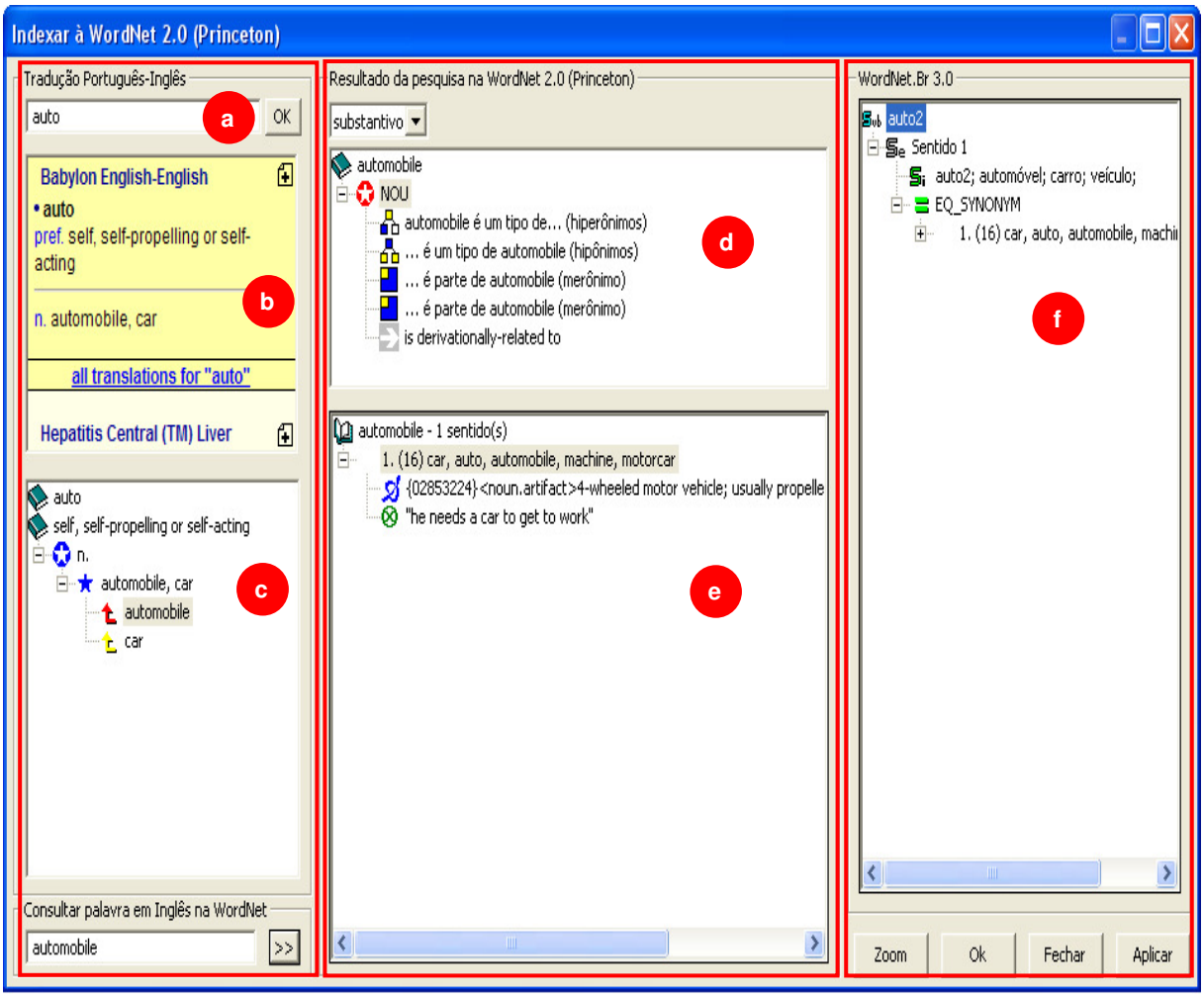


Figura 61: Interface gráfica do editor da WN.Br para o alinhamento léxico-conceitual.

Mais especificamente, o editor auxilia o processo de alinhamento por meio de seis passos (DIAS-DA-SILVA et al, 2006):

(a) seleção manual de uma unidade lexical *x* da WN.Br;

(b) pesquisa automática, em um dicionário bilíngüe, de todas as possíveis formas em AmE que correspondam a *x*; na versão atual, utiliza-se a versão *on-line* do dicionário PB-AmE Babylon[85];

(c) seleção manual de uma unidade *y* do AmE;

(d) busca automática por todos os *synsets* da WN.Pr que contêm *y*;

[85] Disponível em http://www1.uol.com.br/babylon/.

(e) seleção do *synset* da WN.Pr e identificação do tipo de relação de equivalência; tais tarefas podem ser feitas com o auxílio de algumas ferramentas *on-line* (isto é, dicionários, bases de dados lexicais e *corpora*) acopladas ao editor, sendo que, atualmente, o único recurso acoplado é a versão *on-line* do *´Michaelis: moderno dicionário da Língua Portuguesa´* (WEISZFLOG, 1998);

(f) associação do *synset* da WN.Pr (selecionado em (e)) ao *synset* da WN.Br que contém *x* por meio da ação de **drag-and-drop** (no PB, *arrastar e soltar*) e seleção do rótulo adequado para a relação identificada também em (e).

Com base na Figura 61, nota-se que cada módulo da interface permite ao lingüista realizar passos específicos, dentre os descritos em (a-f), durante o processo de alinhamento das bases WN.Br e WN.Pr.

Mais especificamente, o processo de alinhamento tem início com a seleção manual de uma unidade lexical armazenada na base da WN.Br. Em outras palavras, o processo de indexação tem como ponto de partida uma expressão lexicalizada no PB já armazenada na referida base (passo (a)). Por exemplo, considera-se a unidade lexical *auto* (como elemento constitutivo do *synset* BR00001462={auto2; automóvel; carro; veículo}. Equipado com um dicionário bilíngüe, o editor busca automaticamente todas as possíveis equivalências no AmE. No caso, o dicionário bilíngüe retorna *automobile* e *car* (passo (b)). Após a análise do lingüista, este seleciona uma das formas do AmE (passo (c)); no caso, *automobile*. Assim que a forma do AmE é selecionada, o editor automaticamente busca e exibe todos os *synsets* na WN.Pr que contêm tal forma; no caso, o editor retorna apenas {car; auto; automobile; machine; motorcar} (passo (d)). Na seqüência, o lingüista analisa (com base em dicionários monolíngues, *corpora* e no seu próprio conhecimento lingüístico) o *synsets* e verifica qual o tipo de relação que se estabelece entre {auto2; automóvel; carro; veículo} e {car; auto; automobile; machine; motorcar} (passo (e)). Em seguida, o *synset* da WN.Pr {car; auto; automobile; machine; motorcar} é "arrastado" para a área do terceiro módulo e "solto" sobre o *synset* do PB {auto2; automóvel; carro; veículo}, criando-se, por padrão, a relação de equivalência sinonímica entre esses *synsets*, representada pela inserção automática do rótulo EQ_SYNONYM (passo (f)).

Vale ressaltar que as relações EQ_SYNONYM estabelecidas entre os conceitos lexicalizados no AmE e no PB pertencentes ao domínio dos "veículos com rodas" constituem uma das colaborações deste trabalho para o desenvolvimento da WN.Br.

A Figura 62 ilustra outro tipo de alinhamento estabelecido com o auxílio da extensão do editor da WN.Br.

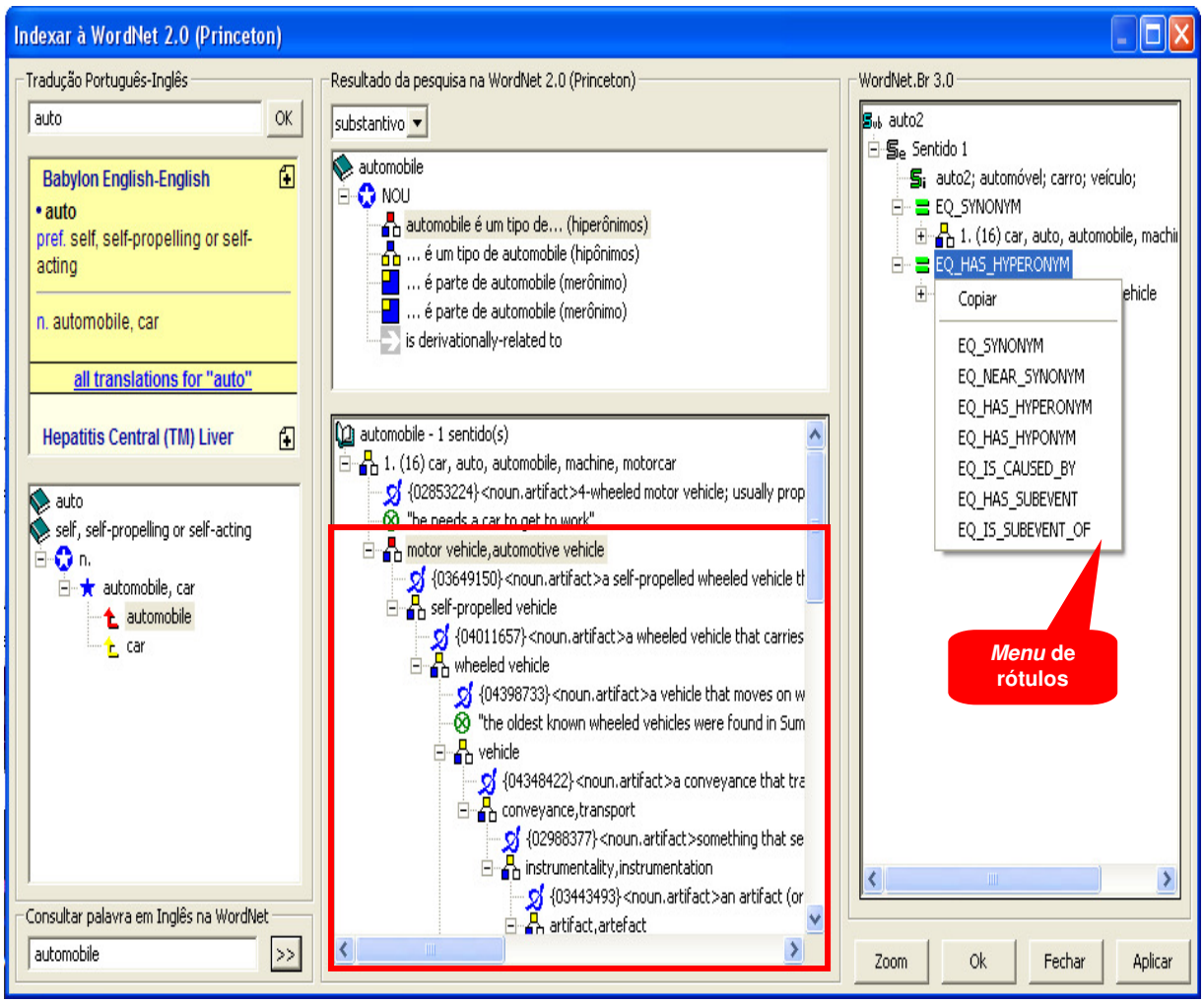


Figura 62: O editor da WN.Br e o alinhamento por EQ_HAS_HYPERONYM.

Nessa Figura, nota-se que, ao se clicar na informação hiperonímica contida no campo superior do módulo central, o editor exibe, no campo inferior do mesmo módulo, os conceitos/ *synsets* hiperônimos do *synset* em questão (ou seja, os *synsets* mais gerais), os quais estão destacados pelo retângulo vermelho na Figura 62. Dessa forma, é possível "arrastar", por exemplo, o *synset* {motor vehicle; automotive vehicle}, que é o *synset* imediatamente superior a {car; auto; automobile; machine; motorcar}, para a área do terceiro módulo e "soltá-lo" sobre o *synset* da WN.Br {auto2; automóvel; carro; veículo}, criando-se, assim, uma indexação entre eles. Por fim, por meio do *menu* de relações, que se abre ao clique do botão direito do *mouse* sobre o rótulo da relação, seleciona-se o rótulo EQ_HAS_HYPERONYM para especificar o alinhamento entre {auto2; automóvel; carro; veículo}, *synset*-fonte, e {motor vehicle; automotive vehicle}, *synset*-alvo.

Assim, o *synset* {auto2; automóvel; carro; veículo} alinha-se ao *synset* {car; auto; automobile; machine; motorcar} pela relação EQ_SYNONYM e ao *synset* {motor vehicle;

automotive vehicle} pela relação EQ_HYPERONYM. Vale ressaltar que esses relacionamentos também podem ser obtidos deste trabalho.

7.5. Síntese da Seção VII

Nesta Seção, os dados construídos no domínio lingüístico foram inseridos no editor de ontologia Protégé-OWL, que possibilitou, por conseguinte, a construção da base de dados REBECA. Nessa base, uma parte do léxico do AmE que recobre o domínio dos "veículos com rodas" está alinhada por meio de uma interlíngua estruturada (ontologia) à parte do léxico do PB que recobre os conceitos pertencentes ao mesmo domínio conceitual. Nesse editor, (i) os conceitos da ontologia foram inseridos como "classes", (ii) os demais conceitos, que se vinculam aos da interlíngua pelas relações de PARS (parte-todo) e PURP (propósito ou função), foram inseridos como "propriedades" das classes, e, por fim, (iii) as expressões lingüísticas (lexicalizações e SLRs) foram inseridas como "instâncias" das classe

Além disso, as contribuições para o desenvolvimento da WN.Br também foram descritas nesta Seção. Mais especificamente, essas contribuições vão desde o fornecimento de informações ou dados para a expansão e refinamento da base quanto à proposição de uma ferramenta que auxilia a tarefa de alinhamento das bases *wordnets* brasileira e norte-americana.

