Base de Conhecimento Léxico-Ontológico para o Português do Brasil: uma proposta de modelo

Claudia Zavaglia[1]

[1]Universidade Estadual Paulista – UNESP/IBILCE – Câmpus de São José do Rio Preto

{zavaglia@lem.ibilce.unesp.br; czavaglia@terra.com.br}

Abstract. *This paper is an attempt to describe the construction of a Lexical Knowledge Base (LKB) from homograph names for Brazilian Portuguese, the lexical items of which occur linked by different types of semantic relations.*

Resumo. *Este trabalho objetiva descrever a elaboração de uma Base de Conhecimento Lexical (BCL), a partir de nomes homógrafos para o português do Brasil, cujos itens lexicais encontram-se vinculados por diferentes tipos de relações semânticas.*

1. Introdução

O léxico é um dos Recursos Lingüísticos primários no que diz respeito à Engenharia da Linguagem. De fato, qualquer sistema aplicativo, para analisar ou processar uma língua natural, não pode prescindir do léxico. Este, por sua vez, para que seja utilizado por uma máquina, deve conter informações adequadas e codificadas para que o programa computacional ou o algoritmo possa decodificá-las e utilizá-las. As informações contidas em um léxico podem ser de vários níveis lingüísticos (morfológico, sintático, semântico) sendo que para cada um deles existe um tipo de codificação de dados em diferentes etapas.

Para o Processamento de Língua Natural (PLN), a elaboração de recursos lexicais que contenham informações semânticas faz-se importante para sistemas que tratem da desambiguação dos sentidos das palavras, como por exemplo, a Tradução Automática, a Recuperação de Informação, Sistemas de Busca, entre outros. A semântica é capaz de resolver muitos casos de homografia na linguagem falada e escrita. Tendo em vista a pragmática do discurso e o seu poder de desambiguação, a ambigüidade gerada pelos homônimos na fala é satisfatoriamente resolvida. Ao contrário, em um contexto de escrita, a ambigüidade é um dos grandes inimigos da interpretação correta de um texto. O homem, enquanto falante de uma língua, possui intuições interpretativas que o levam a resolver certas ambigüidades de uma língua natural de forma até mesmo inconsciente. Inversamente, o computador não possui tais intuições e um dos maiores desafios dos lingüistas computacionais é justamente esse, ou seja, tentar transportar para a máquina os mesmos mecanismos de interpretação desambiguadora próprios dos seres humanos.

2. Objetivos

O modelo de representação aqui proposto contém informações do tipo semânticas e morfossintáticas. Essas últimas restringiram-se à classe gramatical, ao

gênero e ao número das palavras[1]. Em contrapartida, privilegiamos o tipo de informação que diz respeito ao significado, introduzindo uma série de relações semânticas entre as palavras que têm o escopo de, justamente, resgatar de forma minuciosa o significado de cada item lexical em questão.

Nos mesmos moldes de SIMPLE (Lenci, 1999) e ItalWordNet (ITC-irst, 2000) (e suas antecessoras WorNet e EuroWordNet), em que se procurou esquematizar por meio de correlações cada hipônimo ao seu hiperônimo (e vice-versa) gerando, assim, um sistema de hereditariedade do tipo lexical, realizamos um esforço de individualizar os hipônimos e os hiperônimos de um exemplário de formas homônimas freqüentes[2] com o intuito de estabelecer um sistema de hereditariedade semântica. Por conseguinte, um item homônimo é identificado, caracterizado e desambiguado a partir das características que herda de seu hipônimo (ou das outras relações semânticas com as quais mantém ligação) que, por sua vez, herda de seu hiperônimo. O modelo semântico que ora expomos não pretende definir de modo direto o significado de um homônimo, ao contrário; tenciona sugerir o significado para cada item homógrafo, bem como para suas ocorrências polissêmicas, por meio de itens léxicos interligados a cada uma das formas homônimas, que têm por escopo delimitar o seu campo significativo.

