A HOMONÍMIA NO PORTUGUÊS: TRATAMENTO SEMÂNTICO SEGUNDO A ESTRUTURA *QUALIA* DE PUSTEJOVSKY COM VISTAS A IMPLEMENTAÇÕES COMPUTACIONAIS

Claudia ZAVAGLIA[1]

■ RESUMO: No presente trabalho, apresentamos uma proposta de tratamento semântico de formas ambíguas do português do Brasil, no caso, lexias homógrafas, com o escopo de oferecermos estratégias lingüísticas para a sua implementação computacional em Sistemas de Processamento das Línguas Naturais (SPLN). O Léxico Gerativo de Pustejovsky foi usado como modelo teórico. Nesse modelo, a Estrutura *Qualia* – EQ (e os papéis Formal, Télico, Agentivo e Constitutivo) foi selecionada como um dos expedientes lingüístico-semânticos para a realização da desambiguação das formas homônimas. Para que os dados analisados e tratados pudessem ser manipulados, elaboramos uma Base de Conhecimento Lexical (BCL) cujo repertório lingüístico possui seus itens lexicais correlacionados e interligados por diferentes tipos de relações semânticas presentes na EQ.

■ PALAVRAS-CHAVE: Homonímia; estrutura *Qualia*; léxico computacional; base de conhecimento lexical; processamento das línguas naturais

Introdução

De acordo com Biderman (1996, p.27), "o léxico é o lugar da estocagem da significação e dos conteúdos significantes da linguagem humana". A mesma autora ressalta em obra anterior:

> o léxico pode ser considerado como tesouro vocabular de uma determinada língua. Ele inclui a nomenclatura de todos os conceitos lingüísticos e não-lingüísticos e de todos os referentes do mundo físico e do universo cultural, criado por todas as culturas humanas atuais e do passado (BIDERMAN, 1981, p.138).

[1] Departamento de Letras Modernas – Instituto de Biociências, Letras e Ciências Exatas – UNESP – 15054-000 – São José do Rio Preto – SP – Brasil. E-mail: zavaglia@lem.ibilce.unesp.br.

Com efeito, o léxico de uma língua abrange todas as palavras desse sistema lingüístico, inclusive as gramaticais, que se encontram contempladas ou não em dicionários.

O léxico das línguas naturais foi gerado por um processo de nomeação, a partir do momento em que o homem, por meio das palavras, passou a dar nome a todas as entidades que faziam parte do mundo que o circundava (BIDERMAN, 1998c).

Desde há muito tempo, o léxico tem sido relacionado com a memória humana. De fato, as entradas lexicais em um dicionário são como registros da memória e muito provavelmente a estruturação do léxico se assemelha àquela da memória, fato esse que permite uma recuperação rápida e veloz das palavras que o constituem (BIDERMAN, 1981, p.28). Com efeito, fala-se de Léxico Mental, isto é, as palavras que se encontram estocadas na mente humana.

Quando nos referimos ao tratamento automático das línguas naturais, seja de variadas formas ou por variados mecanismos, estamos tratando essencial e primordialmente da estruturação de itens lexicais de uma maneira formal, ou seja, de codificação e decodificação de dados. Na forma como serão armazenados esses dados, seja em listas imensas de palavras, seja por analisadores morfológicos, seja por índices alfanuméricos em forma de códigos, ou de qualquer outro modo, verificar-se-á uma dependência da estruturação lingüístico-formal dos mesmos. A propósito, Biderman (1998b) diz:

> [...] o léxico está associado ao conhecimento e o processo de nomeação em qualquer língua resulta de uma operação perceptiva e cognitiva. Assim sendo, no aparato lingüístico da memória humana, o léxico é o lugar do conhecimento sob o rótulo sintético de palavras – os signos lingüísticos. Eis por que precisamos começar a trabalhar com esta imensa galáxia de signos que devemos conhecer melhor. É preciso desvendar o mistério de como se estrutura o léxico da nossa língua (BIDERMAN, 1998b, p.179).

É nesse sentido que o computador está fadado a incompletudes, já que um, dentre tantos, mistérios sobre a mente humana ainda é, justamente, a forma como são estocados os dados na memória do homem. Por conseguinte, a forma de armazenamento de dados na memória de uma máquina contém as mesmas (e talvez bem mais) obscuridades que o não-conhecimento sobre mecanismos mentais humanos gera para os pesquisadores. Nesse sentido, Button et al. (1998) afirmam:

> A analogia entre "a mente" e "o computador" também foi contestada com base no fato de que se equivoca sobre o funcionamento dos computadores e sobre a natureza dos programas que os dirigem (são neles executados). O ponto até onde os computadores podem ser usados para simularem atividades humanas dá uma impressão enganosa do grau em que o computador está realmente "igualando" o desempenho simulado. Os computadores podem ser capazes de gerar séries de palavras, símbolos matemáticos etc., que correspondem corretamente aos requisitos da linguagem humana, sistemas de cálculo etc., mas – para dizê-lo

> muito grosseiramente, por enquanto – a diferença crucial entre a simulação da máquina e o desempenho humano é que esse último envolve o entendimento do que as séries de palavras e fórmulas significam, ao passo que a primeira, não (BUTTON et al., 1998, p.12).

Biderman (1981, p.139) sugere:

> [...] em virtude do número elevadíssimo dos elementos do léxico e da complexidade combinatória resultante desse número, é necessário supor que o cérebro organiza uma estrutura dos dados léxicos de grande funcionalidade, para que ele possa recuperar em frações mínimas de segundo (100 a 700 milissegundos) não só o significado de uma palavra, mas todas as suas características gramaticais e os usos que lhe são adequados, conforme o contexto do discurso, a situação momentânea e o registro lingüístico requerido pela situação, pelo interlocutor e pelo assunto.

Essa mesma autora nos diz que, provavelmente, o léxico está encadeado em redes semânticas, i.e., a sua integração está estruturada por vários campos léxicos. E ainda: "os padrões neuronais da memória léxica devem ter estabelecido redes de ligações entre os lexemas de modo funcional" (BIDERMAN, 1981, p.139). Da sua proposta de Rede Semântica e Campo Léxico (BIDERMAN, 1981, p.140), a autora deduziu que a associação entre signos é estabelecida de duas maneiras: por contigüidade/similaridade e por oposição de contrários. Inferimos, portanto, que as relações semânticas da sinonímia e da antonímia fazem parte, essencialmente, do conjunto de estruturação do léxico mental de um ser humano.

Em consonância, Bogaards (1994, p.70-71) diz que as relações entre os elementos do léxico mental são de dois tipos: (i) relações intrínsecas, que se baseiam nos diferentes tipos de informações lingüísticas (semânticas, morfológicas, fonológicas etc.) e (ii) relações associativas, que são baseadas na co-ocorrência freqüente de itens. No que diz respeito à natureza semântica das relações intrínsecas, podemos descrevê-las em termos de sinonímia, antonímia e hiponímia; por sua vez, as relações associativas baseiam-se no conhecimento de mundo e no conhecimento enciclopédico: à palavra *guerra* são associadas palavras como *morte*, *combate*, *miséria* etc.

Tais concepções nos levam a acreditar que uma das maneiras de se ordenar a estruturação de léxico em computadores poderia ser via Redes Semânticas e Associações Semânticas. Essas redes poderiam organizar-se por meio de relações semânticas (hiperonímia, sinonímia, antonímia, meronímia) já que, ao que tudo indica, essa seria a forma que, possivelmente, mais se assemelharia à estocagem de dados na mente humana.