Seção VIII

Considerações finais e etapas futuras

Devido à globalização e à conseqüente demanda por sistemas de PLN como tradutores automáticos e sistemas de recuperação de informação multilíngüe, o desenvolvimento de recursos lexicais multilíngües faz-se premente.

Nesse cenário, destaca-se a base EuroWordNet, em que as unidades lexicais de várias línguas européias, organizadas em conjuntos de sinônimos (os *synsets*), estão alinhados por meio dos conceitos que expressam.

Para o processamento automático do PB, os recursos de natureza léxico-conceitual são bastante escassos. A base bilíngüe resultante do alinhamento em andamento das *wordnets* brasileira e americana será o único recurso dessa natureza para o par de línguas inglês norte-americano/ português brasileiro.

Do ponto de vista lingüístico, a construção de bases multilíngües e/ou bilíngües em que as unidades de línguas distintas estão alinhadas em função do conceito que expressam requer a tarefa nada trivial de identificação dos padrões de lexicalização das línguas envolvidas, ou seja, da associação regular entre as unidades lexicais e conceitos. Diz-se "nada trivial" porque essa tarefa precisa lidar com (i) os conceitos, que são elementos de natureza abstrata e complexa, com (ii) as unidades lexicais, cuja definição é questão de conflito entre os lingüistas, e com (iii) as próprias divergências léxico-conceituais existentes nas línguas. Do ponto de vista computacional, por sua vez, a construção de bases multilíngües e/ou bilíngües envolve questões como a escolha do formalismo semântico e a arquitetura da base, que pode ser baseada em uma interlíngua estruturada ou não-estruturada.

Assim, diante desse cenário, este trabalho teve como objetivo geral contribuir, do ponto de vista lingüístico, para a especificação de padrões de lexicalização e, do ponto de vista lingüístico-computacional, para o desenvolvimento de recursos léxico-conceituais para o processamento do PB. Para tanto, baseou-se na concepção de PLN como"uma engenharia do conhecimento lingüístico" e na metodologia tripartida, em que as atividades de pesquisa ficam distribuídas nos domínios lingüístico, lingüístico-computacional e implementacional. As atividades relativas ao domínio implementacional, em especial, não foram realizadas porque não fazem parte do escopo deste trabalho.

No domínio lingüístico, partiu-se de um conjunto de conceitos lexicalizados no AmE para a identificação das lexicalizações (ou não) dos mesmos no PB. Posteriormente, realizou-se a indexação das unidades do AmE, codificadas em *synsets*, às unidades do PB, também organizadas em *synsets,* por meio de uma interlíngua estruturada.

Mais especificamente, selecionaram-se conceitos do tipo "objetos (conceituais) concretos discretos", que intuitivamente categorizam referentes que são perceptíveis pelos sentidos, localizados no tempo e no espaço, contáveis e indivisíveis e que prototipicamente são expressos por nomes. Selecionaram-se os "objetos concretos discretos" pertencentes ao domínio dos "veículos com rodas", sendo extraídos da base da WN.Pr e delimitados em função dos construtos teóricos do modelo de RC MultiNet e de definições informais (as glosas). A escolha por esse domínio foi pautada em sua delimitação bem-definida e de extensão reduzida. O modelo de representação formal MultiNet, por sua vez, foi escolhido principalmente (i) pela sua homogeneidade, isto é, seus meios de representação são capazes de expressar conceitos lexicalizados e não-lexicalizados (expressos por meio de sintagmas e sentenças) da mesma forma, e (ii) pela sua adequação cognitiva, isto é, todo conceito tem uma representação única por meio da qual toda a informação a ele associada torna-se acessível. Para facilitar a manipulação do grande volume de dados com o qual se lidou, foi proposto um *template* conceitual para cada conceito sob análise. Esse *template* contém diferentes campos de preenchimento, denominados: (i) unidade conceitual, em que se registra o conceito em questão por meio de um símbolo mnemônico; (ii) glosa, em que são armazenadas as definições informais dos conceitos; (iii) tipo conceitual; (iv) traços semânticos; (v) atributos multidimensionais; e (vi) relações conceituais. Com exceção dos campos unidade lexical e glosa, os demais campos refletem diretamente os construtos do MultiNet. Para delimitação do conceitos, que implica o preenchimento dos *templates* conceituais, foram utilizados recursos lexicográficos (dicionários monolíngües), computacional (a base da WN.Pr) e textuais do AmE (*corpora*), sendo que a análise dos mesmos foi feita manualmente.

A partir da delimitação dos conceitos (e conseqüentemente do preenchimento dos *templates*), foram identificadas as lexicalizações dos mesmos no PB. Para tanto, partiu-se de uma tipologia específica de unidades lexicais e de alguns critérios de identificação. Além disso, considerou-se a noção de "sintagma livre recorrente" como alternativa para as lacunas lexicais. As unidades lexicais do PB (e os SLRs) foram identificadas pelo método manual de análise de recursos (i) lexicográficos convencionais (dicionários bilíngües AmE-PB e monolíngües e de sinônimos do PB, (ii) computacionais (as bases da WN.Pr e WN.Br), e (iii) textuais (ou seja, *corpora* do PB), e codificadas em *synsets*. As lexicalizações também foram

armazenadas em um *template* composto pelo campo "expressões lingüísticas", em que foram registradas as unidades do AmE (extraídas da WN.Pr) e as unidades e SLRs do PB. Ao final da delimitação dos conceitos e identificação das lexicalizações do PB, construiu-se um "*template* léxico-conceitual" para cada um dos conceitos originalmente extraídos da WN.Pr.

Quanto à expressão lexical dos conceitos no PB, verificou-se que, dos 205 conceitos analisados, 84 deles são lexicalizados no PB. Para os demais conceitos, o PB apresenta lacunas lexicais.

Para ilustar o tipo de diferença léxico-conceitual identificada, considere-se a Figura 63, em que a estrutura hierárquica ou hiponímica extraída da WN.Pr reflete a combinação de conceitos lexicalizados (*synsets*) e não-lexicalizados (em letras maiúsculas) no AmE. Mais especificamente, percebe-se que no AmE há o conceito <wagon>, que agrupa conceitos mais específicos como os expressos pelos *synsets* {bandwagon}, {cart}, entre outros. Esse conceito parece não ser lexicalizado no PB. Como conseqüência, percebe-se uma organização léxico-conceitual mais plana no PB.

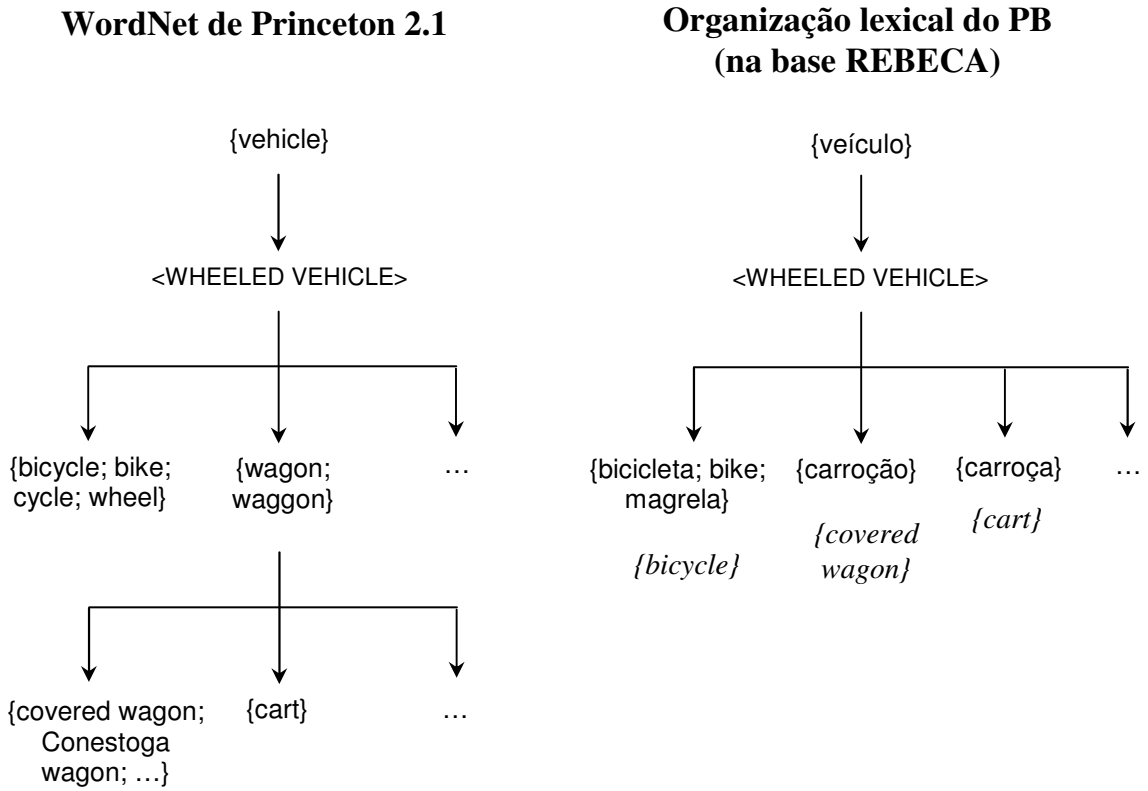


Figura 63: Um exemplo de diferentes estruturações léxico-conceituais.

Dessa forma, pode-se dizer que o trabalho ora apresentado contribui para a identificação dos padrões de lexicalização do PB, posto que um conjunto de conceitos lexicalizados,

pertencentes ao domínio dos "veículos com rodas", foi identificado. Em outras palavras, buscou-se capturar aqui a informação disponível sobre conceitualizações que são lexicalizadas no PB.

O alinhamento ou indexação dos conceitos, em especial, foi feito por meio de uma interlíngua estruturada, que nada mais é do que o próprio conjunto de conceitos extraídos da WN.Pr. Assim, os dados extraídos da WN.Pr, na forma de *synsets* estruturados por meio da relação de hiponímia, forneceram, ao mesmo tempo: (i) a parcela do léxico do AmE que recobre o domínio conceitual em questão, e (ii) o conjunto de conceitos que compõem a interlíngua. Os conceitos da interlíngua, em especial, foram representados formalmente por meio do modelo de RC MultiNet, que se fundamenta principalmente na tecnologia de representação das redes semânticas. Segundo o MultiNet, um conceito é representado em função de certos atributos multidimensionais e das relações que estabelece com outros conceitos. Dessa forma, pode-se dizer que os conceitos da interlíngua, ao serem representados pelo modelo MultiNet, tornam-se elementos estruturados e formais, ou seja, constituem uma "ontologia". Vale ressaltar que a interlíngua, no caso, organiza-se principalmente em função das relações de hiperonímia/ hiponímia.

No domínio lingüístico-computacional, os dados levantados no domínio lingüístico foram inseridos em uma ferramenta computacional de auxílio à criação e manutenção de ontologias, que pemitiu, a construção de uma base de dados léxico-conceituais bilíngüe. O alinhamento foi feito especificamente com o auxílio do editor de ontologia Protégé-OWL, amplamente utilizado no âmbito do PLN e da Engenharia Ontológica. Vale ressaltar que o editor que implementa o próprio paradigma MultiNet não está disponível e não foi possível encontrar na literatura outro que tenha sido construído com base nas redes semânticas. Dessa forma, escolheu-se um que fosse capaz de representar, mesmo que de forma adaptada, os tipos de informação contidos no *template* léxico-conceitual. Nesse sentido, o Protégé-OWL, construído com base no formato da linguagem OWL, mostrou-se bastante viável. Nesse editor, (i) os conceitos da ontologia foram inseridos como "classes", (ii) os demais conceitos, que se vinculam aos da interlíngua pelas relações de PARS (parte-todo) e PURP (propósito ou função), foram inseridos como "propriedades" das classes, e, por fim, (iii) as lexicalizações, assim como os SLRs, foram inseridas como "instâncias" das classes.

Uma vez inseridos no editor, os dados foram armazenados em uma base de dados denominada REBECA, que se caracteriza principalmente por:

(a) armazenar conceitos lexicalizados e, por isso, refletir as lexicalizações e as relações entre as unidades lexicais do PB;

(b) utilizar uma interlíngua hierarquicamente estruturada que, devido à representação de seus conceitos pelos construtos formais de um paradigma de representação do conhecimento (o MultiNet), pode ser considerada uma ontologia lingüisticamente motivada; em outras palavras, nessa base, as parcelas dos léxicos das referidas línguas estão indexadas por meio dessa ontologia;
(c) englobar apenas conceitos do tipo "objeto concreto discreto" e pertencentes ao domínio dos "veículos com rodas";
(d) fornecer, além da representação formal dos conceitos da ontologia, glosas para cada um deles;
(e) fornecer uma frase-exemplo para cada unidade lexical de ambas as línguas, que serve de contexto mínimo para essas unidades;
(f) fornecer expressões alternativas (SLRs) para as lacunas do PB;
(g) estar no formato OWL e, por isso, ser aplicável não só em algumas tarefas do PLN, como recuperação de informação multilíngüe, mas também em bibliotecas digitais, na própria Web semântica, na gestão de conhecimento, no comércio eletrônico, etc., e, por fim, ser exportável do editor Protégé-OWL para outros formatos ou linguagens.