3. A Estrutura Qualia

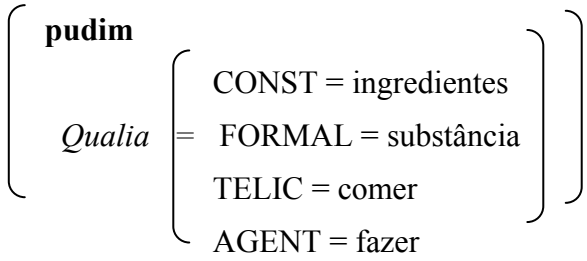
Dada a suposição de que múltiplas dimensões do significado são necessárias para começar a caracterizar unidades lexicais em um nível semântico, a Estrutura Qualia (Pustejovsky, 1995) tem sido utilizada como um dos princípios cruciais de organização para a representação e interpretação do significado lexical de uma frase em sistemas computacionais de complexidade variada. De fato, ela é capaz de suprir o vocabulário básico para expressar aspectos diferentes do significado lexical (word-meaning). A Estrutura Qualia especifica quatro papéis essenciais do significado de uma palavra: **Constitutivo ou Partes Constituintes** (*Constitutive*), i.e., aquele que exprime a relação entre um objeto e suas partes constituintes; **Formal** (*Formal*), ou seja, aquele que distingue o objeto em um domínio mais amplo; **Télico** (*Telic*), aquele que expressa o objetivo/escopo e a função do objeto; **Agentivo** (*Agentive*), i.e., aquele que considera fatores envolvidos na origem do objeto.

Os quatro papéis essenciais da Estrutura Qualia representam as dimensões múltiplas do significado lexical. Com efeito, Qualia é a estrutura representacional para expressar partes do aspecto componencial do significado lexical, na medida em que resgata ou captura diferentes graus de complexidade entre itens lexicais. Ademais, sustenta um conjunto de inferências disponível para default, quer dizer, essas inferências têm de ser usadas de modo geral, como se fossem um padrão a ser seguido. Vejamos o exemplo do item lexical "pudim":

---

[1] Fato esse justificado pela existência de léxicos computacionais para o português do Brasil que possuem informações morfossintáticas bastante detalhadas (Cf. Nunes et al., 1996).

[2] Os homônimos foram extraídos do *Dicionário de Freqüências do Português Contemporâneo*, de Maria Tereza Camargo Biderman, versão em disquetes, 1997.



Cada um dos quatro papéis Qualia é representado como uma relação que está em alternância com o topo da hierarquia de outras relações específicas, representando os subtipos de informação de um dado papel. Essa hierarquia nos quatro papéis Qualia é chamada de Conjunto de Qualia Ampliado (*Extended Qualia Set*). Por conseguinte, para cada um dos quatro papéis Qualia foi especificado um Conjunto de Qualia Ampliado, ou seja, foram especificados subtipos para cada um desses papéis (Lenci, 1999).

Em nosso modelo de representação[3], para que fosse possível resgatar as dimensões do significado de um item homônimo, lançamos mão de uma codificação de base relacional, a partir das possibilidades decomposicionais que nos oferece a noção da Estrutura *Qualia* de Pustejovsky (1995) e da Estrutura *Qualia* Ampliada de SIMPLE (Cf. Lenci, 1999). Desse modo, a ambigüidade semântica entre formas homônimas é tratada por meio de papéis formais, constitutivos, télicos e agentivos, de acordo com a informação lingüística que cada unidade homônima carrega consigo. Por meio da caracterização das informações nesses quatro tipos de papéis, o significado da **FORMA[1]** ou **FORMA[2]** ou **FORMA[3]** é recuperado de forma desambiguada. Além disso, a relação semântica que o item homônimo mantém com um outro item léxico oferece indícios para a sua desambiguação. E ainda, a categorização em uma base ontológica[4] é capaz, ainda de suprir eventuais ambigüidades que o conceito do item homônimo possa gerar, dependendo do contexto no qual encontrar-se-á inserido.