Em conformidade, Bezerra (2002, p.3) enuncia:

> Em nossa memória de longo tempo, ou memória profunda, armazenamos as unidades lexicais da língua que falamos associadas em diversas combinações: sintagmáticas, paradigmáticas, hiponímicas, conceituais, discursivas, dependendo dos modelos da língua que fa-

lamos e de nossas experiências anteriores ou de nossos esquemas culturais. Como é no léxico que se cruzam informações fonético-fonológicas, semânticas, sintáticas e pragmáticas, deve-se considerá-lo em relação à linguagem em geral, como uma competência, neste caso, lexical, que o falante deve desenvolver, para ampliar sua competência comunicativa.

Elaborar repertórios lexicais para serem tratados computacionalmente contribuiria, não somente para as ciências que se interessam por processamento automático de línguas naturais, mas também para a formação de acervos lexicais para a memória de computadores, e, conseqüentemente, para a composição de seus "conhecimentos" que pudessem servir a toda sorte de pesquisadores.

Sendo o léxico de uma língua essencialmente abrangente e complexo, seria de se esperar que fenômenos lingüísticos igualmente complexos e abrangentes caracterizassem e fizessem parte da língua natural à qual ele se encontra vinculado.

Um desses fenômenos é a homonímia, além da polissemia, da sinonímia, entre outros. A homonímia e a polissemia causam o fenômeno da ambigüidade; por conseguinte, temos de considerá-lo como característico de uma língua natural.

Devemos observar, porém, que a ambigüidade não existe do ponto de vista do produtor do discurso, mas sim do seu receptor. De fato, quando um falante produz um texto, muito provavelmente, não se dá conta de um significado alternativo que possa existir no interior de seu discurso, seja ele falado ou escrito; ao contrário, ele tem bem claro em sua mente o que deseja expressar, como afirma Leffa (1998).

Ora, será no âmbito do léxico, bem como dos fenômenos lingüísticos geradores de ambigüidades interpretativas, que um estudioso deparar-se-á com inúmeros empecilhos ao aventurar-se a descrever os seus mecanismos para o Processamento das Línguas Naturais (doravante PLN).

Em conformidade, Carvalho (2001, p.1) ressalta:

A ambigüidade (lexical, estrutural), intrínseca a qualquer língua natural, é um dos aspectos que maiores problemas colocam ao processamento automático de um texto. A nível lexical, a ambigüidade é provocada pela homografia, que existe em qualquer língua natural, mas que é particularmente abundante no caso das línguas que, como o português, têm um sistema morfológico bastante desenvolvido.

Dessa forma, o fenômeno da homonímia causa sérios obstáculos para o desenvolvimento do PLN, máxime para casos de homografia, e lingüistas computacionais tentam, insistentemente, buscar meios de fazer com que a máquina disponha de mecanismos interpretativos de desambiguação que se aproximem daqueles que o homem possui. Com efeito, Carvalho (2001, p.3) põe em relevância esse problema quando diz:

Ainda que os vários casos de homografia de que temos vindo a falar não levantem, em geral, problemas aos falantes da língua, eles representam, retomando a idéia com que iniciámos o capítulo, um obstáculo à quase totalidade das operações efectuadas ao nível do tra-

tamento automático de textos escritos. A fiabilidade dos resultados de uma operação de análise extremamente simples, como por exemplo, a localização num texto dos adjectivos que ocupem uma posição pré-nominal, através da expressão regular: <A><N> está fortemente condicionada pela existência de homografia entre as categorias descritas nessa expressão e outras categorias gramaticais.

Objetivos

Com o presente artigo, apresentamos uma proposta para o tratamento de itens lexicais homônimos da língua portuguesa do Brasil, com vistas à sua implementação computacional, por meio de Base de dados relacionais, mais especificamente uma Base de Conhecimento Lexical (doravante BCL). A hipótese principal que se faz é que o fenômeno da homonímia é passível de tratamentos computacionais e que podemos manipulá-lo em implementações para base de dados lexicais com eficiência. Ressaltamos que o problema da homonímia gramatical é resolvido, e satisfatoriamente, por sistemas computacionais que realizam análises morfossintáticas automáticas (*parsers*) que possuam desambiguadores. A máquina é capaz de produzir soluções de desambiguação sintática de uma maneira bastante aceitável. Entretanto, tais sistemas não dão conta de outros problemas de ambigüidade, como a homonímia semântica e a polissemia. Tal fato ocorre porque a máquina não tem a capacidade de relacionar semanticamente itens lexicais em meio a construções sintáticas ou inseridos em um contexto, como faz o homem, de forma inerente. Como situa Carvalho (2001, p.38): "As máquinas não têm competência lingüística, pelo que 'é preciso dizer-lhes tudo, e é preciso dizer-lhes tudo de forma completa, explícita e coerente'" (RANCHHOD apud CARVALHO, 2001. p.38).

Assim, a ineficiência de desambiguadores[2] de tipo gramatical justifica a proposta de uma Base de dados conceitual, que será proposta com a finalidade de suprir as necessidades de um analisador sintático[3], além de atender possíveis novos sistemas que realizem tratamento semântico.

Em PLN, no que diz respeito à ambigüidade lexical, por exemplo, causada pela homonímia, o computador terá pelo menos duas possibilidades de interpretação para uma mesma forma. Para os casos de homonímia categorial, os resultados podem ser desastrosos se, ao invés de classificar uma forma contextualizada como verbo, o com-

2 Carvalho (2001) aponta alguns casos problemáticos de não resolução de ambigüidades causadas pela homografia, com a aplicação de gramáticas para a desambiguação, em análises lexicais. Por exemplo, nos contextos seguintes, a máquina não etiquetou corretamente as palavras "muda" e "só" nos exemplos: [...] *visíveis e* ***a moda muda*** *muito mais rapidamente. Não existe, penso aquela* [...] e *sim,* [...] *que* ***a sua alma só*** *entra em actividade vulcânica quando o político* [...]. "As palavras "muda" e "só" foram reconhecidas como adjetivos, quando, na realidade, se trata de uma forma do verbo *mudar* e do advérbio *só*, respectivamente" (CARVALHO, 2001, p.93).

3 Zavaglia (1999) cita vários casos de homografia categorial que não foram satisfatoriamente tratados pelo *parser* do Revisor Gramatical *ReGra* entre substantivo X adjetivo (lexia "cara", "tinta", "vaga", "polêmica", "fluxo", "queda") e entre substantivo X verbo (lexia "ajuda"), por exemplo.

putador categorizá-la como substantivo, por exemplo. Para Revisores gramaticais automáticos, tais interpretações errôneas interferem na performance da ferramenta, gerando insatisfação para seus usuários.

Unir informações semânticas às informações de uma gramática formal, ou seja, dotá-la de uma base de conhecimento de mundo, é um caminho para amenizar problemas de ambigüidade em PLN, segundo a literatura atual. Desse modo, na gramática formal seriam introduzidos marcadores semânticos que permitiriam à máquina resolver casos de ambigüidade segundo um esquema de compartilhamento ou não-compartilhamento de dados. De fato, Medeiros (1999, p.8) diz: "Os aspectos semânticos devem ser contemplados para solucionar problemas não resolvidos pela análise sintática, como, por exemplo, o da ambigüidade lexical e estrutural, e o das sinonímias".

Ainda que, no presente, não saibamos com precisão quais serão os resultados (positivos ou negativos) de suas aplicações, temos a certeza de que informações de cunho meramente sintáticas ou morfossintáticas não mais satisfazem pesquisadores em Lingüística Computacional, pois são insuficientes no PLN. De fato, somente com a elaboração de Base de dados conceituais poder-se-ão obter análises de textos com bons resultados.