Uma vez que uma parcela do léxico do AmE e uma parcela do léxico do PB estão alinhadas à interlíngua, a base REBECA propicia a observação das diferenças na lexicalização e no relacionamento léxico-conceitual interno às línguas, como ilustra o exemplo da Figura 63. Nessa base, é possível ir de um conceito em uma língua para o conceito na outra língua, que está vinculado ao mesmo elemento da interlíngua. Assim, verifica-se que, nessa base, não somente o reflexo dos padrões de lexicalização do AmE e do PB pode ser verificado, mas também a materialização de ontologias em léxicos computacionais e da própria constituição do léxico mental. Dessa forma, evidencia-se seu potencial de uso em várias aplicações do PLN, por exemplo, na recuperação de informação multilíngue, pela expansão de unidades lexicais de uma língua a unidades lexicais relacionadas em outra língua via a interlíngua estruturada. Por fim, agrega-se, a esse potencial tecnológico, um outro: a geração totalmente automática de um dicionário digital bilíngüe AmE-PB do domínio dos "veículos com rodas".

Uma característica aparentemente problemática da REBECA reside no fato de que, a inserção, na ontologia, de conceitos específicos de outras línguas pode implicar certa reestruturação da ontologia. No entanto, por outro lado, evita-se o número excessivo de *links* entre as expressões lingüísticas (*synsets*) e a interlíngua, como acontece no projeto EuroWordNet.

Ainda no domínio lingüístico-computacional, o editor Protégé-OWL, mais especificamente o seu *plug-in* de visualização TGVizTab, auxiliou a tarefa de alinhamento dos conceitos nos moldes da EuroWordNet, ou seja, por meio de um índice não-estruturado. Esse índice é composto pelos mesmos conceitos que compõem a ontologia da base REBECA, com a diferença de que este não apresenta estrutura ou organização. As indexações resultantes dessa fase podem ser diretamente aplicadas ao alinhamento das bases *wordnets* brasileira e norte-americana.

Além dos alinhamentos segundo o modelo EuroWordNet, este trabalho contribui diretamente para o desenvolvimento da WN.Br ao realizar as seguintes tarefas:

(a) refinamento de *synsets* (já armazenados);
(b) montagem de novos *synsets*;
(c) elaboração de glosas para os *synsets* novos e/ou refinados;
(d) a seleção de frases-exemplos para cada unidade lexical dos *synsets*.

Assim como ocorre com a WN.Pr, um dos recursos lexicais mais utilizados no processamento do AmE, a WN.Br, quando terminada, poderá ser utilizada em várias tarefas do PLN que envolvam o PB.

Como desenvolvimento futuro deste trabalho, propõem-se as tarefas descritas na seqüência:

(a) *Refinamento do domínio conceitual dos "veículos com rodas"*

Apesar de ter sido construída com base em recursos lexicográficos e textuais do AmE e, por isso, ser considerada extensa, a WN.Pr não armazena alguns conceitos lexicalizados no AmE pertencentes ao domínio aqui investigado. Esse é o caso, por exemplo, do conceito "caminhão equipado com reservatório fechado para transporte de combustível", lexicalizado por *tank truck*, e "caminhão equipado com reservatório fechado para transporte de água", lexicalizado por *tank wagon.* Esses dois conceitos também são lexicalizados no PB. O primeiro é expresso pelas unidades *caminhão-tanque* e *carro-tanque*, e o segundo pelos unidades *caminhão-cisterna*, *carro-pipa* e *caminhão-pipa*.

Assim, propõe-se refinar o domínio dos "veículos com rodas", buscando identificar outros padrões de lexicalização do AmE que não constam da base da WN.Pr, tomada aqui como ponto de partida. Para tanto, poder-se-á seguir uma análise semasiológica, em que se parte delexicais extraídas de *corpora* do AmE para a identificação dos conceitos a elas subjacentes e conseqüente montagem de *synsets*.

Mais especificamente, uma metodologia possível para o refinamento do domínio em questão consiste na utilização do *plug-in* do Protégé denominado OntoLT, cujos resultados parecem promissores (BUITELAAR et al., 2003, 2004). De um modo geral, esse *software* acoplável ao referido editor caracteriza-se pela (i) extração de unidades lexicais de um *corpus* anotado lingüisticamente (isto é, *corpus* em que elementos como a categoria sintática das palavras, sintagmas e estrutura de argumentos dos predicadores estão explicitamente marcados por etiquetas específicas) e (ii) identificação de possíveis relações conceituais (principalmente a de hiponímia) entre elas, ou seja, identificação das unidades lexicais enquanto `Classes` ou `Properties` no editor Protégé-OWL. A identificação das unidades/ conceitos no OntoLT é feita tanto com base em informações lingüísticas (isto é, padrões léxico-sintáticos, p.ex.: "SN *such as* SN") quanto em informações estatísticas.

Vale ressaltar que, uma vez identificados, os conceitos serão (i) delimitados em função dos construtos do MultiNet (tipos conceitual, traços semânticos, atributos, etc.) e de glosas. Estas, em especial, poderão ser elaboradas com o subsídio dos dicionários monolíngües do AmE. As unidades sinônimas, que comporão os *synsets*, poderão ser identificadas manualmente nos próprios dados extraídos pelo OntoLT e em dicionários do AmE.

Uma vez identificados, os conceitos poderão ser inseridos na interlíngua ou ontologia da base REBECA e na WN.Br.

(b) *Investigação dos conceitos (lexicalizados) "específicos" do PB e inclusão dos mesmos na ontologia da base REBECA*

Como mencionado, a ontologia da base REBECA é composta apenas por conceitos do AmE e, por isso, é lingüisticamente motivada por esse sistema lingüístico. Em outras palavras, há conceitos lexicalizados no PB que não estão presentes na interlíngua, ou seja, na WN.Pr (2.1). Como ilustração, citam-se os conceitos "carrinho de madeira que se compõe de uma tábua montada sobre esse mecanismo" e "carroça pequena", lexicalizados no PB pelos potenciais *synsets* {carrinho de rolimã; rolimã} e {carrocim}, respectivamente.

Assim, propõe-se identificar os conceitos lexicalizados no PB e, na seqüência, as lexicalizações (ou não) dos mesmos no AmE. Dessa forma, o PB será a língua-fonte e o AmE, a língua-alvo. Conseqüentemente, a investigação dos padrões de lexicalização poderá resultar na identificação das lexicalizações dos conceitos (lexicalizados no PB) ou na verificacão da existência de lacunas no AmE.

A investigação dos conceitos do PB poderá ser feita com o auxílio do *plug-in* do Protégé-OWL denominado OntoLP (RIBEIRO Jr., 2006). Esse *plug-in* é a adaptação do OntoLT para

o PB. Assim como o OntoLT, o OntoLP caracteriza-se pela extração de unidades lexicais de um *corpus* anotado lingüisticamente e identificação de possíveis relações conceituais entre elas. Por se basear em conhecimento lingüístico, certas modificações foram feitas para que o OntoLP pudesse tratar especificamente o PB. Nesse *software*, por exemplo, o *corpus,* previamente construído, é anotado pelo *parser* PALAVRAS (BICK, 2000), desenvolvido especificamente para o PB, e os padrões léxico-conceituais também foram modificados para atender às características do PB.

Uma vez inseridos na base REBECA, os dados resultantes dessa investigação ampliarão a ontologia/ interlíngua de tal forma que esta passa a ser capaz de expressar a lexicalização (ou não), no AmE, dos conceitos oriundos do PB. Os mesmos dados também poderão ser inseridos na WN.Br

(c) *Investigação de outros domínios conceituais e conseqüente extensão da base REBECA*

A metodologia aplicada neste trabalho para a identificação dos conceitos lexicalizados no AmE e no PB poderá ser empregada na (i) ampliação do domínio dos "veículos com rodas" para o domínio dos "veículos" e (ii) na investigação de outros domínios conceituais (p.ex.: o domínio dos recipientes, dos alimentos, etc.). Essa metodologia, que se baseia especialmente em informações extraídas de recursos lexicográficos, pode ser estendida pela utilização de informações provenientes de *corpora*, como as que são extraídas pelo *plug-in* OntoLP.

Dessa forma, novos padrões de lexicalização do PB serão identificados, assim como possíveis diferenças léxico-conceituais entre o AmE e o PB. Os dados gerados por essas novas investigações poderão, naturalmente, ser incluídos nas bases REBECA e WN.Br.

Referências bibliográficas

AIRES, R. V. X. **Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil.** São Carlos, 2000. Dissertação (Mestrado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2000.

AGIRRE E., ALDEZABAL I., POCIELLO E. Lexicalization and multiword expressions in the Basque WordNet. In: INTERNATIONAL WORDNET CONFERENCE, 3, Jeju Island. **Proceedings…** Jeju Island (Korea), 2006, p 131-138. ISBN 80-210-3915-9

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.

ALLGAYER, J. REDDIG-SIEKMANN, C. What KL-ONE lookalikes need to cope with natural language: scope and aspect of plural noun phrases. **Sorts and types in artificial intelligence**, p. 240–285, 1990.

ALLAN, K. **Natural language semantics.** Oxford: Oxford Blackwell Publishers, 2001.

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

ALMEIDA, J. Ambiguidade lexical. **Alfa**, São Paulo, n. 34, p. 187-193, 1990.

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.

ALUISIO, S.M., PINHEIRO, G.M., FINGER, NUNES, M.G.V., TAGNIN, S.E. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: CORPUS LINGUISTICS CONFERENCE, 2003, UK. **Proceedings…** UK, 2003, p. 14-21.

ALLWOOD, J. et al. **Logic and linguistics.** Cambridge: CUP, 1977.

ANTONUIOU, G., HARMELEN, F van. Web ontology language: OWL. In: STAAB, S., STUDER, R. (Eds.). **Handbook on ontologies.** International Handbooks on Information Systems. Berlin, Heidelberg: Springer-Verlag, 2004, p.67-92.

ATKINS, S, ZAMPOLLI, A. **Computational approaches to the lexicon.** Oxford: Oxford University Press, 1994.

AUSTIN, J. L. **How to do things with words.** New York: Oxford University Press, 1965.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley FrameNet project. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS

(COLING/ACL), 17, Montreal, Quebec (Canada). **Proceedings...** Montreal, Quebec (Canada), 1998, pp. 86-90.

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

BARWISE, J., PERRY. J. **Situations and attitudes.** Cambridge: Cambridge University Press, 1983.

BELIAEVA, L. N., PIOTROWSKI, R. G., SOKOLOVA, S. V. Principles of linguistic automata and their information bases design. **Terminology and Knowledge Engineering,** v. 2, p. 419-425, 1990.

BENTIVOGLI, L., PIANTA, E. Detecting hidden multiwords in bilingual dictionaries. In: EURALEX INTERNATIONAL CONGRESS, 10, Copenhagen, Denmark. **Proceedings...** Copenhagen (Denmark), 2002, p. 14-17.

___. Beyond lexical units: enriching wordNets with *phrasets*". In: EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL'03) (Research Note Sessions), 10, Budapest, Hungary. **Proceedings...** Budapest (Hungary), 2003, p. 67-70.

___. E. Extending wordnet with syntagmatic information. In: GLOBAL WORDNET CONFERENCE, 2, 2004, Brno, Czech Republic. **Proceedings...** Brno, Czech Republic, 2004. p. 47–53.

___; 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.

BERBER SARDINHA, A. P. **Lingüística de corpus**. São Paulo, Barueri: Editora Manole, 2004.

BERNERS-LEE et al. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**. 2001. Disponível em: <http://www.cs.nyu.edu/rgrimm/teaching/reading/semantic-web.pdf>. Acesso em: 25 jan. 2008.

BICK, E. **The parsing system PALAVRAS: automatic grammatical analysis of portuguese in a constraint grammar framework**. 2000. PhD Thesis. Arhus University, 2000.

BIDERMAN, M. T. C. A unidade lexical e o lema no dicionário de língua. **Corpo e Voz,** Ano XV, n. 1, p. 71-76, 1997.

___. O conceito lingüístico de palavra. **Palavra**, Rio de Janeiro: Grypho, n. 5, p. 81-97, 1999.

___. **Teoria lingüística**: t**eoria lexical e lingüística computacional** São Paulo: Martins Fontes, 2001.

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

BLACKBURN, S. **Dicionário Oxford de Filosofia.** Rio de Janeiro: Jorge Zahar, 1997.

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

BOGURAEV, B., BRISCOE, T. (Eds.). **Computational Lexicography for Natural Language Processing**, London: Longman, 1989.

BOLSHAKOV, I., GELBUKH, A. **Computational linguistics: models, resources and applications**. México City: Centro de Investigación en Computación/ Instituto Politécnico Nacional, 2004.

BORBA, F. da S. **Organização de dicionários: uma introdução à lexicografia.** São Paulo: Editora UNESP, 2003.

___. **Introdução aos estudos linguísticos.** Campinas: Editora Pontes, 1991.

___. (Coord.) **Dicionário gramatical de verbos do português contemporâneo.** São Paulo: Editora Unesp, 1991.

BORST, W. N. **Construction of engineering ontologies**. Holanda, 1997. Tese (Doutorado). Disponível em: <http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf>. Acesso em: 05 abril 2006.