Cada papel da Estrutura *Qualia* possui as seguintes relações semânticas no Conjunto de *Qualia* Ampliado:

Tabela 1. Relações Semânticas

FORMAL <é_um>; <é_um_sinônimo>; <é_um_antônimo>
CONSTITUTIVO <é_um_membro_de>; <contém>; < quantifica>; < vive_em>;
<atividade_constitutiva>; < está_em>; < tem_como_cor>; < tem_como_membro>;
< feito_de>; < produzido_por>; < é_parte_de>; < propriedade_de>; < medido_por>
TÉLICO <é_uma_atividade_de>; < objeto_da_atividade>; < é_a_habilidade_de>;
< usado_para>; < usado por>; < destinado_a>; < usado_contra>.

[3] Parte desse trabalho foi desenvolvido no *Istituto di Linguistica Computazionale di Pisa (ILC)*, sob a orientação de Nicoletta Calzolari, por ocasião de nosso estágio sanduíche com bolsa de estudos oferecida pelo CNPq, de janeiro a abril de 2000.

[4] Trabalho esse realizado por ocasião de nossa tese de doutoramento (Zavaglia, 2002).

AGENTIVO <experiência_agentiva>; <resultado_de>; <origem>; <derivado_de>

4. Ontologias

Para Gruber (1993), ontologias compartilham e reutilizam o conhecimento de mundo. Com efeito, segundo o autor: "o termo ontologia significa uma especificação de conceitos, isto é, uma ontologia é uma descrição formal dos conceitos e das relações existentes entre estes em um determinado domínio" (*apud* Braga et al., 2002).

Segundo Ortiz (2000:2), a semântica baseada em ontologia em PLN serve:

a) de suporte para a tradução de lacunas léxicas;
b) de suporte para a desambiguação, tanto léxica como estrutural;
c) para um tratamento adequado do fenômeno da sinonímia.

Em consonância, Tiscornia (1995:1) diz que para o desenvolvimento de aplicativos computacionais é necessária a individualização dos modelos dos mecanismos cognitivos humanos e do processo de formação do conhecimento, e que a ontologia formal, uma das mais recentes abordagens da modulação do conhecimento, é, na verdade, uma revisitação de teorias filosóficas e lingüísticas. Nesse sentido, as categorias ontológicas são "subdivisões de um sistema de classificação utilizadas para catalogar conhecimento, por exemplo, em uma base de dados" (Tiscornia, 1995:4).

Ressaltamos que, atualmente, no campo do PLN, principalmente em Sistemas de Bases de Conhecimento Lexical, é consensual que a inclusão desse tipo de repositório semântico, i.e., do tipo ontológico para a representação do significado, é essencial. Existe a necessidade de se oferecer de forma estruturada e organizada um léxico comum utilizado em conformidade por uma determinada comunidade. O uso de ontologias tem sido amplamente empregado em representações do conhecimento de domínios restritos, máxime para sistemas de busca de informação e indexação de documentos, onde a sua aplicação pode ser mais eficaz por tratar, justamente, de conjuntos léxicos de número finito. Em uma Base de Conhecimento Lexical – BCL, por exemplo, o uso de uma ontologia pode servir como recurso de apoio à informação contida no repositório lexical dessa base para ser possível resgatar o significado de um item léxico de forma unívoca. De fato, os recursos lingüístico-classificatórios que a utilização de uma ontologia pode oferecer para um lingüista e/ou lexicógrafo servem para que ele possa dar conta de individualizar univocamente, dentre os diversos significados ou diversas acepções atribuíveis a um mesmo item lexical, o significado pertinente no interior do feixe de sentidos polissêmicos que a palavra comporta, neutralizando, dessa maneira, a polissemia própria a esse mesmo item lexical. Vejamos a seguir, como, de fato, o uso de uma ontologia pode servir como recurso de desambiguação de formas homônimas:

mercúrio [0_1 / 0_2a / 0_2b]

"mercúrio$0_1": Elemento químico metálico, obtido de um mineral vermelho e brilhante