A adoção do modelo sugerido por Pustejovsky (1995) deveu-se a pelo menos quatro componentes nele contidos: (i) atualização teórica, (ii) representatividade do significado, (iii) natureza computacional, (iv) aplicabilidade (Cf. Projeto SIMPLE em LENCI, 1999). A idéia de que o Léxico Gerativo (LG) é capaz de dar conta do conhecimento semântico global que temos sobre as palavras, segundo o próprio autor, faz dele um modelo adequado para solucionar o problema da representação lexical que envolve o fenômeno da homonímia. Admitindo-se, portanto, que tal suposição seja verdadeira, tentaremos mostrar que a homonímia pode ser, realmente, definida conforme os parâmetros de um dos aspectos dessa teoria.

Investigação teórica: o fenômeno da homonímia e modelo semântico adotado

Da investigação teórica que realizamos, detalhada em Zavaglia (2002), constatamos que a homonímia, enquanto fenômeno de uma língua natural, não é mais intrigante e enigmática do que a sua própria definição, ou seja, a sua compreensão e a sua delimitação. Para defini-la, os autores oscilam entre critérios diacrônicos, convergência fonética, divergência semântica, influência estrangeira, polissemia homonímica, critérios sintáticos e morfológicos, distinções estilísticas e sociais, ortografia, entre outros.

Por conseguinte, definimos como parâmetros teóricos de nossas pesquisas, no que diz respeito ao fenômeno da homonímia, os seguintes postulados:

(I) A homonímia é o fenômeno lingüístico em que se tem a identidade de duas lexias no plano da expressão, ou seja, formas perfeitamente iguais que se distinguem semanticamente (um significante para dois significados, no plano do conteúdo) ou a identidade de duas construções gramaticais, gerando a ambigüidade. O primeiro refere-se à homonímia lexical e o segundo à homonímia estrutural.

(II) Para a homonímia lexical, a igualdade de formas pode se realizar tanto graficamente como fonicamente. No primeiro caso, as lexias possuem identidade de grafia (homografia) e, no segundo, identidade de som (homofonia). E assim, temos lexias homógrafas que: (i) são distintas quanto ao seu significado e idênticas, tanto oralmente como gramaticalmente, caso esse denominado de Homonímia Semântica; como: ***cabo*$_1$**: "Militar que tem posição superior ao soldado e inferior ao sargento" *X* ***cabo*$_2$**: "Extremidade de um objeto que serve para para segurar"; ***colônia*$_1$**: "País ou região dependente de um outro país em situação econômico-política superior" *X* ***colônia*$_2$**: "Líquido que serve para se perfurmar que possui uma essência menos concentrada do que a do seu extrato"; ***parábola*$_1$**: "História que contém um fundo moral ou religioso" *X* ***parábola*$_2$**: "Curva cujos pontos são eqüidistantes de um ponto fixo e de uma reta fixa"; ***calar*$_1$**: "Emudecer, não falar" *X* ***calar*$_2$**:"Penetrar, repercutir"; (ii) são distintas quanto ao fato de pertencerem a classes gramaticais diversas e serem idênticas oralmente caso esse denominado de Homonímia Categorial, como: ***caça*$_1$** (substantivo) X ***caça*$_2$** (verbo); ***calça*$_1$** (substantivo) X ***calça*$_2$** (verbo); (iii) são distintas quanto ao seu étimo e idênticas oral e graficamente, caso esse denominado de Homonímia Etimológica, como: ***manga*$_1$**: "Fruto" [Do *malaiala* manga.] X ***manga*$_2$**: "Parte do vestuário" [Do lat. *manica*, 'manga de túnica'.]; (iv) são distintas na sua realização oral, caso esse denominado de Homonímia Heterófona[4], nas quais o substantivo realiza-se fonicamente como [e] e o verbo como [ɛ], para a vogal "e", como nos seguintes exemplos: ***acerto*$_1$** (substantivo) *X* ***acerto*$_2$** (verbo); ***começo*$_1$** (substantivo) *X* ***começo*$_2$** (verbo)

(III) As lexias homófonas são aquelas distintas na grafia e idênticas no som, como por exemplo: ***sensor***: "dispositivo" *X* ***censor***: *"crítico"*; ***cessão***: "ato de ceder" *X* ***seção***: "segmento, divisão" *X* ***sessão***: "espaço de tempo que dura uma reunião, um ato"

(IV) Já a homonímia estrutural realiza-se quando temos duas construções gramaticais idênticas com sentidos diferentes: ***Entrei no carro andando*** (entrei no carro que andava) *X* ***Entrei no carro andando*** (entrei no carro enquanto eu andava).

[4] Forma que possui grafia idêntica a de uma outra forma e ambas se pronunciam diferentemente.

A organização conceitual da BCL assemelha-se àquela de um *thesaurus*, já que os itens lexicais se encontram correlacionados e interligados por diferentes tipos de relações semânticas. Atualmente, para os estudos de PLN, o levantamento e a identificação das relações léxico-semânticas entre as palavras são de extrema importância, já que fundamentais como fontes de recursos lingüísticos para implementação computacional. De fato "a informação lexical e semântica é instrumento indispensável para programas que analisam e 'compreendem' textos em língua natural" (DEL FIORENTINO, 1995, p.8).

Em nossa concepção, pressupomos que um dos maiores problemas de ambigüidade interpretativa para o PLN, a saber, o fenômeno da homonímia, pode ser tratado a partir de subsídios lingüísticos oferecidos à máquina, tais como as relações semânticas de itens lexicais em redes de significação.

As definições, bem como as entradas das formas homógrafas, que constam da amostragem de nossa pesquisa, foram extraídas do *Dicionário Didático de Português* de Biderman (1998a), doravante DDBI. A escolha desse dicionário foi devida, principalmente, à cuidadosa elaboração dos verbetes para os homônimos, por parte de sua autora, bem como ao critério de delimitação para uma forma homônima, a saber: aquele de base semântica. Além disso, em sua elaboração, a autora utilizou-se de um *corpus* representativo da língua portuguesa do Brasil e valeu-se de dados de freqüência lexical para a constituição da sua nomenclatura. Vejamos o exemplo:

ato[1] s.m. a-to. Ação humana, considerada do ponto de vista objetivo e não durante o seu transcurso. *Todo mundo é responsável por seus atos. Este foi um grande ato de coragem! Este bombeiro merece uma medalha. O chefe dispensou o funcionário; foi um ato justo.*
ato[2] s.m. a-to. 1. Solenidade ou cerimônia para marcar um fato. *Os grevistas marcaram um ato público para às 16:00h.* 2. Decisão publica emitida por uma autoridade e publicada em diário oficial. *O governador admitiu novos funcionários, através de uma ato administrativo, que o diário oficial publicou ontem.* 3. Evento que se registra, porque representa um acordo comercial, ou de natureza permanente entre as partes. *O marido e a mulher devem estar presentes no ato de venda de uma propriedade comum ao casal.*
ato[3] s.m. a-to. 1. Cada uma das divisões de uma peça de teatro. *O drama está dividido em dois atos.* 2. Momento considerado como dramático. *O último ato da vida deste ditador será provavelmente sangrento.*

Tabela 1 – Exemplo extraído do DDBI (BIDERMAN, 1998a, p117).

ATO é uma forma que possui 3 conceitos ou significados diferentes não interligados entre si, i.e., um significante com três significados; trata-se, portanto, de formas homônimas. O *sentido*[1] é explicado ou definido por "ação humana, considerada do ponto de vista objetivo e não durante o seu transcurso". Já o *sentido*[2] possui três acepções, separadas entre si por caracteres numéricos; trata-se de uma forma homógrafa

que possui três sentidos correlatos entre si e, portanto, polissêmicos. Por sua vez, o *sentido*^{3} possui duas definições, sendo considerada também um vocábulo polissêmico, como exemplificado para o *sentido*^{2}.