BRACHMAN, R. J. On the epistemological status of semantic networks. In: FINDLER, N. V. (Ed.). **Associative Networks.** New York: Academic Press, 1979, p. 3-50.

BRACHMAN, R. J, LEVESQUE, H (Eds.). **Readings in knowledge representation.** San Mateo, Ca: Morgan Kaufmann, 1985.

___. **Knowledge representation and reasoning**. San Mateo, Ca: Morgan Kaufmann, 2004.

BRACHMAN, R.J., SCHMOLZE, J.G. An overview of the KL-ONE knowledge representation system. **Cognitive Science**, v. 9, n. 2, p. 171- 216, 1985.

BRISCOE, T. Lexical issues in natural language processing. In: Klein, E., Veltman, F. (Eds.). **Natural Language and Speech**. Spinger-Verlag, 1991, p. 39-68.

BUITELAAR, P., OLEJNIK, D., SINTEK, M. Ontolt: A Protégé plug-in for ontology extraction from text. In: DEMO SESSION OF THE INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC), 2003, Sanibel Island, Florida. **Proceedings...** Sanibel Island, Florida, 2003.

___. A Protégé plug-in for ontology extraction from text based on linguistic analysis. In: EUROPEAN SEMANTIC WEB SYMPOSIUM (ESWS), 1, 2004, Heraklion, Greece. **Proceedings**... Heraklion, Greece, 2004, p. 31–44.

BUTLER, C. S. **Systemic linguistics: theory and applications**. London: Batsford Academic and Educational, 1985.

CANN, R. **Formal semantics.** Cambridge: Cambridge University Press, 1993.

CARNAP, R. **Meaning and necessity.** Chicago: Chicago Phoenix Books, 1958.

CARROL, J. Parsing. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 12, p. 233-248.

CASELI, H.M.; NUNES, M.G.V.; FORCADA, M.L. LIHLA: A lexical aligner based on language-independent heuristics. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL (ENIA), 5, 2005, São Leopoldo-RS. **Proceedings...**. São Leopoldo, 2005, p.641-650.

CAVAZZA, M., ZWEIGENBAUM, P. Lexical semantics: dictionary ou encyclopedia? In: SAINT-DIZIER, P, VIEGAS, E. (Eds.) **Computational Lexical Semantics**. Cambridge: CUP, 1995, cap. 16, p. 336-348.

CERCONE, N., MCCALLA, G. (Eds.). **The knowledge frontier: essays in the representation of knowledge** (Symbolic Computation Series) New York: Springer-Verlag, 1987.

CHANDRASEKARAN, B. Ontologies: What are they? Why do we need them? **IEEE Intelligent Systems** (Special Issue on Ontologies), 14(1), p. 20-26, 1999.

CHIERCHIA, G. **Semântica.** Tradução de Luiz Arthur Pagani, Ligia Negri e Rodolfo Ilari. Campinas, SP: Editora da UNICAMP; Londrina, Pr: EDUEL, 2003. Tradução de Semântica.

CHIERCHIA, G., MCCONNEL-GINET, S. **Meaning and Grammar: an introduction to Semantics.** Cambridge, Mass.:The MIT Press, 1990.

COLLINS, A. M., QUILLIAN, M. R. Retrieval time from semantic memory. **Journal of verbal behavior and verbal learning**, 8, p. 240-247, 1969.

COPELAND, C., J. DURAND, S. KRAUWER, MAEGAARD, B. (Eds.). **The Eurotra formal specifications.** Luxembourg: Office for Official Publications of the European Community, 1991.

COPELAND, J. **What is artificial intelligence?** Publicação Maio de 2000. Disponível em <http://www.alanturing.net> Acesso em: 10 set. 2006.

CORCHO, O. et al. WebODE: an integrated workbench for ontology representation, reasoning, and Exchange. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT (EKAW): ONTOLOGIES AND THE SEMANTIC WEB, 13, 2002, Siguenza, Spain. **Proceedings ...** (Lecture Notes in Computer Science, 2473), 2002, p.138-153.

COSERIU, E., GECKLER, H. **Trends in structural semantics**. Tübinger: Gunter Narr Verlag, 1981.

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

CRUSE, A. **Lexical semantics.** Cambridge: Cambridge University Press, 1986.

___. **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.

CULLER, J. **Saussure.** London: Fontana, 1976.

DAVIS, R., SHROBE, H., SZOLOVITS, P. What is knowledge representation? **AI Magazine,** v. 14, n. 1, p.17-33, 1993.

DIAS-DA-SILVA, B. C. **A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais.** Araraquara, 1996. 272p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.

____. Bridging the gap between linguistic theory and natural language processing. In: INTERNATIONAL CONGRESS OF LINGUISTICS, 16, 1997, Paris. **Proceedings…** Oxford: Elsevier Sciences, 1998, n. 16, p. 1-10.

___. O estudo lingüístico-computacional da linguagem. **Letras de Hoje,** Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

___ et al. A construção de um thesaurus eletrônico para o português do Brasil. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE – PROPOR, 5, 2000, Atibaia, SP. **Proceedings…** Atibaia, 2000, p. 1-10.

___, OLIVEIRA, M. F., MORAES, H. R. Groundwork for the development of the Brazilian Portuguese Wordnet. In: INTERNATIONAL CONFERENCE PORTUGAL FOR NATURAL LANGUAGE PROCESSING **–** PorTAL, 3, 2002, Faro. **Proceedings…** Faro, 2002. p. 189-196.

___, MORAES, H. R. A construção de thesaurus eletrônico para o português do Brasil. **Alfa**, São Paulo; Editora da UNESP, v. 47(2), p. 101-115, 2003.

___, DI FELIPPO, A., HASEGAWA, R. Methods and tools for encoding the WordNet.Br sentences, concept glosses, and conceptual-semantic relations. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE – PROPOR, 7, 2006, Itatiaia. **Proceedings…** Rio de Janeiro, 2006, p. 120-130. ISBN 3-540-34045-9

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: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION **–** LREC, 6, 2008. Marrakech, Morocco. **Proceedings...** Marrakech, 2008.

DI FELIPPO, A. **Representação lingüístico-computacional dos adjetivos valenciais do Português**. Araraquara, 2004, 120p. Dissertação (Mestrado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 2004.

___. Ontologias lingüísticas aplicadas ao processamento automático de línguas naturais: o caso das redes wordnets. In: SIMPÓSIO NACIONAL DE LETRAS E LINGÜÍSTICA (I SIMPOSÓSIO INTERCIONAL DE LETRAS E LINGÜÍSTICA), 11, Uberlândia. **Anais...** Uberlândia, Minas Gerais: Universidade Federal de Uberlândia, Brasil. 22 a 24 de Novembro, 2006.

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

DIK, S. C. **The theory of functional grammar**. Berlin, New York: Mouton de Gruyter, 1997.

DORR, B.J. JORDAN, P.W., BENOIT, J.W. A survey of current research in machine translation. **Advances in Computers,** v. 49, p. 1-68, 1999.

DOWTY, D. R. et al. **Introducion to Montague semantics**. Dordrecht, Reidel, 1981.

FAUSTINO, S. **Wittgenstein, o eu e sua gramática**. São Paulo: Editora Ática, 1995.

FELLBAUM, C. A semantic network of English: the mother of all wordnets. **Computers and the Humanities**, Dordrecht: Kluwer Academic Publichers, v. 32, p. 209-220, 1998a.

___ (Ed.). **WordNet: an electronic lexical database**. Cambridge, MA: MIT Press, 1998b.

FELTRIM, V. D. **Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes Web de auxilio à escrita acadêmica em português.** São Carlos, 2004, 169p. Tese (Doutorado em Ciência da Computação) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

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: Ed. Positivo, 2004. 1 CD-ROM

FILLMORE, C. J. The case for case. In: BACH, E., HARMS, R. (Ed.). **Universals in linguistic theory**. New York: Holt, Rinehart & Winston, 1968, p. 1-88.

___. Frame semantics and the nature of language. In: CONFERENCE ON THE ORIGIN AND DEVELOPMENT OF LANGUAGE AND SPEECH, 1976. **Annals of the New York Academy of Sciences**, v. 280, p. 20-32, 1976.

___. Topics in lexical semantics. In: COLE, R. W. (Ed.) **Current issues in linguistic theory.** Bloomington: Indiana University Press, 1977, p. 76-138.

___. WOOTERS, C., BAKER, C. F. Building a large lexical databank which provides deep semantics. In: PACIFIC ASIAN CONFERENCE ON LANGUAGE, INFORMATION AND COMPUTATION – PACLIC, 15, 2001, Hong Kong. **Proceedings**…Hong Kong, 2001. Disponível em <http://www.icsi.berkeley.edu/ framenet/papers/dsemlex16.ps.gz>. Acessado em 20 Julho de 2007.

FRAWLEY, W. In defense of the dictionary. A response to Haiman. **Lingua**, 55, p.53-61, 1981.

FREGE, G. On sense and nominatum. In: MARTINICH (Ed.). **The philosophy of language.** New York & Oxford: Oxford University Press, 1990, p. 190-202. Tradução de Über Sinn und Bedeutung.

GANGEMI, A., GUARINO, N., OLTRAMARI, A. Conceptual analysis of lexical taxonomies: thecase of WordNet top-level. In: INTERNATIONAL CONFERENCE OF FORMAL ONTOLOGY IN INFORMATION SYSTEM - FOIS, 2, 2001, Ogunquit, Maine, USA. **Proceedings...** 2001, Ogunquit, Maine, USA, p. 285-296.

GEERAERTS, D. Cognitive grammar and the history of lexical semantics. In: B. RUDZKA-OSTYN (Ed.) **Topics in Cognitive Linguistics.** Amsterdam and Philadelphia: John Benjamins, 1988, p. 647-77.

___. (Ed.). **Words and other wonders:** papers on lexical and semantic topics. Berlin: Mouton de Gruyter, 2006.

GENNARI J. H. et al. The evolution of Protégé: an environment for knowledge-based systems development. **International Journal of Human Computer-Studies**, 58(1), p.89-123, 2003. Disponível em <http://smi.stanford.edu/smi-web/reports/SMI-2002-0943.pdf>.

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

GLOCK, H-J. **A Wittgenstein dictionary**, Oxford: B. Blackwell, 1998.

GODDARD, C. **Semantic analysis: a practical introduction**. Oxford: Oxford University Press, 1998.

GONZALO, J., VERDEJO, F., PETERS, C., CALZOLARI, N. Applying EuroWordNet to cross-language text retrieval. **Computers and the Humanities**, Dordrecht: Kluwer Academic Publichers, v. 32, p. 185-207, 1998.

GOODENOUGH, W. H. Componential analysis and the study of meaning. **Language**, 32, p. 195-216, 1956.

GRICE, H. P. Logic and conversation. In: COLE, P., MORGAN, J. (Eds.). **Syntax and semantics 3: speech acts**. New York: Academic Press, 1975, p. 41–58.

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

___. Information extraction. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 30, p. 545-559.

___. CALZOLARI, N. Lexicons. In: ____. **Survey of the state of the art in human language technology**. New York: Cambridge University Press, 1997.

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

HAIMAN, J. Dictionaries and encyclopedias. **Lingua**, 50, p. 329-357, 1980.

HALLIDAY, M.A.K. **An introduction to functional grammar**, London: Edward Arnold, 1985.

HALLIDAY, M., HASAN, R. **Cohesion in English.** London: Longman, 1976.

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.

HAYES-ROTH, F. Expert systems. In: SHAPIRO, E. (Ed.). **Encyclopedia of artificial intelligence.** New York, Wiley, 1990, p. 287-298.

HELBIG, H., HEROLD, C. MESNET: a multilayered extended semantic network. **Informatik-Bericht,** 185. FernUniversität Hagen, Hagen, Germany, 1995.

___, SCHULZ, M. Knowledge representation with MESNET: a multilayered extended semantic network. In: AAAI SPRING SYMPOSIUM ON ONTOLOGICAL ENGINEERING, 1997, Califórnia. **Proceedings …** Stanford, Califórnia, 1997, p. 64-72.

___, GNÖRLICH, C. Multilayered extended semantic networks as a language for meaning representation in NLP systems. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS – CICLING'02, 3, 2002, Mexico City, Mexico. **Proceedings ...** (Lecture Notes in Computer Science, n. 2276), Berlin: Springer, 2002, p. 69-85.

___. **Knowledge representation and semantics for natural language.** Berlin, Heidelberg: Springer-Verlag, 2006.

HIRST, G. **Semantic interpretation and the resolution of ambiguity**. Cambridge: Cambridge University Press, 1992.

___. Ontology and the lexicon. In: STAAB, S., STUDER, R. (Eds.). **Handbook on ontologies.** International Handbooks on Information Systems. Berlin, Heidelberg: Springer-Verlag, 2004, p.209-229.

HORRIDGE, M. et al. **A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools**. The University Of Manchester, 2004. Disponível em <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf>.

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

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

HOVY, E. Text summarization. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 32, p. 583-598.

HUTCHINS, W. J. **Machine translation: past, presence, future**. Ellis Horwwod/Wiley, Chichester/New York, 1986.

___. Machine translation: general overview. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 27, p. 501-511.

___. SOMERS, H. L. **An introduction to machine translation**. London: Academic Press, 1997.

JACKENDOFF, R. **Semantics and cognition.** Cambridge, Mass.: MIT Press, 1983.