Tipo:	[Substância Natural]
Supertipo:	[Substância]
Domínio:	*Química*

"mercúrio$0_2a": Um dos planetas do sistema solar

Tipo:	[Objeto Natural]
Supertipo:	[Entidade Concreta]
Domínio:	*Astronomia*

"mercúrio$0_2b": Deus da indústria e do comércio na religião romana

Tipo:	[Ideologia]
Supertipo:	[Humano]
Domínio:	*Mitologia*

vale [0_1/0_2]

"vale$0_1": Espaço alongado de terra entre montanhas, montes

Tipo:	[Localização em 3 D]
Supertipo:	[Localização]
Domínio:	*Geografia*

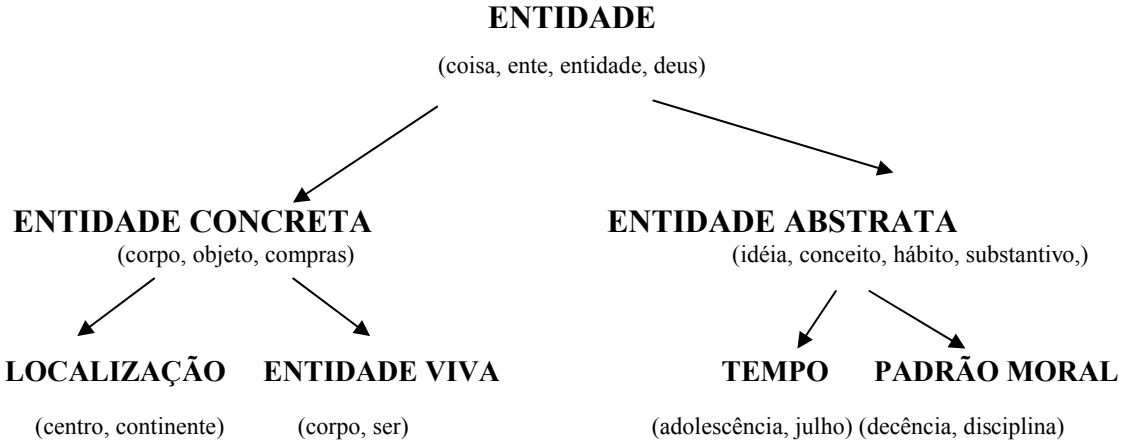
"vale$0_2": Documento sem valor legal que comprova retirada ou adiantamento de dinheiro

Tipo:	[Dinheiro]
Supertipo:	[Manufaturado]
Domínio:	*Finanças*

Como pudemos observar, para cada acepção das formas homônimas acima descritas, realizamos uma representação ontológica diversa, fato esse que auxilia a desambiguação semântica das mesmas.

Uma Ontologia possui categorias que buscam catalogar o conhecimento de mundo em símbolos de uma língua natural, isto é, as palavras. Essas categorias são também chamadas de classes, relações ou tipos. Toda categoria reúne um conjunto de objetos com natureza e propriedades comuns; as categorias têm por princípio e escopo organizar e tornar mais clara e simples a base de conhecimento.

A taxonomia mais comum de uma ontologia é do tipo hereditária em que classes e sub-classes mantêm relações hierárquicas em forma de árvores:



Por meio desse exemplo de categorias arbóreas, a taxonomia hierárquica se verifica a partir do momento em que temos axiomas do tipo:

(1) Todo animal terrestre é um animal, que por sua vez, é uma entidade viva, uma entidade concreta e uma entidade: *um cachorro é um animal, ser vivo e concreto*.

(2) Toda área é uma localização, que por sua vez, é uma entidade concreta e é uma entidade: *uma praia localiza-se em uma cidade que é uma entidade concreta*.

Os membros de uma mesma categoria ou sub-categoria carregam algumas propriedades em comum: na sub-categoria "animal terrestre", por exemplo, seus membros "boi", cachorro", "coelho" possuem patas, andam, não falam; propriedades em comum são, portanto, herdadas pela inserção de uma palavra em uma ou em outra categoria.