Os sentidos das entradas apresentadas na Tabela (1) são identificados do seguinte modo: ato 0_1 / 0_2a / 0_2b / 0_2c / 0_3a / 0_3b, em que:

a) cada forma homógrafa é classificada por meio de um código de identificação que contém duas partes separadas por um traço: X_X;
b) a primeira parte é, justamente, a classificação-identificação de forma homógrafa, a partir de um caractere numérico, a saber: 0_X, em que "0" representa "forma homógrafa";
c) a segunda parte do código corresponde ao número de ocorrências da homografia para aquela forma, a partir de um caractere também numérico: 0_1; 0_2 (identifica uma forma homógrafa que possui dois significados);
d) a segunda parte do código pode conter, também, um caractere alfabético que identifica a ocorrência da polissemia para cada forma homógrafa: 0_1a, 0_1b; 0_2 (identifica uma forma homógrafa que possui dois significados, cujo primeiro é polissêmico, com dois sentidos), indicada e registrada no interior do verbete, geralmente, por meio de caracteres numéricos.

No âmbito computacional, os termos *genus* e *differentia*, emprestados de Aristóteles[5], foram introduzidos por Amsler (1980), a partir do momento em que toda definição de uma entrada lexical (ou *definiendum*) de um dicionário padrão, segundo o autor, pode ser analisada como uma seqüência constituída por um termo indicador do *genus* e por um outro indicador da *differentia*. Os dicionários convencionais possuem uma tipologia de definição própria, i. e., apresentam um item lexical considerado como sendo o núcleo da definição que é antecedido ou seguido por modificadores.

Os modificadores do *genus terminus* têm papel importante e fundamental na definição do conceito da entrada lexical. De fato, eles constituem as *differentiae* da definição e oferecem indícios de significação que delimitam o conceito no interior da definição expresso pelo *genus terminus*.

Por taxionomia entende-se "palavras baseadas em relações específicas existentes, geralmente, entre o *definiendum* e o *genus terminus* em uma definição lexicográfica padrão" (CALZOLARI et al., 1991, p.25).

Partindo-se da afirmação de que as definições contidas em um dicionário possuem informações semânticas (AMSLER, 1980), buscamos estabelecer a taxionomia existente para as formas homógrafas com categoria idêntica, no caso, nome. Como ponto de partida para essa estruturação, foram identificados os *genus terminus* de ca-

[5] Dentre muitos de seus feitos e contribuições, (o estabelecimento da lógica formal, por exemplo), Aristóteles estabeleceu a distinção entre atributos: o gênero, a espécie, a diferença, o próprio e o acidente. Segundo o filósofo, o gênero se refere à classe mais ampla a que o sujeito pode pertencer: "O homem é um **animal**" e a diferença permite situar o sujeito relativamente às subclasses em que se divide o gênero: "O homem é animal **racional**" (ARISTÓTELES 2000, p17, grifo do autor).

da entrada homógrafa, bem como para cada acepção semântica. De fato, uma forma homógrafa, no interior de sua definição, possui significados polissêmicos, como, por exemplo, em: *ato*[1] (uma acepção), *ato*[2] (três acepções) e *ato*[3] (duas acepções), como descrito anteriormente.

A nossa metodologia partiu do pressuposto de que a definição de cada um dos sentidos de um lema contém pelo menos uma relação semântica entre o próprio lema e *o genus terminus* ou também entre o lema e a *differentia* das definições (DEL FIORENTINO, 1995).

A extração do *genus terminus* das definições das entradas lexicais de um dicionário é uma etapa importante e essencial para que se realize uma organização taxionômica de um repertório lexical, segundo uma estrutura hereditária em termos de hiperonímia. Com efeito, o *genus terminus* será localizado no vértice dessa estrutura.

Uma língua natural utiliza-se de uma enorme variedade de realizações lexicais e/ou sintáticas para expressar os conceitos do mundo elaborados nessa língua. De fato, o léxico de uma língua, bem como a sua realização sintática, é imensurável; até hoje, tem-se, efetivamente, apenas aproximações de realizações lexicais e não confirmações de números finitos. Um dicionário buscará se servir, portanto, de todos os recursos lingüísticos de que uma língua possui para poder expressar o conceito de um item lexical. Tais conceitos são definidos por meio de relações semânticas que os itens lexicais da definição do *definiendum* mantêm entre si.

A partir das definições do DDBI (BIDERMAN, 1998a) formalizadas, procedemos à extração do *genus terminus* e da sua relação semântica com o *definiendum*, ou seja, a taxionomia. Alguns tipos de relações semânticas, e seus princípios, interessam de modo particular ao modelo semântico que propomos para a organização da base de dados lexical. O significado de cada item lexical pertencente a essa base é estruturado, justamente, a partir das relações semânticas que o conceito desse item lexical mantém com outro item lexical. De fato, os conceitos são interligados em uma "cadeia significativa", ou seja, por meio de associações. Cada item lexical situa-se em um determinado lugar dessa cadeia e todos eles são correlacionados, por meio de conexões, àqueles com os quais possui pelo menos uma relação semântica.

Segundo Picoche (1992, p.138):

> [...] é um fato biológico que os homens sejam aptos a perceber diversos níveis de abstração e a passar facilmente de um para o outro; é uma propriedade universal da linguagem humana ser capaz de explicar e de condensar, de poder exprimir em mais de uma palavra aquilo que é dito em uma palavra (expansion – expansão) e de poder [...] resumir em uma palavra aquilo que é dito com mais de uma palavra (condensation – condensação).

Tal afirmação não se aplica a um computador, ao contrário. A máquina perceberá níveis de abstração se a ela forem oferecidos dados para tal[6], fato esse que vale tam-

[6] Se isso for realmente possível.

bém para as capacidades de expansão e condensação. A mesma autora ainda diz que todo homem que é dotado da fala manipula espontaneamente conjuntos de sinônimos e até mesmo as suas equivalências. Em contrapartida, a máquina deverá apresentar procedimentos artificiais para a manipulação desses sinônimos e de seus equivalentes.

A decomposição do significado proposta por Pustejovsky (1995) em sua teoria é capaz de oferecer caminhos para que uma máquina recupere um conjunto de sinônimos e/ou equivalentes para uma determinada unidade léxica. Com efeito, a partir do momento em que a definição de um item lexical apresenta o seu conteúdo por meio de relações de significação com outros itens lexicais em uma cadeia significativa, itens sinônimos (se existirem) podem ser recuperados para uma unidade léxica. O mesmo aplicar-se-á à busca/recuperação de itens hiperônimos, hipônimos, antônimos, merônimos. Com efeito, para o autor, um léxico gerativo é caracterizado como um sistema computacional que envolve, no mínimo, quatro níveis de representação: (i) Estrutura Argumental (*Argument Structure*), em que se tem a especificação do número e do tipo de argumentos lógicos e como eles são realizados sintaticamente; (ii) Estrutura de Evento (*Event Structure*), na qual há a definição do tipo de evento de um item lexical e uma frase. Inclui eventos do tipo ESTADO, PROCESSO e TRANSIÇÃO que podem ter uma estrutura de subeventos; (iii) Estrutura *Qualia* (*Qualia Structure*) que inclui modos de explicação compostos pelos papéis FORMAL, CONSTITUTIVO, TÉLICO e AGENTIVO e (iv) Estrutura de Herança Lexical (*Lexical Inheritance Structure*), em que se tem a identificação de como uma estrutura lexical se relaciona com outras estruturas e a sua contribuição para a organização global do léxico.

Assim, Pustejovsky (1995, p.62) propõe que a semântica de um item lexical "**a**" seja definida como uma estrutura composta por quatro componentes:

$$\alpha = < \mathbf{A}, \mathbf{e}, \lambda, \acute{\mathbf{Y}} >^{7} \text{ em que:}$$

A é a estrutura argumental; **e** a especificação do tipo de evento; λ estabelece o vínculo desses dois parâmetros na Estrutura *Qualia* e **Ý** determina qual informação é hereditária na estrutura lexical global.