___.**Foundations of language: brain, meaning, grammar, evolution.** Oxford, Oxford University Press, 2002.

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

JANSSEN, M. Multilingual lexical databases, lexical gaps, and SIMuLLLDA. **International Journal of Lexicography**, 17, p. 136 – 154, 2004.

JURAFSKY, D., MARTIN, J.H. **Speech and language processing:** an introduction to natural language processing, computational linguistics and speech recognition. Upper Saddle River, New Jersey: Prentice Hall, 2000.

KAPLAN, R. M. Syntax. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 04, p. 70-90.

KATZ, J., FODOR, J. A. The structure of a semantic theory. **Language**, 39, p. 170-210, 1963.

KATZ, J., POSTAL, P. **An integrated theory of linguistic description.** Cambridge, Mass.: MIT Press, 1964.

KAY, M. Parsing in functional unification grammar. In: DOWTY, D. R. et al. (Eds.). **Natural language parsing.** Cambridge: CUP, 1985, p. 251-278.

KILGARIFF, A., Grefenstette, G. Introduction to the special issue on the Web as Corpus. **Computational Linguistics**, 29, 2003.

KLAVANS, J. Computational linguistics. In: O'GRADY, W. et al. **Contemporary linguistics**. New York: St. Martin's Press, 1989, cap. 15, p. 413-447.

KNUBLAUCH, H. et al., The protégé OWL plugin: an open development environment for semantic web applications. In: International Semantic Web Conference (ISWC2004), 3, 2004, Hiroshima, Japon. **Proceedings...** (Lecture Notes in Computer Science, 3298), 2004, p. 229-243.

KOZAKI, K. et al. Hozo: an environment for building/using ontologies based on a fundamental consideration of "role" and "relationship": In: INTERNATIONAL CONFERENCE KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT (EKAW): ONTOLOGIES AND THE SEMANTIC WEB, 13, 2002, Siguenza, Spain. **Proceedings ...** (Lecture Notes in Computer Science, 2473), 2002, p.213-218.

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.

LANGACKER, R .W. **Foundations of cognitive grammar**, vol. 1: theoretical prerequisites. Stanford: Stanford University Press, 1987.

___. **Foundations of cognitive grammar**, vol. 2: descriptive application. Stanford: Stanford University Press, 1991.

LEECH, G. **Semantics.** Middlesex, Penguin Books Ltd., Hardmondswort, 1976.

___. **Principles of pragmatics.** London: Longman, 1983.

LEHRER, A. **Semantic fields and lexical structure.** Amsterdam and New York: North Holland, 1974.

LEITE, D.S. et al. Extractive automatic summarization: does more linguistic knowledge make a difference? In: WORKSHOP ON TEXTGRAPHS-2 GRAPH-BASED ALGORITHMS FOR NATURAL LANGUAGE PROCESSING, 2007. Rochester, USA. **Proceedings…** Rochester, 2007, p. 17-24.

LENAT, D. GUHA, R. **Building large knowledge based systems: representation and inference in the Cyc project.** Addison-Wesley Publishing, 1990.

LEVELING, J., HELBIG, H. A robust natural language interface for access to bibliographic databases. In: WORLD MULTICONFERENCE ON SYSTEMICS, CYBERNETICS AND INFORMATICS – SCI'02, 6, 2002, Orlando. **Proceedings …** Orlando: Florida: International Institute of Informatics and Systemics (IIIS), 2002, p. 133-138.

___, J. Natural language access to the GIRT4 data. In: CROSS-LANGUAGE SYSTEM EVALUATION CAMPAIGN – CLEF'03, 2003, Trondheim, Norway. **Proceedings…** Trondheim, Norway, 2003, p. 253-262.

___. 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'04, 2004, Avignon, France: Le Centre de Hautes Etudes Internationales d'informatique Documentaire. **Proceedings …** Avignon, France, 2004, p. 431-446.

LEVELT, W. J. M. **Speaking: to intention to articulation**. Cambridge, Mass.: The MIT Press, 1992.

LÔBNER, S. **Understanding semantics.** Oxford: Oxford University Press, 2002.

LOUNSBURRY, F. A semantic analysis of Pawnee kinship usage. **Language,** 32, p. 159-194, 1956.

LYONS, J. **Structural semantics**. Oxford, Basil Blackwell, 1963.

___. **Semantics**. Cambridge: Cambridge University Press, vol. 2, 1977.

___. **Introdução à lingüística teórica.** Tradução de Rosa V. Mattos e Hélio Pimentel. São Paulo: Ed. da USP, 1979.

___. **Linguagem e lingüística.** Rio de Janeiro: Zahar, 1981.

MAGNINI, B., SPERANZA, M. Merging global and specialized linguistic ontologies. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION - LREC, 3, 2002, Las Palmas. **Proceedings**… Las Palmas: University of Las Palmas, 2002.

MAHESH, K., NIRENBURG, S., BOWIE, J., FARWELL, D. An Assessment of Cyc for Natural Language. **Memoranda in Computer and Cognitive Science MCCS-96-302**. Las Cruces, N.M.: New Mexico State University, 1996.

MARTINS, R.T. **A nova língua do imperador.** Campinas, 2004. 296p. Tese (Doutorado) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2004.

___, et al. Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. **Natural Language Engineering**, 4, p. 287-307, 1998.

MICROSOFT PRESS. **Microsoft press dicionário de informática**. Rio de Janeiro: Editora Campus, 805 p., 1998.

MILLER, G. A., CHARLES, W. G. Contextual correlates of semantic similarity. **Language and cognitive processes**, 6, p. 1-28, 1991.

___, FELLBAUM, C. Semantic networks of English. **Cognition**, Amsterdam: Elsevier, v. 41, p. 197 – 229, 1991.

___, JOHNSON-LAIRD, P. N. **Language and perception.** Cambridge, Mass.: The MIT Press, 1976.

MINSKY, M. A framework for representing knowledge. In: HAUGELAND, J. (Ed.). **Mind design.** Cambridge, Mass.: The MIT Press, 1975, p. 95-128.

MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics.** Oxford, New York: Oxford University Express, 2004.

MIZOGUCHI, R. Ontology engineering environments. In: STAAB, S., STUDER, R. (Eds.). **Handbook on ontologies.** International Handbooks on Information Systems. Berlin, Heidelberg: Springer-Verlag, 2004, p. 275-298.

MORAES, H. R. **Aspectos sintaticamente relevantes do significado lexical: estudo dos verbos de movimento.** Araraquara, 2008. 172p. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 2008.

MORATO, J., MARZAL, M.A., LLORENS, J., MOREIRO, J. WordNet Applications. In: GLOBAL WORDNET CONFERENCE, 2, 2004, Brno, Czech Republic. **Proceedings...** Brno, Czech Republic, 2004. p. 270-278.

MORLEY, B. WebCorp: a tool for online linguistic information retrieval and analysis. In: A. RENOUF, A., Kehoe, A. (Eds.). **The changing face of corpus Linguistics**. Amsterdam: Rodopi, 2006.

MOURA, H. M. de M. **Significação e contexto: uma introdução a questões de semântica e pragmática.** Florianópolis: Editora Insular, 2000.

MÜLLER, A, L. et al. (Orgs.) **Semântica formal**. São Paulo, Ed. Contexto, 2003.

MUNIZ, M. C. M. **A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB**. São Carlos, 2004. 72p. Dissertação (Mestrado em Ciência da Computação) - Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, 2004.

NILES, I., PEASE A. Origins of the IEEE standard upper ontology. In: WORKSHOP ON THE IEEE STANDARD UPPER ONTOLOGY (Working Notes of the IJCAI-2001), 2001, Seattle, Washington. **Proceedings …** Seattle, Washington, 2001.

NIRENBURG, S. Knowledge and choices in machine translation. In: NIRENBURG, S. (Org.). **Machine translation – theoretical and methodological issues**. Cambridge: Cambridge University Press, 1989, p. 1-15.

___ et al. **Machine translation**. San Mateo: Morgan Kaufmann, 1992.

___, RASKIN, V. **Ontological semantics.** Cambridge, MA: MIT Press, 2004

NUGUES, P. M. (2006). **An introduction to language processing with perl and prolog.** Springer-Verlag, 2006.

NUNES, M.G.V. et al. A construção de um léxico da língua portuguesa do Brasil para suporte à correção automática de textos. **Relatório Técnico ICMC-USP**, 42, 36p, 1996.

OGDEN, C.K., RICHARDS, I. A. **The meaning of meaning**. New York: Harcourt Brace Javanovitch, 1946.

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

PARDO, T.A.S.: **SENTER: um segmentador sentencial automático para o português do Brasil.** Relatório técnico NILC-TR-06-01, São Carlos, SP, 6p., 2006.

___; RINO, L.H.M.; NUNES, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In: INTERNATIONAL WORKSHOP ON THE COMPUTATIONAL PROCESSING OF PORTUGUESE - PorTal, 3, 2002, Faro. **Proceedings…** Faro, 2002, p. 210-218.

___, NUNES, M. G. V. Review and evaluation of DiZer - an automatic discourse analyzer for Brazilian Portuguese. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF WRITTEN AND SPOKEN PORTUGUESE – PROPOR, 7, 2006, Itatiaia. **Proceedings…** Rio de Janeiro, 2006, p. 180-189. ISBN 3-540-34045-9

PARTEE, B. H. Semantics: mathematics or psychology. In BÄUERLE, R. et al. (Eds.). **Semantics from different points of view.** Berlin: Springer-Verlag, 1979, p. 1-14.

PERINI, M. A. **Gramática descritiva do português.** 3ª edição. São Paulo: Editora Ática, 1999.

PEETERS, B. Setting the scene: some recent milestones in the lexicon-encyclopedia debate. In: _____. (Ed.) **The lexicon-encyclopedia interface**. New York/Amsterdam: Elsevier, 2001, p. 1-52.

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.

POLLARD, C., SAG, I. **Head-driven phrase structure grammar**. Chicago: University of Chicago Press, 1994.

POTTIER, B. **Linguistique générale: problèmes et methods**. Paris: Ed. Klincksieck, 1974.

PUSTEJOVSKY, J. **The generative lexicon**. 2ª ed. Cambridge: Mass.: The MIT Press, 1996.

___. Type construction and the logic of concepts. In: BOUILLON, P., BUSA, F. (Eds.). **The syntax of word meaning**. Cambridge: Cambridge University Press, 2001.

PUTNAM, H. The meaning of "meaning". In: GUNDERSON, K. (Ed.). **Language, Mind and Knowledge.** Minneapolis: University of Minnesota Press, 1975, p. 131–193.

QUILLIAN, M. R. Word concepts: a theory and simulation of some basic semantic capabilities. **Behavioral Science**, 12, p. 410-430, 1967.

___. Semantic Memory. In: MINSKY, M. (Ed.). **Semantic information processing.** Cambridge, Mass.: MIT Press, 1968, p. 227-270.

RASKIN, V. Linguistic and encyclopedic knowledge in text processing. **Quaderni di Semantica**, 6, p. 92-102, 1985.

RIBEIRO Jr., L. C.; VIEIRA, R. . Geração de ontologias para a web semântica a partir de textos da língua portuguesa. In: WORKSHOP DE TESES E DISSERTAÇÕES EM INTELIGÊNCIA ARTIFICIAL (WTDIA)/ SBIA-IBERAMIA, 3, 2006, Ribeirão Preto. **Proceedings ...** Ribeirão Preto, SP, 2006.

RIGAU, G. Automatic acquisition of lexical knowledge from MRDs. Tesis doctoral, Departament de Llenguatges i Sistemes Informàtics, UPC, Barcelona, 1998.

RINO, L.H.M. et al. A comparison of automatic summarization systems for Brazilian Portuguese texts. In: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE – SBIA. 17, 2004, São Luís-MA. **Proceedings...** São Luís, 2004, p. 235-244.

ROCA, S. C. Individuación e información parte-todo. Representación para el procesamiento computacional del lenguaje. **Estudios de Lingüística Española,** v. 08, 2000. Disponível em <http://elies.rediris.es/elies8/>. Acesso em: 10 jun. 2005.

ROSCH, E. Natural categories. **Cognitive Psychology**, 4, p. 328-350, 1973.

ROSCH, E., MERVIS, C. Family resemblances: studies in the internal structure of categories. **Cognitive Psychology**, 7, p. 573-605, 1975.

SAG I. A., BALDWIN T., BOND, F., COPESTAKE A., AND FLICKINGER, D. Multiword expressions: A pain in the neck for NLP. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS – CICLING, 3, 2002, Mexico City. **Proceedings…** Mexico City, 2002, p. 1–15.

SANFILIPPO, A. Lexicons for constraint-based grammars. In: Cole, R. A (Ed). **Survey of the state of the art in human language technology**. Oregon: Graduate Institute, p. 118-121, 1995.

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

SAUSSURE, F. de. **Curso de lingüística geral**. Tradução Antônio Chelini, José Paulo Paes, Isidoro Blikstein. 25ª ed. São Paulo: Cultrix, 1999.

SCHANK, R. C., ABELSON, R. P. **Scripts, plans, goals and understanding.** Hillsdale, N. J.: Lawrence Erlbaum Associates, 1977.

SEARLE, J. R. **Expression and meaning.** Cambridge: Cambridge University Press, 1979.

SÉRASSET, G. An interlingual lexical organisation based on acceptions. In: ICLA, 1994, Penang, Malaysia. **Proceedings ...** Penang, Malaysia, 1994, p. 20-33.