5. Base de Conhecimento Lexical

De posse de todas as informações[5] que julgamos necessárias para a construção do paradigma da nossa Base de Conhecimento Lexical (doravante BCL), a saber: (i) **informação ontológica** (subdividida em Tipo, que corresponde ao hipônimo; Supertipo, que corresponde ao hiperônimo e Domínio); (ii) **informação Qualia** (papéis Formal, Agentivo, Télico e Constitutivo); (iii) **informação morfossintática** (Rep_PDD, i.e., Representação das partes do discurso e Rep_Morf. i.e., Representação morfológica); (iv) **informação definicional**, i.e., a definição extraída do dicionário de base, representada por Glossário; (v) **informação pragmática**, i.e., a contextualização do uso do item homônimo, representada por Exemplo; permitimo-nos legitimar o seguinte modelo de BCL, que ora visualizamos por meio do exemplo da forma homônima *banco*:

Tabela 2. Entrada da BCL

BANCO	
HomoU[6]:	"banco$0_1[7]"
SemU[8]:	<banco>
Tipo:	[Mobília]
Supertipo:	[Manufaturado]
Domínio:	*Móveis (Mobiliaria)*
Formal:	*é_um*(<banco>,<objeto>)

[5] Neste artigo, tratamos somente das informações concernentes à Estrutura *Qualia* e à Ontologia. As outras informações encontram-se descritas e discutidas em nossa tese de doutoramento.

[6] HomoU = Unidade Homônima.

[7] Essa simbologia indica que "banco" é um forma homógrafa (representado por "0") e esta é a sua primeira forma (representado por "1").

[8] SemU = Unidade Semântica.

Agentivo:	<Nil[9]>
Constitutivo:	*feito_de*(<banco>,<pedra>) *feito_de*(<banco>,<madeira>) *é_parte_de*(<banco>,<mobília>)
Télico:	*usado_para*(<banco>,<sentar>)
Glossário[10]:	Objeto alongado, com ou sem encosto, em que várias pessoas podem assentar-se
Exemplo[11]:	*Não sei se por causa do vinho, quando me larguei, ou me largaram no **banco** traseiro do carro, pareceu-me ver, sentado na calçada, meu superego arrancando os cabelos* (CP)
Rep_PDD:	NOME
Rep_Morfo:	MASC SING

⇨[12]

HomoU:	"banco$0_2[13]"
SemU:	<banco>
Tipo:	[Local Construído]
Supertipo:	[Localização]
Domínio:	*Sistema Bancário*
Formal:	*é_um*(<banco>,<empresa>)
Agentivo:	<Nil>
Constitutivo:	*está_em*(<banco>,<cidade>)
Télico:	*usado_para*(<banco>,<depositar_dinheiro>) *usado_para*(<banco>,<emprestar_dinheiro>)
Glossário:	Empresa financeira que opera com dinheiro, títulos e outros valores, onde se deposita dinheiro e que pode emprestar dinheiro
Exemplo:	*Dessa vez desceu um senhor engravatado, coisa difícil por ali, com ares de gerente de **banco*** (CP)

[9] O símbolo <Nil> é usado quando o elemento não sofre variação na composição.
[10] O Dicionário usado como base para este trabalho foi o *Dicionário Didático de Português* de Maria Tereza Camargo Biderman. São Paulo: Editora Ática, 1998.
[11] A contextualização da forma homônima foi extraída do *Corpus* Principal (CP) existente no Centro de Estudos Lexicográficos da UNESP de Araraquara.
[12] Essa flecha indica que as duas tabelas encontram-se correlacionadas.
[13] Essa simbologia indica que "banco" é um forma homógrafa (representado por "0") e esta é a sua segunda forma (representado por "2").

Rep_PDD:	NOME
Rep_Morfo:	MASC SING

Esse modelo foi implementado computacionalmente pelo *Núcleo Interinstitucional de Lingüística Computacional* (NILC) da USP/São Carlos para cerca de 200 formas homógrafas, cujas interfaces dizem respeito a cinco módulos de representação, a saber: Módulo Lexical, Módulo Morfossintático, Módulo Desambiguação, Módulo Ontológico e Módulo Estrutura *Qualia*. Todos esses módulos estão correlacionados de modo que as informações neles contidas possam ser vinculadas e interconectadas, dependendo do tipo de pesquisa/busca que o usuário pretenda realizar junto ao sistema.