A nosso ver, Pustejovsky (1995) procura recuperar as dimensões do significado de um item lexical a partir dos conceitos individuais de outros itens lexicais, tendo como ponto de partida a natureza do significado inerente e já cristalizado nas unidades léxicas. Neste caso, a afirmação de Richelet (séc. XVII) de que uma definição é "um discurso que explica nitidamente a natureza de uma coisa" (apud PICOCHE, 1992, p.140) é válida e pertinente.

A partir do momento que Pustejovsky (1995) especifica quatro papéis fundamentais do significado de uma palavra na Estrutura *Qualia* (Constitutivo, Formal, Télico e

[7] Adaptação nossa da simbologia da teoria de Pustejovsky (1995).

Agentivo), o autor está delimitando o significado por meio de marcas distintivas[8]. De fato, cada um dos aspectos essenciais do significado de um item lexical possui traços que os especificam:

- 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 identifica 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.

Retomando os mesmos exemplos citados acima, confiramos:

ato$0_1

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \ldots\ldots \\ \text{FORMAL} = \text{ação} \\ \text{TELIC} = \text{agir} \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

ato$0_2a

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \ldots\ldots \\ \text{FORMAL} = \text{solenidade} \\ \text{TELIC} = \text{comemorar} \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

ato$0_2b

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \ldots\ldots \\ \text{FORMAL} = \text{decisão} \\ \text{TELIC} = \text{decidir} \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

ato$0_2c

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \ldots\ldots \\ \text{FORMAL} = \text{evento} \\ \text{TELIC} = \text{acordar} \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

ato$0_3a

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \text{peça} \\ \text{FORMAL} = \text{divisão} \\ \text{TELIC} = \text{dividir} \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

ato$0_3b

$$\begin{bmatrix} \textit{Qualia} = \begin{bmatrix} \text{CONST} = \text{vida} \\ \text{FORMAL} = \ldots\ldots \\ \text{TELIC} = \ldots\ldots \\ \text{AGENT} = \ldots\ldots \end{bmatrix} \end{bmatrix}$$

Para Picoche (1992, p.140), em uma definição lingüística, o que importa é a especificidade, ou seja, a indicação de traços distintivos pertinentes a um item lexical que o diferenciará de outros itens lexicais.

Base de conhecimento lexical: uma sugestão de modelo

Em nossa proposta, os dados que figuram na Base de Conhecimento Lexical – BCL se encontram dispostos de modo a poderem ser utilizados em uma rede semântica em sistemas computacionais, uma vez que possuem características e propriedades da mesma.

A elaboração de recursos lexicais que contenham informações semânticas faz-se importante para sistemas que tratam da desambiguação dos sentidos das palavras,

[8] Cabe lembrar que não necessariamente todos os papéis *qualia* devem estar preenchidos.

como por exemplo, a Tradução Automática, a Recuperação de Informação, Motores 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.

O modelo de representação aqui proposto contém informações de tipo semântico e morfossintático. Essas últimas restringiram-se à classe gramatical, ao gênero e ao número das palavras. Em contrapartida, privilegiamos o tipo de informações semânticas, 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.

Em PLN, sabe-se da importância que se atribui a esses dois tipos de componentes, dado que pesquisadores na área afirmam que a sintaxe não prescinde (e não deve prescindir) da semântica em análises automáticas. De fato, Salton (apud MEDEIROS, 1999, p.64) afirma que "a sintaxe sozinha não resolve muitas ambigüidades que complicam a tarefa de análise de conteúdo" e ainda, Binot (1991, p.61) ressalta que "essa necessidade de informação semântica é reconhecida há muito tempo: a resolução de ambigüidades, elipses, atos do discurso deve apoiar-se no sentido das palavras e no contexto do discurso". Da mesma forma Hagége e Duarte (1995) defendem que analisadores somente sintáticos ou somente semânticos dão conta apenas de uma parte do tratamento da linguagem e que nos dias de hoje ninguém nega a necessidade de considerar a língua de um ponto de vista sintático e semântico.

Nos mesmos moldes de SIMPLE (LENCI, 1999) e ItalWordNet (e suas antecessoras WordNet e EuroWordNet, (CALZOLARI, 2000), 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, neste trabalho, realizou-se um esforço de individualizar os hipônimos e os hiperônimos das formas homônimas, 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 aqui proposto não pretende definir de modo direto o signifi-

cado de cada item homógrafo. Pretende tão somente sugerir o significado para cada item homógrafo, bem como para suas ocorrências polissêmicas, por meio de termos interligados a cada ocorrência homógrafa que têm por escopo delimitar o seu campo significativo.

Dada a suposição de que múltiplas dimensões do significado são necessárias para começar a caracterizar unidades lexicais em um nível semântico, a Estrutura *Qualia* tem sido utilizada[9] 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. O objetivo geral é ir além de uma hierarquia dimensional, resgatando, assim, o padrão de relações de hiponímia e hiperonímia.

Informações baseadas na Estrutura *Qualia* podem ser especificadas por todas as partes do discurso, embora, em primeira instância, ela pareça ser mais diretamente adequada para a caracterização dos nomes (LENCI, 1999). Justifica-se, dessa forma, o fato de termos nos detido na codificação de formas homônimas cuja categoria é a do nome.

A Estrutura *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 e 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.

Em SIMPLE (LENCI, 1999), a Estrutura *Qualia* é usada como sintaxe básica para a construção do significado lexical (PUSTEJOVSKY apud LENCI, 1999). Cada papel *Qualia* pode ser visto como um elemento independente ou uma dimensão independente do vocabulário para a descrição semântica. A partir da compreensão do papel da Estrutura *Qualia*, é possível formular um conjunto de questões que, de uma perspectiva teórica, são o núcleo da pesquisa em semântica lexical e, de uma perspectiva prática, permite realizar uma codificação sistemática em larga escala.

O modelo de Léxico Gerativo impôs alguns requisitos para a representação do aspecto componencial do significado lexical. Para satisfazer essas exigências, os papéis *Qualia*, no projeto SIMPLE, foram implementados como relações entre unidades semânticas (SemU) e, em um número mais restrito de casos, como *valued features* (características de valor). Tal fato levou ao desenvolvimento de uma estratégia representacional que permite a lexicógrafos, por exemplo, representarem ou codificarem uma riqueza de relações semânticas existentes em uma língua natural, na medida em que

[9] Um exemplo da utilização da Estrutura *Qualia* como representação do significado pode ser visto em Hathout (1996) onde estão as especificações da elaboração de uma base de conhecimento lexical para o domínio da química, na qual as informações específicas das entidades desse domínio correspondem ao papel Formal da Estrutura *Qualia*.

mantém a estrutura básica de propriedades dos tipos semânticos dados em termos de Estrutura *Qualia*.

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 uma dado *Quale*. Essa hierarquia nos quatro papéis *Qualia* é chamada de Conjunto de *Qualia* Ampliado (*Extended Qualia Set*). Para cada um dos quatro papéis *Qualia* foi especificado um Conjunto de *Qualia* Ampliado, ou seja, foram especificados subtipos de um dado papel *Qualia* que são coerentes com a sua interpretação.

A partir dos itens lexicais "nadador" e "peixe", vejamos algumas razões lingüísticas para que seja incluído o Conjunto de *Qualia* Ampliado na captura de similaridades entre palavras pertencentes às mais diversas áreas conceptuais.