___. 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.

SIEWIERSKA, A. **Functional grammar.** London-New York: Routledge, 1991.

SINCLAIR, J. **Looking up.** London-Glasgow: Collins, 1987.

SLOCUM, J. A Survey of machine translation: its history, current status, and future prospects. In: SLOCUM, J. (Org). **Machine translation systems**. Cambridge: Cambridge University Press, 1985, p.1-41.

SMITH, E. E. Theories of semantic memory. In: ESTES, W. K. (Ed.). **Handbook of learning and cognitive processes,** 5. Hillsdale, N.J: Lawrence Erlbaum Ass., 1978.

SMITH, M. K. et al. (Eds.). **OWL Web ontology language guide.** W3C Recommendation, 2004. Disponível em <http://www.w3.org/TR/owl-guide/>.

SOMERS, H. Machine translation: latest developments. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 28, p. 512-528.

SPARCK-JONES, K., WILLET, P. **Readings in information retrieval.** São Francisco: Morgan Kaufmann, 1997.

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

TAYLOR, J. R. **Linguistic categorization: prototypes in linguistic theory.** Oxford: Clarendon Press, 1985.

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.

___. Force dynamics in language and cognition. **Cognitive Science,** 12, p. 49-100, 1988.

TZOUKERMAN, E., KLAVANS, J. L., STRZALKOWSKI, T. Informaton retrieval. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2004, cap. 29, p. 529-544.

USCHOLD, M., GRUNINGER, M. Ontologies: principles, methods and applications. **Knowledge Engineering Review**, 11(2), p. 93-155, 1996.

ULLMANN, S. **Précis de sémantique française**. Bern: Francke, 1952.

___. **The principles of semantics.** 3ª ed. Oxford: Basil Blackwell, 1963.

UNIVERSITY OF CHICAGO. **Philologic user manual**, 2006. Disponível em: <http://philologic.uchicago.edu/manual>.

VARELI, G. B., ZAMPOLLI, A. **Survey of the state of the art in human language technology.** Cambridge: CUP, 1997.

VILELA, M.; SILVA, F. The position of the adjective in Portuguese: centre and periphery of the adjective class. In: SILVA, A. da S. et al. (Orgs.). **Linguagem, cultura, cognição: estudos de linguística cognitiva.** Coimbra: Almedina, p.661-690, 2004.

VOSSEN, P. Introduction to EuroWordNet. **Computers and the Humanities**, Dordrecht: Kluwer Academic Publichers, v. 32, p. 73-89, 1998.

___ et al. Compatibility in interpretation of relations in EuroWordNet. **Computers and the Humanities,** Dordrecht: Kluwer Academic Publichers, v. 32, p. 153-184, 1998.

___, PETERS, W., GONZALO, J. Towards a universal index of meaning. In: WORKSHOP OF THE SPECIAL INTEREST GROUP ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON THE LEXICON OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS – SIGLEX/ACL, 1999, College Park, Maryland. **Proceedings**... Maryland: University of Maryland, 1999, p. 81-90.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics**. Oxford, New York: Oxford University Express, 2004, cap. 11, p. 219-232.

WIERZBICKA, A. Dictionaries and encyclopedias. How to draw the line. In: DAVIS, P. W. (Ed.). **Alternative linguistics: descriptive and theoretical modes.** Amsterdam: John Benjamins, 1995, p. 289-315.

WEISZFLOG, W. **Michaelis: moderno dicionário da Língua Portuguesa**. Editora Melhoramentos, 1998. Disponível em < http://michaelis.uol.com.br/moderno/portugues/index.php>.

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

____. **Semantics. primes and universals.** Oxford: Oxford University Press, 1996.

WILKS Y. A., SLATOR, B. M., GUTHRIE, L. M. **Electric words: dicionaries, computers, and meanings.** Cambridge, Mass: ACL-MIT Press, 1996.

WILKINS, A. J. Conjoint frequency, category size, and categorization time. **Journal of verbal learning and verbal behavior,** 10, p. 382-385, 1971.

WINOGRAD, T. **Understanding natural language.** New York: Academic Press, 1972.

WITTGENSTEIN, L. **Investigações filosóficas**. São Paulo: Ed. Abril Cultural, 1979.

WOODS, W. A. What's in a link: foundations for semantic networks. In: BRACHMAN, R., LEVESQUE, H (Eds.). **Readings in knowledge representation.** Palo Alto: Morgan Kaufmann, 1985, p. 217-242.

ZAVAGLIA, C. **Análise da homonímia no português: tratamento semântico com vistas a procedimentos computacionais.** Araraquara, 2002, v I. 199p., v II, 360p. Tese (Doutorado em Letras) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 2002.

___. Produção de ontologias específicas: a modelagem da Onto-Eco. **Estudos Lingüísticos,** 28, p. 1182-1187, 2005.

APÊNDICE 1: As unidades lexicais do PB e suas respectivas frases-exemplo.

Unidades lexicais	***Synsets***	**Frases-exemplo**
ambulância	{ambulância}	Ele seguiu até um posto de gasolina e foi levado para o Hospital Souza Aguiar por uma F]ambulância dos bombeiros.
angledozer	{angledozer; aplanadora; niveladora; patrol; patrola}	O I]angledozer é de construção parecida à do buldôzer com a diferença de que o avental de varredura pode orientar-se segundo ângulos diversos com relação ao eixo da marcha.
aplanadora	{angledozer; aplanadora; niveladora; patrol; patrola}	Dona Maria, o seu marido foi atropelado por uma I]aplanadora!
aranha	{aranha}	Íamos os dois, na A]aranha sacolejante, ao trote largo do Paputinga.
armão	{armão}	O féretro do líder palestino foi tirado da mesquita em um I]armão puxado por seis cavalos negros e escoltado pela guarda cerimonial do exército egípcio.
auto	{auto; automóvel; carro}	O automóvel se deteve e o animal grunhia, e os grunhidos estremeciam o I]auto.
automóvel elétrico	{automóvel elétrico; carro elétrico}	A GM americana vai lançar o primeiro I]automóvel elétrico em série do mundo.
automóvel	{auto; automóvel; carro}	Quando um F]automóvel em movimento freia bruscamente, seus ocupantes tendem a se chocar contra o banco ou o vidro da frente.
baratinha	{baratinha; roadster}	O primeiro veículo do empresário foi uma I]"baratinha" Ford Roadster, ano 1929, cor vermelha e preta.
berlinda	{berlinda}	O Brasil só conheceu a I]berlinda a partir da segunda metade do século XVIII.
bicicleta	{bicicleta; bike; magrela}	Ricardo entrega pizza com uma F]bicicleta e não é registrado.
biga	{biga}	Ganhava principalmente por sabotar várias I]bigas adversárias antes da corrida,ou então jogava sua biga contra as outras.
bike	{bicicleta; bike; magrela}	Desde a semana passada, ele faz de F]bike o trajeto entre sua casa, no Jardim Marajoara (zona sul), e o escritório, na avenida Faria Lima.
blindado	{blindado}	De longe, alguns soldados em um I]blindado do exército apenas descansavam dentro do veículo.
bonde	{bonde; trâmuei; tramway}	Para qualquer pessoa que conheceu a velha Nova Orleans, viajar num F]bonde da linha St. Charles é como fazer uma viagem numa máquina do tempo.
buggy	{buggy}	O carro, em exposição na porta do Palácio de Congressos de Lyon, quartel-general da reunião deste ano do G-7, parece um I]buggy, pelo tamanho e design, mas, ao contrário deste, é coberto.

buldôzer	{buldôzer; trator de lâmina}	Em 1997, a gente acordou em estado de choque, com o barulho de um I]buldôzer que estava arrasando as árvores nativas a apenas 18 metros de distância da nossa casa.
cabriolé	{cabriolé}	No Rio de Janeiro, o primeiro I]cabriolé chegou em 1839.
caleche	{caleche}	No dia seguinte, de manhã, o senhor Ivan Ivanovitch ordenou ao lacaio que preparasse a I]caleche.
calhambeque	{calhambeque; lata}	Perder o I]calhambeque para o banco foi até bom.
camburão	{camburão}	Fomos sumariamente jogados num velho F]camburão que nos despejou no quartel da PE, na rua Barão de Mesquita, que se tornaria um dos emblemas da repressão.
caminhão	{caminhão}	Silva, depois de vender sua casa para a Vale, voltou a Serra Pelada com um F]caminhão para buscar sua família e seus móveis e pertences.
caminhão-baú	{caminhão-baú}	Um I]caminhão-baú pegou fogo após arrastar uma moto durante cerca de 50 metros pela Marginal Tietê.
caminhão-cegonha	{caminhão-cegonha; cegonha}	Todos os três carros estavam sendo transportados em um FI]caminhão-cegonha que foi interceptado pela PF na quarta-feira na BR-040, perto de Sete Lagoas (MG).
caminhonete	{caminhonete; picape; pickup}	Dois homens chegaram em uma I]caminhonete e outros seis em um Gol.
carreta	{carreta; jamanta}	Os motoristas do ônibus e da F]carreta morreram na hora.
carretão	{carretão}	Ele foi encontrado em 1784 e no ano seguinte foi colocado em um I]carretão puxado por 14 bois.
carrinho de golfe	{carrinho de golfe}	O ator americano Bill Murray foi pego pela polícia sueca dirigindo, em estado de embriaguez, um I]carrinho de golfe pelas ruas de Estocolmo.
carrinho de mão	{carrinho de mão; carriola; carro de mão}	O I]carrinho de mão foi especialmente projetado para o manuseio de cargas como caixas e sacos, com agilidade e segurança.
carriola	{carrinho de mão; carriola; carro de mão}	É nesse cenário que Michael parte carregando sua mãe em uma I]carriola improvisada em direção ao interior do país.
carro de assalto	{carro de assalto}	Os alemães equiparam o veículo a que deram o nome de I]carro de assalto, com um canhão de origem russa de calibre 57mm.
carro de boi	{carro de boi}	Meu pai era colono em uma fazenda e tinha F]carro de boi.
carro de corrida	{carro de corrida}	Fantástico protótipo de um I]carro de corrida da Fórmula 1 em super exposição.

carro de mão	{carrinho de mão; carriola; carro de mão}	Todos os dias, o agricultor João Esteves de Lima, 67, sai pelo menos duas vezes com um I]carro de mão onde transporta quatro botijões em busca de água.
carro de praça	{carro de praça; táxi}	O descumprimento da tabela por parte do proprietário do I]carro de praça poderia levar o infrator ao pagamento de pesada multa.
carro elétrico	{automóvel elétrico; carro elétrico}	No começo de 2008 chegará ao mercado um I]carro elétrico diferente dos outros.
carro esporte	{carro esporte}	Em 1987, Paul Roupinian ganhou dos pais um F]carro esporte da Nissan.
carro fúnebre	{carro fúnebre}	O I]carro fúnebre com o corpo de Slobodan Milosevic chegou à praça da cidade de Pozarevac para uma cerimônia que precederá o enterro do ex-presidente iugoslavo.
carro	{auto; automóvel; carro}	É como o F]carro, que recebe o combustível (gasolina ou álcool) e, em seguida, queima-o, para obter a energia de que precisa para andar.
carro	{carro; vagão}	Acoplado à locomotiva estava o I]carro de passageiros da 1ª classe em madeira de 1914, da antiga São Paulo Railway (SPR), companhia inglesa responsável pela construção da ferrovia e pelo acampamento de trabalhadores que deu origem à vila.
carroça	{carroça}	Um casal de atores conta as peripécias e bravatas de Tonhedo, um solitário andarilho puxador de F]carroça.
carroção	{carroção}	Marta deitou-o em um I]carroção e levou-o até a cidade de Nephi, onde o colocou no trem e levou-o até Provo.
carro-de-combate	{carro-de-combate; tanque}	Este aprendizado propiciou um sonho maior que foi o de conceber um I]carro-de-combate médio totalmente brasileiro.
carro-forte	{carro-forte}	Em um assalto meticulosamente planejado, executado com armamento pesado, uma quadrilha levou todos os valores contidos em um I]carro-forte da empresa Prosegur, com sede em Campinas.
carro-guincho	{carro-guincho; carro-socorro; guincho; reboque}	Em alguns lugares somente um I]carro-guincho autorizado pode rebocar um carro.
carro-leito	{carro-leito; vagão-dormitório; vagão-leito}	Os trens, em geral, possuem vagões confortáveis, além de I]carro-leito e vagão-bar.
carro-madrinha	{carro-madrinha}	O modelo SL será o I]carro-madrinha dos GPs e a perua C, a ambulância.
carro-salão	{carro-salão}	O I]carro-salão é um carro típico do estilo Budd para compor os trens completos.
carro-socorro	{carro-guincho; carro-socorro; guincho; reboque}	Logo chegou um I]carro-socorro de uma oficina e consertou o estrago.