O escopo desse trabalho, ou seja, a versão computacional, foi incitado essencialmente, por dois motivos: (i) o fato de podermos demonstrar que a efetivação de nossa proposta poderia ser real e que, ao contrário, não estaria fadada ao mundo "virtual". Por conseqüinte, convalidamos as análises lingüísticas que realizamos para a construção da base lingüística cujas entradas são itens homônimos, uma vez que elas foram capazes de sustentar uma implementação computacional; (ii) o fato de podermos demonstrar as vantagens de termos informações de natureza diversificada armazenadas em uma base de dados eletrônica. Dentre essas utilidades destacamos: (i) a recuperação veloz de informações lingüísticas variadas sobre itens homônimos; (ii) a realização de buscas especializadas de certas informações lingüísticas, por meio da geração automática de listas, que poderão servir a diversos tipos de pesquisa; (iii) a possibilidade potencial de utilização das informações lingüísticas contidas no repositório lexical da BCL para aplicações em Sistemas de Processamento de Línguas Naturais, em Motores de Busca, Parsers Semânticos, Desambiguadores, Tradução Automática, Taggers, etc. Com efeito, o fato de termos incluído uma gama variada de informações lingüísticas de natureza pluridimensional (lexical, morfossintática, ontológica, *qualia*, desambiguadora) permite prever uma aplicação variada.

Como trabalho futuro, iniciamos a elaboração de uma *Base Léxico-Ontológica Computacional (português) do Subdomínio da Ecologia – BLOC-Eco,* a partir dos resultados alcançados em Zavaglia (2002), que prevê, além da estruturação de uma Ontologia Específica para a Ecologia, a construção de um *corpus* especial para a extração dos termos que serão distribuídos nas categorias e subcategorias da ontologia por meio de relações semânticas.

Referências Bibliográficas

Braga, J. L.; Torres, K. S.; Botelho, F.C. (2002) "Reengenharia e Visualização de Conceitos no WordNet". Universidade Federal de Viçosa, http://www.sbc.org.br/reic/edicoes/2002e2/cientificos/ReengenhariaEVisualizacaoDeConceitosNoWordNet.pdf.

Gruber, T. R. (1993) "Toward principles for the design of ontologies used for knowledge sharing". Presented at the Padua workshop on Formal Ontology, March 1993, to appear in na edited collection by Nicola Guarino, http://kslweb.stanford.edu/KSL_Abstracts/KSL-93-04.html.

ITC-IRST (2000). “ItalWordNet: Rete semantico-lessicale per l’italiano”. Trento: Consorzio Pisa Ricerche, Istituto per la Ricerca Scientifica e Tecnologica (ITC-irst).

Lenci, Alessandro et. al.(1999) “SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica”. Linguistic Specifications - Deliverable D2.1. University of Pisa and Institute of Computational Linguistics of CNR, Pisa, December.

Nunes, M.G.V.; Vieira, F.M.C.; Zavaglia, C.; Sossolote, C.R.C.; Hernandez, J. A. (1996) “Construção de um Léxico para a Língua Portuguesa do Brasil para Suporte à Correção Automática de Textos”. In: Relatório Técnico do ICMSC-USP, nro.42, 36p.

Ortiz, A. M. (2000) Diseño e implementación de un Lexicón Computacional para lexicografía y Traducción Automática. “Estudios de Lingüística Española”. Vol.9, http://elies.rediris.es/elies9/index.htm.

Pustejovsky, James. (1995) “The Generative Lexicon”. Cambridge, The MIT Press.

Tiscornia, D. (1995) “Una metodologia per la rappresentazione della conoscenza giuridica; l’ontologia formale applicata al diritto”. Articolo per conferenza di filosofia del diritto. Bologna.

Zavaglia, C. (2002) “Análise da Homonímia no português*:* tratamento semântico com vistas a procedimentos computacionais”. Tese de Doutorado. Araraquara: [s.n].