Um nadador é claramente um indivíduo cuja função típica é "nadar" (nos exemplos que seguem, os termos entre "<" e ">" são de unidades semânticas (SemUs)):

$$\left[\begin{array}{l} \textbf{nadador} \\ \\ \textit{Qualia} = \left[\begin{array}{l} \text{FORMAL = isa(<nadador>,<pessoa>)} \\ \text{TELIC = is\_the\_activity\_of}^{10}\text{ (<nadador>,<nadar>)} \end{array}\right] \end{array}\right]$$

No processo de decodificação da semântica do item lexical "peixe", pode-se querer codificar a informação de que uma das atividades típicas de um peixe é nadar. Permutando-se os dois nomes com um adjetivo, poder-se-á perceber o comportamento lingüístico diferente dos dois itens lexicais:

(1) *um velho nadador*
(i) uma pessoa que é velha e que nada.
(ii) uma pessoa que nada há muito tempo.

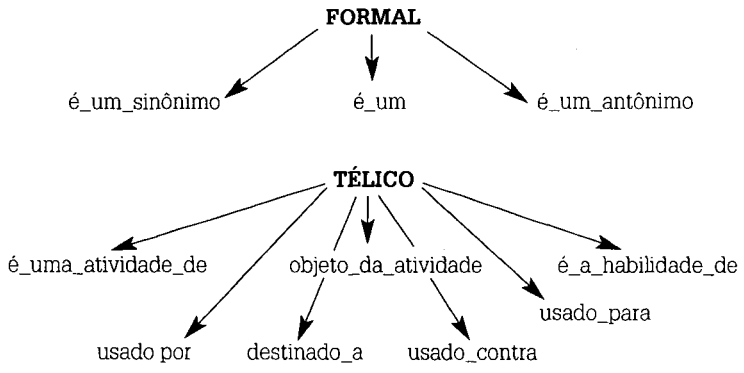
(2) *um velho peixe*
(i) um peixe que é velho.
(ii) um peixe que nada há muito tempo. **

A informação de que um peixe "nada" não faz parte corretamente da dimensão télica, i.e, não funciona como um objetivo hereditário. A propriedade de nadar não acrescenta uma informação télica para o item, mas especifica o "peixe" na sua dimensão constitutiva/física. Por esta razão, a informação de que um peixe nada é expressa no papel Constitutivo de *Qualia*, por meio da relação Constitutive_activity:

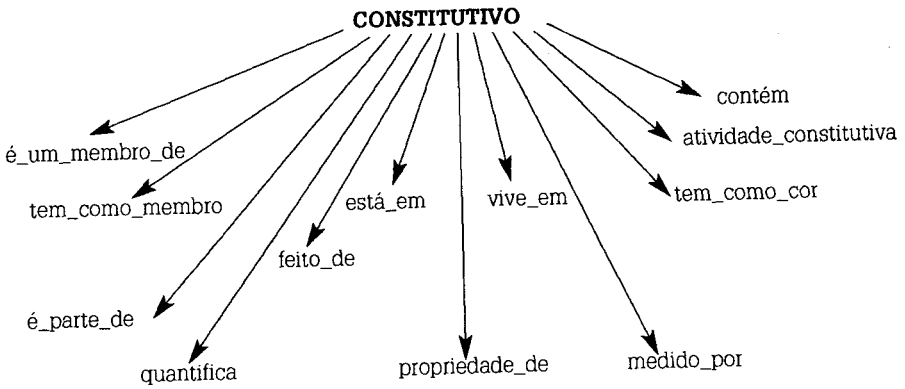
[10] É uma relação da *Qualia* Ampliada e significa "É_a_atividade_de".

$$\left[\begin{array}{l} \textbf{peixe} \\ \\ \textit{Qualia} = \left[\begin{array}{l} \text{FORMAL} = \text{ISA}(\langle\text{peixe}\rangle,\langle\text{animal}\rangle) \\ \text{CONST} = \text{Constitutive\_activity}^{11}\ (\langle\text{peixe}\rangle,\langle\text{nadar}\rangle) \end{array}\right] \end{array}\right]$$

Em nosso modelo de representação, 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 (LENCI, 1999). Desse modo, a ambigüidade semântica entre formas homônimas será 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} será recuperado de forma desambiguada. Além disso, a relação semântica que o item homônimo mantém com um outro item lexical de um repertório lexical oferecerá indícios para a sua desambiguação. E ainda, a formalização em uma base ontológica poderá, ainda, suprir eventuais ambigüidades que o conceito do item homônimo poderá gerar, dependendo do contexto no qual encontrar-se-á inserido. Os valores dos papéis *Qualia* da Unidade Semântica (SemU) são apresentados por meio de relações entre SemU e outras SemUs que especificam a natureza dessas relações. O conjunto de relações proposto para representar a informação *Qualia* contém as relações que estão disponíveis no Léxico Gerativo e também as que foram introduzidas em SIMPLE. De forma esquemática, cada papel da Estrutura *Qualia* possui as seguintes relações semânticas:



[11] É uma relação da *Qualia* Ampliada e significa "Atividade_constitutiva".



De posse de todas as informações que julgamos necessárias para a construção do paradigma da nossa BCL, a saber,

- informação ontológica[12] (subdividida em *Tipo*, que corresponde ao hipônimo; *Supertipo*, que corresponde ao hiperônimo e *Domínio*);
- informação *Qualia* (papéis Formal, Agentivo, Télico e Constitutivo);
- informação morfossintática (*Rep_PDD*, i.e., Representação das partes do discurso e *Rep_Morf.* i.e., Representação morfológica);
- informação definicional, i.e., a definição extraída do dicionário de base, representada por *Glossário*;
- informação pragmática[13], 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*:

12 Para esse tipo de trabalho, elaboramos uma Ontologia de conceitos que procura representar o conhecimento de mundo por meio de categorias de representação, divididas em Classes Fundamentais (Tipo e Supertipo) e Domínios. Como amostragem, temos as categorias "1. Entidade", "1.1. Entidade Concreta", "1.1.1. Localização" "1.1.2. Manufaturado"; "1.1.3. Alimentos", "1.1.4. Entidade Viva"; "1.2. Entidade Abstrata", "1.2.1. Tempo", "1.2.2. Fato Cognitivo", "1.2.3. Padrão Moral", "1.2.3.4. Doutrina"; "2. Escopo"; "3. Agentivo"; "4. Constitutivo"; "5. Propriedade"; "6. Representação"; "7. Evento" para as Classes Fundamentais e "Alimento", "Agricultura/Pesca/Silvicultura", "Negócios", "Serviços", "Atividades Artesanais", "Indústria de Transformação", "Construção", "Política e Governo", entre outros, para Domínios. A estrutura arbórea completa dessa Ontologia, com exemplificação de inserções de lexias para cada categoria ou sub-categoria, pode ser vista em Zavaglia (2002).