carruagem	{carruagem; coche; sege}	Por volta de 1730, as autoridades de Berna também proibiram as saídas de F]carruagem a partir das 21h, aplicando multas pesadas a quem não respeitasse a lei.
carruagem	{carruagem}	Há quem prefira, ainda, ser conduzida por um dos padrinhos, no próprio carro dele, ou chegar em uma I]carruagem, com belos cavalos à frente.
caterpílar	{caterpílar}	Depois, o I]caterpílar passa e aplaina.
cegonha	{caminhão-cegonha; cegonha}	Coisa rara em lançamento de automóvel: o Toyota Corolla que será lançado na próxima semana já está sendo levado pelas I]cegonhas às concessionárias da marca.
charrete	{charrete}	Tornou-se mascate, vendendo roupas de cama, mesa e banho de porta em porta,com uma I]charrete.
conche	{carruagem; coche; sege}	Os cavaleiros entraram na pista em um I]coche puxado por quatro cavalos.
cupê	{cupê}	O C]cupê estará à venda no mercado europeu este mês, por US$ 150 mil .
diligência	{diligência}	Tem enforcamento, tiroteios, desavenças entre pistoleiros inimigos, duelos, assaltos à I]diligência, enfim, tudo aquilo que estamos acostumados a ver nos velhos filmes de bangue-bangue.
empilhadeira	{empilhadeira}	O operador de F]empilhadeira Celso Froner, 26, da metalúrgica Case, aprova a greve porque já sofreu na pele.
esqueite	{esqueite; skate; skateboard}	As intensas chuvas ocorridas desde o início do ano atrapalharam bastante o andamento das obras de construção da pista de I]esqueite.
fiacre	{fiacre}	Quando Jules Blois dirigia-se em um I]fiacre para o local designado para o combate, os cavalos assustaram-se com a aparição súbita de um vulto e empinaram-se, derrubando por terra, Blois e sua comitiva.
furgão	{van; furgão}	Hoje é viável, por exemplo, um tetraplégico, dependendo do grau da lesão, conduzir um F]furgão.
gôndola	{gôndola}	Ainda neste ano são lançados seis vagões frigoríficos com ref. de 2015 e I]gôndola 2020, além do 1º vagão limpa trilhos 2098, hoje fora de linha.
guincho	{carro-guincho; carro-socorro; guincho; reboque}	O turista recebe orientação de casas de câmbio, hospitais, clínicas médicas, odontológicas, serviço de F]guincho e oficinas mecânicas.
hatch	{hatch}	Esteticamente, o I]hatch é um carro com um visual agressivo e jovem.

jamanta	{carreta; jamanta}	Pouco depois, é uma F]jamanta que viaja pelos ares, antes de explodir junto ao carro dos nossos heróis.
jinriquixá	{jinriquixá}	Num tempo que pareceu eterno, o marido chegou com um I]jinriquixá, uma espécie de carrinho, puxado por um homem.
jipe	{jipe}	Os americanos consideram que o I]jipe foi o verdadeiro herói da Segunda Guerra Mundial.
kart	{kart}	"A sensação é sempre muito boa de correr em um F]kart.", afirmou Prost.
lambreta	{lambreta; motoneta; scooter; vespa}	Não pense, como muita gente, que a I]lambreta grande é a boa.
landau	{landau; landô}	A esperá-los, estava o infante D. Manuel, que viera das Necessidades num I]landau (carruagem aberta) com o visconde de Asseca, o Presidente do Ministério João Franco e elementos do governo e da corte.
landô	{landau; landô}	O A]landô – outro carro de luxo, carruagem de quatro rodas, com dupla capota que se erguia e abaixava - serviria de desembarque da Família Baptista, no Cais Pharoux.
lata	{calhambeque; lata}	Estava vendo a hora daquela I]lata quebrar na subida do pontilhão.
limusine	{limusine}	Dentro da I]limusine, um executivo tira um sarro com o motorista da Kombi.
locomotiva	{locomotiva; locomotora}	Os maquinistas afirmaram que a F]locomotiva poderia ter sido deslocada para o trilho do trem com problema e, dessa forma, evitado que ele se movesse.
locomotora	{locomotiva; locomotora}	Por ser tão antiga quanto uma I]locomotora que se desvia dos trilhos rumo a ribanceira...
magrela	{bicicleta; bike; magrela}	Na hora de comprar uma I]magrela, o melhor é procurar um profissional.
maria-fumaça	{maria-fumaça}	Essa I]maria-fumaça faz, nos finais de semana, o percurso entre Tiradentes e São João Del Rey.
mobilete	{mobilete}	Segundo noticiaram os jornais diários, tudo começou quando dois assaltantes, em uma I]mobilete, entraram na pizzaria e exigiram as chaves do carro de um casal que estava no recinto.
monociclo	{monociclo}	O teatro popular usa tudo que vem à mão, como os pinos que a malabarista maneja com destreza sobre um I]monociclo.
moto	{moto; motoca; motocicleta; motociclo}	O motorista atirava em direção a uma F]moto com dois homens.
motoca	{moto; motoca; motocicleta; motociclo}	Era uma I]motoca muito cara.
motocicleta	{moto; motoca; motocicleta; motociclo}{moto; motoca; motocicleta; motociclo}	Paulo Sanda é empresário e diz que comprou uma F]motocicleta para resolver seu problema com o trânsito de São Paulo.

motociclo	{moto; motoca; motocicleta; motociclo}	Um homem de 55 e uma mulher de 58 anos foram atropelados por um I]motociclo, placa KAJ 4822, na rua dos Cravos, no bairro Jardim das Violetas.
motoneta	{lambreta; motoneta; scooter; vespa}	Alguns vendedores de consórcio ou empresas estariam estimulando a venda de I]motoneta, sob alegação de que, para pilotá-la, não haveria necessidade de se ter Carteira Nacional de Habilitação (CNH) e nem a exigência de ser maior de 18 anos, quando o veículo tem menos de 100 cilindradas.
mountain bike	{mountain bike}	A loja tem uma microfilial no boulevard Geneve, que aluga F]mountain bikes (R$ 6 a hora) e bicicletas de alumínio (R$ 10 a hora).
niveladora	{angledozer; aplanadora; niveladora; patrol; patrola}	Uma I]niveladora estava sendo passada sobre a estrada e a transformou num grande areal de mais de seis quilômetros.
panzer	{panzer}	I]Panzers foram usados em ambos os organismos terrestres que compunham as forças armadas alemãs na Segunda Guerra Mundial: a Waffen SS e a Wehrmacht.
patinete	{patinete}	Fred tentou se matar com um patinete I]infantil de plástico ao descer uma ladeira!
patrol	{angledozer; aplanadora; niveladora; patrol; patrola}	Na mesma ocasião será assinado um convênio para a compra de uma I]patrol que vai trabalhar na manutenção das estradas vicinais do município.
patrola	{angledozer; aplanadora; niveladora; patrol; patrola}	Foi entregue na quarta-feira (20), para a Secretaria do Desenvolvimento Rural mais uma I]patrola, para desempenhar as demandas da Patrulha Agrícola.
perua	{perua; wagon}	Conforme antecipou nosso colunista Fernando Calmon (Alta Roda), além do sedã haverá outros dois produtos derivados, incluindo uma I]wagon (perua) e até um utilitário esporte.
picape	{caminhonete; picape; pickup}	Só a I]picape Chevy 500 mantém carburador.
pickup	{caminhonete; picape; pickup}	Não deixe ninguém ser transportado na caçamba de uma I]pickup, mesmo em uma pickup com caçamba coberta.
pullman	{pullman}	Esquecia-me de relatar que o trem das quatro e meia tinha, entre os seus vagões, um I]pullman e nós estudávamos, com muito interesse, as criaturas estranhas que viajavam nele.
radiopatrulha	{radiopatrulha}	A quadrilha foi surpreendida por uma I]radiopatrulha da Polícia Militar justamente no momento em que saia da casa com os objetos roubados.

reboque	{carro-guincho; carro-socorro; guincho; reboque}	A companhia aérea informou que o A380 estava sendo levado para a pista por um I]reboque, que sofreu uma falha no sistema hidráulico.
roadster	{baratinha; roadster}	A capota retrátil pode ser recolhida por um mecanismo eletro-hidráulico, transformando o cupê em um I]roadster em apenas 16 segundos.
scooter	{lambreta; motoneta; scooter; vespa}	Antes que acabem as férias da Fórmula 1, Rubinho deixou o golfe de lado e resolveu dar uma passeios com uma I]scooter, quando ia rumo aos treinos da F1 em Jerez de la Frontera.
sedã	{sedã}	A Ford aposta tudo no Mondeo, novo I]sedã médio de luxo, carro mundial da marca.
sege	{carruagem; coche; sege}	A pobre louca chega em uma I]sege preta, cortinas corridas para que não a vejam.
skate	{esqueite; skate; skateboard}	Aos dez anos de idade, Ferrugem comprou um F]skate com o dinheiro que o pai deu para abrir uma poupança e não largou mais.
skateboard	{esqueite; skate; skateboard}	Quando tudo parecia perdido, eis que um garoto, provavelmente pula da locomotiva com um I]skateboard do futuro, era o filho do doutor Brown.
stock car	{stock car}	Fizemos uma exibição com um I]stock car durante o horário do almoço, no centro de Curitiba, e isso chamou a atenção de quem passava pelas ruas.
tanque	{carro-de-combate; tanque}	As tropas podem ir a qualquer lugar em um I]tanque.
táxi	{carro de praça; táxi}	No caminho, fiz a mesma pergunta a um motorista de F]táxi .
tênder	{tênder}	Diversas paradas, reabastecimento de água e lenha no I]tênder da locomotiva.
todo-terreno	{todo-terreno}	A Land Rover apresentou novas fotos do I]todo-terreno LRX.
trailer	{trailer}	Um I]trailer de boa qualidade, pequeno (que fosse possível ser rebocado por um Fusca ou similar), sem banheiro, custava na Europa o equivalente ao preço de um Fusca zero quilômetro.
trâmuei	{bonde; trâmuei; tramway}	Também não imaginávamos que seríamos obrigados a fazer chapas de pulmão e proibidos de pegar o I]trâmuei por causa dos tuberculosos que "andavam à solta" pela cidade, ou que iríamos encontrar tantos alemães "ex-nazistas"(dizia-se) buscando refúgio na "suíça brasileira".

tramway	{bonde; trâmuei; tramway}	A história do I]tramway do Guarujá se confunde com a própria história da cidade.
trator de lâmina	{buldôzer; trator de lâmina}	O I]trator de lâmina ou motoniveladora procederá ao espalhamento do material.
trator	{trator}	A instalação custou US$ 75 mil, levou três meses para ser realizada, durou 10 semanas e foi destruída por um I]trator em 58 minutos.
triciclo	{triciclo; velocípede}	Em uma delas vemos crianças escravas catando café sob o olhar de uma criança branca, com seu F]triciclo e vestes elaboradas.
tróica	{tróica}	Em frente ao casebre, num frio medonho, passava uma I]tróica puxada por cavalos com neve pelas canelas carregando o corpo de Vladmir Lenin.
trole	{trole}	Os I]troles, carruagens e similiares eram os "carros particulares" dos abastados proprietários de fazendas, nos áureos tempos do café.
vagão dormitório	{carro-leito; vagão-dormitório; vagão-leito}	Segundo o chefe dos bombeiros de Bourbonnais (cidade de 14 mil habitantes), Mike Harshbarger, a maioria dos mortos estava no F]vagão-dormitório, que se partiu em dois e pegou fogo.
vagão	{carro; vagão}	Alunos estudam em F]vagão dos anos 20 reformado e instalado em fazenda.
vagão-frigorífico	{vagão-frigorífico}	Scully chega assim que eles partem, dirigindo-se ao I]vagão-frigorífico onde Mulder tinha sido visto pela última vez e encontrando só destroços do incêndio.
vagão-leito	{carro-leito; vagão-dormitório; vagão-leito}	Do I]vagão-leito foram retirados colchões e cobertores.
vagão-plataforma	{vagão-plataforma}	O transporte de carretas sobre I]vagão-plataforma combina a economia dos trilhos com o serviço porta-a-porta.
vagão-restaurante	{vagão-restaurante}	No I]vagão-restaurante reina um vazio.
vagão-tanque	{vagão-tanque}	A suposta colisão entre o I]vagão-tanque e o caminhão provocou o vazamento de 20 mil litros de gasolina em um córrego que passa pela Unidade de Produção da ALL, localizada em Canoas, além de óleo diesel na vegetação.
vagoneta	{vagoneta}	Neste forno, quando uma I]vagoneta com o lote de peças cerâmicas está chegando ao final do ciclo, outra vagoneta com uma quantidade igual está sendo iniciada.
vagoneta	{vagoneta}	Os visitantes são acomodados em uma I]vagoneta - o mesmo veículo utilizado pelos mineradores desde a fundação da mina, no século passado.

van	{van; furgão}	A equipe se preparava para entrar em uma I]van quando foi abordada por um suposto bandido que estava com uma pistola dourada.
velocípede	{triciclo; velocípede}	Em um I]velocípede de criança, viajou através do Japão entre 1985 e 1986.
velocípede	{velocípede}	Após numerosas tentativas foi em 12 de Julho de 1817 que se iniciou a história dos I]velocípedes e do ciclismo.
vespa	{lambreta; motoneta; scooter; vespa}	Um certo dia, uma garota estranha surge em sua vida montada em uma I]vespa, acertando sua cabeça em cheio com uma guitarra.
wagon	{perua; wagon}	A Ford apresentou as primeiras imagens oficiais da I]perua Focus Wagon totalmente remodelada, assim como o Focus.

APÊNDICE 2: Os sintagmas livres recorrentes (SLRs) do PB e suas respectivas frases-exemplo.