13 Os exemplos foram extraídos de um *corpus* fundamental de 11 milhões de palavras do Laboratório de Estudos Lexicográficos da Unesp de Araraquara.

banco [0_1 / 0_2]	
HomoU[14]**:**	"banco$0_1"
SemU[15]**:**	<banco>
DesamU[16]**:**	"objeto$P_1"
Tipo:	[Mobília]
Supertipo:	[Manufaturado]
Domínio:	Móveis (Mobiliaria)
Formal:	*é_um*(<banco>,<objeto>)
Agentivo:	<Nil[17]>
Constitutivo:	*feito_de*(<banco>,<pedra>)
	feito_de(<banco>,<madeira>)
	é_parte_de(<banco>,<mobília>)
Télico:	*usado_para*(<banco>,<sentar>)
Glossário:	Objeto alongado, com ou sem encosto, em que várias pessoas podem assentar-se
Exemplo:	*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
⇨[18]	
HomoU:	"banco$0_2"
SemU:	<banco>
DesamU:	"empresa$P_1"
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)
Rep_PDD:	NOME
Rep_Morfo:	MASC SING

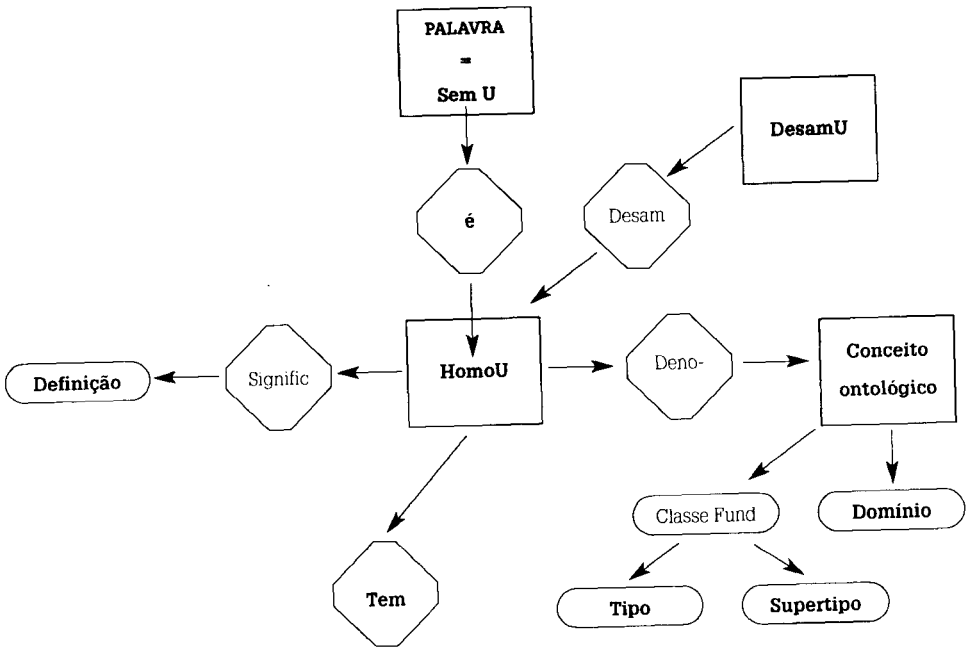
Tabela 2 – Forma homônima BANCO

[14] Unidade Homônima.
[15] Unidade Semântica.
[16] Unidade Desambiguadora.
[17] O símbolo <Nil> é usado quando o elemento não sofre variação na composição.
[18] Essa flecha indica que as duas tabelas encontram-se correlacionadas.

Por meio de uma representação gráfica (diagrama), pretendemos tornar explícitos os vínculos que cada uma dessas informações possui com a unidade homônima em questão, em que:

- Entidades são: **SemU** (Unidade Semântica); **HomoU** (Unidade Homônima); **DesamU** (Unidade Desambiguadora). Toda **SemU** possui pelo menos duas **HomoU**, já que estamos tratando de formas homônimas. Exemplo: a **SemU** <banco> possui duas **HomoU**, a saber: "banco$0_1" e "banco$0_2". Toda **HomoU** possui uma **DesamU**: "banco$0_1" possui a **DesamU** "objeto$P_1" e "banco$0_2" possui a **DesamU** "empresa$P_1".
- Relacionamentos são os vínculos "é", "tem", "significa", "denota", "desambigua".
- Relações semânticas são rótulos de arcos que ligam dois nós. Por exemplo, em "banco$0_1": os nós <banco> e <objeto> são ligados pelo arco de rótulo "é_um"; <banco> e <pedra> por "feito_de"; <banco> e <tecido> por "feito_de"; <banco> e <mobília> por "é_parte_de" e <banco> e <sentar> por "usado_para". Essas relações semânticas se encontram na *Estrutura Qualia*. Vejamos o diagrama:

Diagrama das entidades/relacionamentos



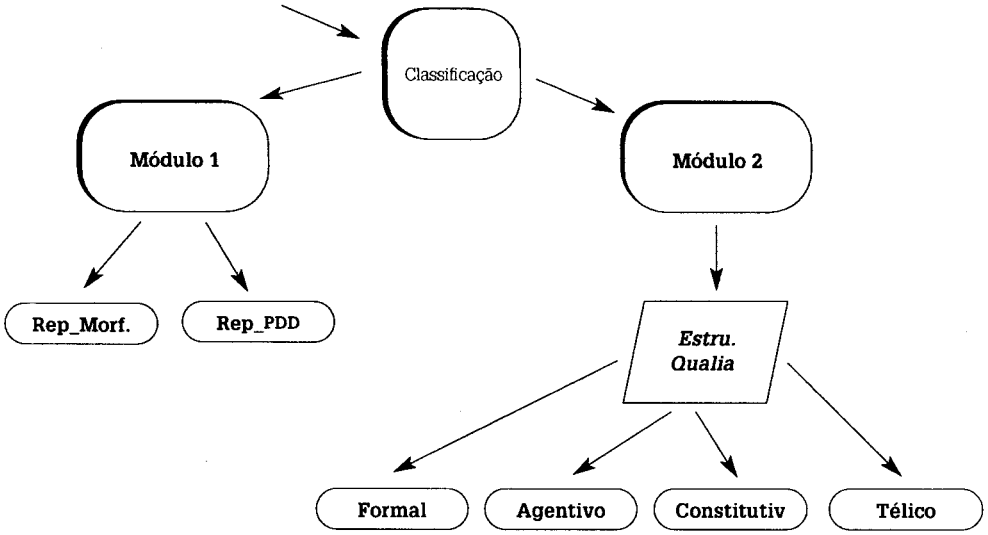


Figura (1) Diagrama Entidades/Relacionamento

Considerações finais

A Estrutura *Qualia* do Léxico Gerativo serviu como estrutura representacional para expressar partes do aspecto componencial do significado lexical, na medida em que se demonstrou capaz de resgatar ou de capturar diferentes graus de complexidade entre itens lexicais e de sustentar, ainda, um conjunto de inferências que está disponível para *default*, ou melhor, essas inferências são usadas como um padrão a ser seguido. Para cada um dos quatro papéis *Qualia*, especificamos um Conjunto de *Qualia* Ampliado, ou seja, esmiuçamos subtipos de um dado papel *Qualia* por meio de diversas relações semânticas, dependendo das características intrínsecas a cada papel *Qualia*. A BCL conta hoje com cerca de 200 formas homônimas de categoria nominal, estruturadas e organizadas segundo o modelo exposto.

Como resultado concreto de nossas pesquisas, análises e investigações, nos propusemos a apresentar uma versão computacional de nosso modelo de Base de Conhecimento Lexical – BCL que foi implementada pelo Núcleo Interinstitucional de Lingüística Computacional – NILC da Universidade de São Paulo – USP/São Carlos que se encontra residente no próprio NILC, em uma máquina servidora, com a seguinte configuração: Pentium II MMX, 266 MHz, 128 Mb RAM, com sistema operacional Windows 2000 Server (ZAVAGLIA, 2002).

Agradecimentos

Ao CNPq pelo auxílio financeiro concedido em ocasião do doutorado sanduíche realizado no *Istituto di Linguistica Computazionale di Pisa* – ILC, onde parte desta pesquisa foi desenvolvida, sob a coordenação da Prof. Nicoletta Calzolari, à coordenadora do NILC – Profa. Dra. Maria das Graças Volpe Nunes e à computóloga Juliana Greghi pelo apoio e estímulo recebidos para a implementação computacional da BCL e à Profa. Dra. Maria Tereza Camargo Biderman, orientadora e incentivadora.

ZAVAGLIA, C. Homonymy in Portuguese: the use of Pustejovsk's Qualia structure approach to foster computational implementations. *Alfa*, São Paulo, v.47, n.2, p.77-99, 2003.