SLRs	***Phrasets***	**Frases-exemplo**
4x4	{veículo com tração nas quatro rodas; 4x4}	Com um I]4x4 e um GPS, é possível explorar os mais novos e belos destinos Eco turisticos no Brasil.
biblioteca ambulante	{biblioteca ambulante}	Para isso, uma Kombi foi transformada em uma I]biblioteca ambulante, com aproximadamente 1600 livros.
bicicleta motorizada	{bicicleta motorizada}	Nícolas Borges lê o JC desde 1982 e prega o prazer de guiar, de uma I]bicicleta motorizada a um superesportivo.
caminhão basculante	{caminhão basculante}	Marcos Alberto de Souza, 19, sofreu traumatismo craniano e facial ao ser atropelado por um I]caminhão basculante, às 11h20 de ontem, na avenida do Café, zona oeste de Ribeirão.
caminhão cegonheiro	{caminhão cegonheiro}	Com os preços do carro zero no Brasil, deviam mudar o nove de I]caminhão cegonheiro para caminhão-sem-vergonheiro.
caminhão de bombeiro	{caminhão de bombeiro; carro de bombeiro}	Mulher tenta roubar I]caminhão de bombeiros seminua.
caminhão de carga	{caminhão de carga}	I]Caminhão de carga bate em poste e deixa grandes estragos.
caminhão de entregas	{caminhão de entregas}	Um I]caminhão de entregas a domicílio, cujo motorista provavelmente sofreu um infarto, invadiu um cortejo fúnebre nesta quinta-feira na Alemanha, matando três.
caminhão de lixo	{caminhão de lixo}	Os contratos que documentam as licitações foram apreendidos pelo Ministério Público, na noite de anteontem, dentro de um F]caminhão de lixo.
caminhão de mudança	{caminhão de mudança}	I]Caminhão de mudança deve ficar fora da regra de carga e descarga, diz Kassab. da Folha Online.
caminhão de som	{caminhão de som}	O F]caminhão de som dirigiu-se à praça central e logo começaram os discursos.
carrinho berço	{carrinho berço}	É um I]carrinho berço e passeio com design inovador, assento acolchoado e várias funções.
carrinho de bebê	{carrinho de bebê}	Chico fez questão de empurrar o I] carinho de bebê da netinha.
carrinho de chá	{carrinho de chá}	O F]carrinho de chá, em madeira pré-tratada, é da Etel Carmona Interiores.

carrinho de compras	{carrinho de compras}	Meio que sem querer, olhei para a imagem refletida que passava empurrando um F] carrinho de compras.
carrinho de sobremesas	{carrinho de sobremesas}	O I]carrinho de sobremesas circula entre as mesas com delícias como doces de frutas regionais.
carro aberto	{carro aberto}	Parece que podemos comparar a visão de mundo de Faulkner à de um homem sentado num F]carro aberto e que olha para trás.
carro beberrão	{carro beberrão; carro gastão}	O economista Rafael Pachoarelli propõe trocar o I]carro beberrão por um mais econômico, jamais levar as crianças ao supermercado e restringir-se à lista de compras essenciais.
carro compacto	{compacto; carro compacto}	A Mercedes-Benz deve instalar a linha de montagem do Swatch, C]carro compacto com design do fabricante dos relógios da marca, em Sarreguemines, cidade do norte da França .
carro conversível	{conversível; carro conversível}	Modelos lançam coleção de lingerie dentro de I] carro conversível.
carro de bombeiro	{caminhão de bombeiro; carro de bombeiro}	Às 16h, todo o grupo inicia o desfile em I]carro de bombeiro que percorrerá as principais ruas de Brusque.
carro de passageiros	{carro de passageiros; vagão de passageiros}	O recurso inicialmente alocado para essa mudança seria de um novo I]vagão de passageiros, devidamente preparado para esse fim.
carro de polícia	{carro de polícia}	Em Sydney, mulher bate em I]carro de polícia estacionado em frente à delegacia.
carro de presos	{carro de presos}	Em 85 minutos de projeção, são 12 choques com a polícia e 5 viagens no I]carro de presos, para a prisão (mais uma na ambulância).
carro de segunda mão	{carro usado; carro de segunda mão; carro seminovo}	É possível comprar um I]carro de segunda mão com a ajuda de um empréstimo.
carro de socorro	{carro de socorro}	O I]carro de socorro empurra o Volks, mas ele só vai até a metade do barranco.
carro envenenado	{carro envenenado}	Marcelo D2 exibe seu I]carro envenenado em festa.
carro esportivo	{carro esportivo}	Essas alterações, porém, não transformaram o Marea em um I]carro esportivo, de comportamento nervoso.
carro gastão	{carro beberrão; carro gastão}	A fama de I]carro gastão, de manutenção cara e com índice de depreciação alto "colou" nos Fiat Marea desde seu lançamento em 1998.

carro reserva	{carro reserva}	Você também pode pedir o I]carro reserva quando o seu. veículo estiver sendo consertado em outra seguradora.
carro seminovo	{carro usado; carro de segunda mão; carro seminovo}	IPVA de F]carros seminovos vai cair pelo terceiro ano.
carro subcompacto	{subcompacto; carro subcompacto}	De acordo com a Forbes, o I]carro subcompacto da Honda é vendido nos Estados Unidos por 15 mil dólares.
carro usado	{carro usado; carro de segunda mão; carro seminovo}	No momento de se adquirir um I]carro usado é importante ver as condições no mesmo.
carroça de leite	{carroça de leite}	Passa todo dia em frente de casa a I]carroça de leite.
casa sobre rodas	{casa sobre rodas}	Mas se o dinheiro está sobrando — e o apartamento já existe —, invista na I]casa sobre rodas.
compacto	{compacto; carro compacto}	A Daihatsu, a sexta maior montadora de veículos do Japão, chega ao mercado brasileiro, com o lançamento de três modelos: o C]compacto Charade, o sedã Applause e o jipe Feroza.
conversível	{conversível; carro conversível}	De forma semelhante, um sedã emerge como mais prototípico do que um I] conversível ou uma limusine.
diligência postal	{diligência postal}	Em Suez, deixou o navio e tornou a I] diligência postal, mantida pela antiga Companhia das Índias Orientais, para conduzir seus oficiais ao Cairo e dali ao Mediterrâneo, onde tomavam a embarcação.
dois lugares	{dois lugares; dois-lugares}	O I]dois lugares em carroceria de alumínio, viria revolucionar o conceito de carros esportivos e abrir a "briga" com os tradicionais europeus.
dois-lugares	{dois lugares; dois-lugares}	Pode ser para os padrões brasileiros, mas o I]dois-lugares faz um bocado de sucesso na Europa e ficou mundialmente conhecido com o filme “O Código Da Vinci”.
locomotiva a diesel	{locomotiva a diesel}	Você simplesmente não sobe na cabine, liga a chave e sai dirigindo uma I] locomotiva a diesel.
locomotiva a vapor	{locomotiva a vapor}	Outros empreendimentos marcaram seu reinado: a primeira I]locomotiva a vapor; a instalação do cabo submarino entre o Brasil e a Europa, a inauguração do telefone e a instituição do selo postal.

locomotiva de manobras	{locomotiva de manobras}	O motor diesel foi empregado, pela primeira vez, nas ferrovias, em 1925, numa I]locomotiva de manobras da Central Railroad de Nova Jersey.
locomotiva elétrica	{locomotiva elétrica}	A I]locomotiva elétrica "Metropolitan Vickers" que se encontrava, desde 1.980, estacionada na empresa SOMA em Sumaré, foi transportada para a sede da Fundação, em São Paulo, em junho de 2.002.
locomotiva tênder	{locomotiva tênder}	Para os trajetos curtos , colocavam-se as reservas sobre o próprio corpo da locomotiva, que era então chamada I]locomotiva tênder.
máquina de erraplanagem	{máquina de terraplanagem}	Eles foram pegos enquanto negociavam a compra de uma I]máquina de terraplanagem.
minicarro	{minicarro}	O Chevrolet Trax é um dos três I]minicarros que serão expostos pela General Motors no Salão de Nova York.
minivan	{minivan}	Toyota Vanguard é um modelo esportivo com capacidade para cinco passageiros, podendo ser transformado em uma I]minivan com sete lugares.
peça antitanque	{peça antitanque}	Aliás, ficou famoso o uso dos Flak 18 e 36 como I] peça antitanque, em episódios da luta na França em 1940 e na África do Norte, entre outros.
Stanley Steamer	{Stanley Steamer}	A Stanley Motor Carriage Company operou (1902-1917) e os carros feitos pela companhia eram chamados de I]Stanley Steamers.
subcompacto	{subcompacto; carro subcompacto}	A Mitsubishi está planejando o lançamento de um subcompacto para mercados estratégicos da marca.
tanque de guerra	{tanque de guerra}	Um soldado russo bêbado em um I]tanque de guerra destrói uma casa e quase acaba com uma loja.
trator florestal	{trator florestal}	O arraste por veículos mais rápidos como o I]trator florestal requer uma equipe de três pessoas: um tratorista, um ajudante no pátio (faz o desengate das toras) e outro ajudante no interior da floresta (procura e enlaça as toras).
utilitário esportivo	{utilitário esportivo; utilitário-esportivo; veículo utilitário esportivo}	Baseado em um I]utilitário esportivo, o Cruise Crosser, apesar de ser uma picape, comporta três fileiras de assentos, algo não muito comum no mercado automobilístico.

utilitário-esportivo	{utilitário esportivo; utilitário-esportivo; veículo utilitário esportivo}	Uma delas já se sabia que seria produzida: um I]utilitário-esportivo de medidas compactas.
vagão carvoeiro	{vagão carvoeiro}	Ouviu a conversa ruidosa dos trabalhadores do I]vagão carvoeiro.
vagão de carga	{vagão de carga; vagão de mercadorias}	A Linha do Norte ficou cortada hoje de madrugada devido ao descarrilame nto de um I]vagão de mercadorias, num incidente que não fez quaisquer feridos.
vagão de gado	{vagão de gado}	A jornalista israelense Amira Hass descreve o momento em que sua mãe, Hannah, marchava de um I]vagão de gado até o campo de concentração nazi de Bergen-Belsen.
vagão de mercadorias	{vagão de carga; vagão de mercadorias}	Um I]vagão de carga de trem, vazio, pesa 10000kg* e se desloca com uma velocidade de 1,20m/s.
vagão de passageiros	{vagão de passageiros}	No local, além de objetos e ferramentas usadas na operação e manutenção da ferrovia, encontram-se preservadas 2 locomotivas à vapor e um I]carro de passageiros.
vagão para fumantes	{vagão para fumantes}	Em trens, se antes havia I]vagão para fumantes, agora não há mais.
vagão-bagageiro	{vagão-bagageiro}	"História, Herança e Esperança," uma exibição permanente do Holocausto, que inclui um I]vagão-bagageiro original da Polônia ocupada pelos Nazis, está no terceiro andar.
vagão-gaiola	{vagão-gaiola}	Um dia chegou, durante o dia, uma locomotiva puxando um I]vagão-gaiola para buscar um touro.
veículo a motor	{veículo motorizado; veículo a motor}	Para conduzir um I]veículo a motor na via pública, é necessário estar legalmente habilitado para o efeito.
veículo anfíbio	{veículo anfíbio}	A última da empresa foi ter quebrado o recorde mundial de travessia do Canal da Mancha em um veículo anfíbio com hidrofólios.
veículo autopropulsad	{veículo autopropulsado; veículo autopropulsionado}	Este I]veículo autopropulsado está concebido para o transporte de cargas, bem como para elevar cargas e pessoas.
veículo autopropulsionado	{veículo autopropulsado; veículo autopropulsionado}	Benz obteve, em janeiro de 1886, uma patente para o primeiro I]veículo autopropulsionado.

veículo blindado	{veículo blindado}	O Decreto Federal nº 3665 e Portaria 003/2001, publicada em 13/03/2001, obrigam todas as pessoas físicas e jurídicas, proprietárias de I]veículos blindados a obterem os registros necessários junto ao Exército.
veículo com roda	{veículo com roda}	Foram os sumérios, que habitavam o sul da Mesopotâmia, que construíram o primeiro I]veículo com roda.
veículo com tração nas quatro rodas	{veículo com tração nas quatro rodas; 4x4}	Em outras palavras, I]veículo com tração nas quatro rodas é um veículo que recebe a força exercida pelo motor tanto nas rodas dianteiras, quanto nas rodas trazeiras.
veículo de reconhecimento	{veículo de reconhecimento}	Desde o inicio da Primeira Guerra Mundial, o Exército norte-americano estava procurando por um I]veículo de reconhecimento para todo terreno, rápido e leve.
veículo motorizado	{veículo motorizado; veículo a motor}	Não use fones de ouvido quando estiver dirigindo, andando de bicicleta ou operando qualquer I]veículo motorizado.
veículo recreativ	{veículo recreativo}	Alunos da faculdade de engenharia da Associação Educacional Dom Bosco, de Resende, projetaram e construíram um I]veículo recreativo, chamado de AEDBaja.
veículo sacolejante	{veículo sacolejante}	Pegamos exatamente o mesmo I] veículo sacolejante, com as mesmas janelas quebradas e as mesmas poltronas esfarrapadas, para mais três horas e meia de estrada.
veículo utilitário esportivo	{utilitário esportivo; utilitário-esportivo; veículo utilitário esportivo}	Pense em um I]veículo utilitário esportivo: para dirigir seu Touareg na estrada, não é preciso acionar o sistema de tração nas quatro rodas.