■ *ABSTRACT: This paper applies Pustejovsky's Qualia structure approach to describe homography in Brazilian Portuguese and highlights specific linguistic strategies for treating the phenomenon within the natural language processing domain. Pustejovsk's quale roles – Formal, Telic, Agentive and Constitutive – were selected as semantic devices to aid natural language processing systems in the task of lexical disambiguation. The proposal was implemented in a toy Lexical-Kwowledge-Base system where lexical items are interrelated by quale roles*

■ *KEYWORDS: Homonymy; Qualia structure; lexical knowledge base; natural language processing.*

Referências bibliográficas

AMSLER, R. A. *The structure of the Merriam-Webster pocket dictionary.* 1980. Dissertation (Phd) – University of Texas, Austin, 1980.

ARISTOTELES: vida e obra. São Paulo: Nova Cultural, 2000. (Os Pensadores).

BEZERRA, M. A. Leitura e escrita: condições para aquisição de vocabulário. *Intercâmbio.* Disponível em: <Http://lael.pucsp.br/intercambio/08bezerra.ps.pdf>. Acesso em: 29 maio 2002.

BIDERMAN, M. T. C. A estruturação mental do léxico. In: *ESTUDOS de filologia e lingüística: em homenagem a Isaac Nicolau Salum.* São Paulo: T. A. Queiroz, Ed. da Universidade de São Paulo, 1981. p. 131-145.

_____. Léxico e vocabulário fundamental. *Alfa*, São Paulo, v. 40, p. 27-46, 1996.

_____. *Dicionário didático de português.* 2 ed. São Paulo: Ática, 1998a.

_____. A face quantitativa da linguagem: um dicionário de freqüências do português. *Alfa*, São Paulo, n.42, n.esp. p.161-181, 1998b.

_____. As ciências do léxico. In: OLIVEIRA, A. M. P. P.; ISQUIERDO, A. N. (Orgs.). *As ciências do léxico*: lexicologia, lexicografia, terminologia. Campo Grande: Ed. UFMS, 1998c. p.11-20.

BINOT, J. L. et al. Représentation sémantique et interpretation dans une interface en langage naturel. *Le Français Moderne*, Paris, v.59, n.1, p. 57-84, 1991.

BOGAARDS, P. *Le vocabulaire dans l'apprentissage des langues étrangères*. France: Hatier/Didier, 1994.

BUTTON, G. et al. *Computadores, mentes e conduta*. Tradução de Roberto Leal Ferreira. São Paulo: Ed. Unesp, 1998.

CALZOLARI, N. et al. Acquiring and representing semantic information in a lexical knowledge base. In: WORKSHOP ON LEXICAL SEMANTICS AND KNOWLEDGE REPRESENTATION, 1., 1991, Berkeley. *Proceedings*... California: ESPRIT BRA-3030 ACQUILEX WP N.016, 1991.

CALZOLARI, N. et al. SI-TAL- Documento di specifiche tecniche di SI-TAL: manuale operativo. In: _____. *ItalwordNet*: rete semantico-lessicale per l'italiano. Piza: Consorzio Pisa Ricerche (CPR), Istituto Trentino di Cultura, Istituto per la Ricerca Scientifica e Tecnologica – (ITC-INST.), 2000. Capitolo 2.

CARVALHO, P. C. Q. da F. *Gramáticas de resolução de ambigüidades resultantes da homografia de nomes e adjetivos*. 2001. Dissertação (Mestrado) – Faculdade de Letras da Universidade de Lisboa, Lisboa, 2001.

DEL FIORENTINO, M. C. *Estrazione di informazione semantica da un dizionario-macchina della lingua italiana*: problemi di disambiguazione e di riorganizzazione delle tassonomie semantiche. 1995. Tesi (Laurea) – Università degli Studi di Pisa, Facoltà di Lettere e Filosofia, Piza, 1995.

HAGÈGE, C.; DUARTE, I. Construção de gramáticas formais para o processamento da linguagem natural. MATEUS, M. H.; BRANCO, A. H. (Org.) *Engenharia da Linguagem*. Lisboa: Colibri, 1995. p. 71-93.

HATHOUT, N. Pour la construction d'une base de connaissances lexicologiques à partir du Trésor de la Langue française: les maqueurs superficiels dans les définitions spécialisées. *Cahier de lexicologie*: Revue Internationale de Lexicologie et de Lexicographie, Paris, v.68, p. 137-173, 1996.

LEFFA, V.J. A resolução da ambigüidade lexical sem apoio do conhecimento de mundo. *Revista Intercâmbio*, São Paulo, v.6, pte1, p. 869-889, 1998.

LENCI, A et. al. *SIMPLE – Semantic Information for Multifunctional Plurilingual Lexica*: linguistic specifications: deliverable D2.1. Pisa: University of Pisa and Institute of Computational Linguistics of CNR, 1999.

MEDEIROS, M. B. B. *Tratamento automático de ambigüidades na recuperação da Informação*. 1999. Tese (Doutorado) – Universidade de Brasília, Brasilia, 1999.

PICOCHE, J. *Precis de lexicologie française*: l'étude et l'enseignement du vocabulaire. Paris: Nathan, 1992.

PUSTEJOVSKY, J. *The generative lexicon*. Cambridge: The MIT Press, 1995.

ZAVAGLIA, C. A homonímia e o computador. *Estudos Lingüísticos*, São Paulo, v.28, p. 738-743, 1999.

_____. *Análise da homonímia no português*: tratamento semântico com vistas a procedimentos computacionais. 2002. Tese (Doutorado) – Faculdade de Ciências de Letras – UNESP, Araraquara, 2002.

Bibliografia consultada

BIDERMAN, M.T.C. Polissemia versus homonímia. In: SEMINÁRIO DO GEL, 38., 1991, Franca. *Anais...* Franca: União das Faculdades Francanas, 1991. p. 283-290.

_____. *Dicionário de freqüências do português contemporâneo.* [S. l: s.n.], 1997. 1 Disquete

_____. O dicionário como norma na sociedade. In: ENCONTRO NACIONAL DO GT DE LEXICOLOGIA, LEXICOGRAFIA E TERMINOLOGIA DA ANPOLL, 1., 1997, Rio de Janeiro. *Anais...* Rio de Janeiro: Ed. Universitária UFPE, 1997. p.161-180.

_____. Os dicionários na contemporaneidade: arquitetura, métodos e técnicas. In: OLIVEIRA, A. M. P. P.; ISQUIERDO, A. N. (Org.). *As ciências do léxico:* lexicologia, lexicografia, terminologia. Campo Grande: Ed. UFMS, 1998. p.129-142.

BOGURAEV, B. et al. Acquisition of lexical knowledge for natural language: processing systems. In: *Techinical Annex. ESPRIT BRA – 3030,* Cambridge (UK), 1988.

CHISHMAN, R. L. de O. *A teoria do léxico gerativo:* uma abordagem crítica. 2000. Tese (Doutorado) – Pontífice Universidade Católica do Rio Grande do Sul, Porto Alegre, 2000.

EVENZ, M. W. (Ed.). *Relational models of the lexicon:* representing knowledge in semantic networks. Cambridge: Cambridge University Press, 1988.

REHFELDT, G. K. *Polissemia e campo semântico:* estudo aplicado aos verbos de movimento. Porto Alegre: EDURGS/FAPA/FAPCCA, 1980.

RICH, E. *Inteligência artificial.* Tradução de Newton Vasconcelos. São Paulo: McGraw-Hill, 1988.

SPANU, A. *Pluridimensionalità delle tassonomie del dizionario.* Pisa, 1995. ILC – LDB n.2 (T.152), CNR – ILC.