A CONSTRUÇÃO DE UM *THESAURUS* ELETRÔNICO PARA O PORTUGUÊS DO BRASIL

Bento Carlos DIAS-DA-SILVA[1]
Helio Roberto de MORAES[2]

■ RESUMO: Este trabalho discute o equacionamento lingüístico da construção de um *Thesaurus Eletrônico para o Português do Brasil*. Na introdução, contextualizamos esse equacionamento no domínio da pesquisa em processamento automático de línguas naturais. Na seqüência, apresentamos, na seção 2, a representação lingüístico-computacional da sinonímia e da antonímia e exemplificamos o processo de compilação dessas relações de sentido em dicionários do português do Brasil. Na seção 3, justificamos a seleção dos dicionários enquanto o *corpus* de referência e propomos uma tipologia dos problemas decorrentes da escolha desse tipo de *corpus* para a montagem do *thesaurus*. Na seção 4, complementamos a discussão com a descrição das principais características de uma ferramenta computacional de autoria, projetada para agilizar o processo de montagem da base de dados lexicais do *thesaurus*: o Editor do *Thesaurus*. Por fim, pontuamos as estatísticas atuais do *thesaurus* e futuros desdobramentos.

■ PALAVRAS-CHAVE: *Thesaurus* eletrônico; sinonímia; antonímia; *WordNet*.

Introdução[3]

Este artigo descreve as principais etapas lingüísticas envolvidas na construção de um *thesaurus*[4] eletrônico para o Português do Brasil (TeP), um tipo específico de dicionário eletrônico de sinônimos e antônimos, que, acoplado a ferramentas computacionais de auxílio à expressão escrita, soma-se a outras obras de referência em meio digital, como dicionários e gramáticas (FLEXNER, 1997). O TeP tem por finalidade oferecer ao usuário da língua portuguesa a opção *on line* de palavras sinônimas e antônimas que ele, por motivos de estilo, precisão, adequação comunicativa, correção ou aprendizagem, desejar substituir (ILARI; GERALDI, 1985).

A construção do TeP tomou por base Saint-Dizier e Viegas (1995) e Dias-da-Silva

1 Departamento de Letras Modernas – Faculdade de Ciências e Letras – UNESP – 14800-901 – Araraquara – SP – Brasil. E-mail: bento@fclar.unesp.br.
2 Programa de Pós-Graduação em Lingüística e Língua Portuguesa – Faculdade de Ciências e Letras (Mestrando-CNPq) – UNESP – 14800-901 – Araraquara – SP – Brasil. E-mail: helio_de_moraes@ig.com.br,
3 Agradecemos ao revisores anônimos pelas sugestões que contribuíram para a lapidação do texto.
4 Projeto desenvolvido no NILC (Núcleo Interinstitucional de Lingüística Computacional) com apoio da FINEP, Programa PADCT-III-CDT/ MCT, PROCESSO RC: 3.1.3-0012/98 - Convênio: 8.8.98.059.00.

(1998) que, além de proporem uma metodologia específica, fornecem os subsídios lingüísticos e computacionais essenciais para o desenvolvimento de projetos interdisciplinares de elaboração de sistemas computacionais que visam à simulação de fenômenos e fatos da linguagem humana.

Como decorrência desse recorte teórico-metodológico, a complexa tarefa de compilação do TeP foi decomposta em um conjunto de atividades complementares, agrupadas, segundo sua natureza, em três domínios: lingüístico, da representação e da implementação (DIAS-DA-SILVA, 1998). Nos domínios lingüístico e da representação, as atividades de pesquisa concentraram-se na fundamentação, delimitação, extração, filtragem e representação formal do que denominamos "conhecimentos lingüístico e metalingüístico", em oposição aos "conhecimentos computacional e metacomputacional".

No domínio da implementação, as atividades, fundamentadas em estratégias e resultados de discussões delineados nos domínios anteriores, foram subdivididas em três tarefas distintas. A primeira, eminentemente computacional, consistiu na implementação de uma ferramenta computacional de autoria para a montagem da base de dados lexicais do TeP, isto é, uma base relacional de dados, no sentido computacional do termo, que contém os dados e a representação computacional interna do TeP. Essa ferramenta desempenhou três funções bastante distintas: função de editor, que possibilitou ao lingüista inserir e editar os verbetes do thesaurus; função de sistema de coleta e gerenciamento de dados, pela qual a ferramenta armazena os verbetes inseridos pelo lingüista sob a forma de uma base relacional de dados; função de gerador, que, a partir dos verbetes inseridos na base, torna-a capaz de gerar automaticamente novos verbetes. A segunda tarefa, essencialmente lingüística, concentrou-se na inserção de verbetes na base, tarefa que, como veremos oportunamente, consistiu em inserir conjuntos de sinônimos e antônimos. A terceira, por fim, também computacional, consistiu na implementação de rotinas computacionais cuja finalidade é converter a base no TeP propriamente dito.

Neste trabalho, restringimos a discussão a três dos principais problemas enfrentados nos Domínios Lingüístico e Representacional: (i) a especificação de uma representação lingüístico-computacional das relações de sinonímia e antonímia, que são as relações constitutivas e estruturadoras do TeP; (ii) o processo de extração dessas relações de um conjunto de quatro dicionários do português, selecionados como o *corpus* de referência; (iii) a caracterização e o enfrentamento dos problemas mais recorrentes durante o processo de refinamento do processo de compilação de sinônimos e antônimos para a montagem do TeP. Para complementar a exposição, esboçamos o editor do *thesaurus*, uma ferramenta computacional de autoria que auxilia na montagem dos verbetes e contribui para minimizar as principais inconsistências observadas em obras semelhantes publicadas em meio impresso.

A próxima seção apresenta a representação formal das relações de sinonímia e antonímia que tornou possível a implementação do editor e da conseqüente montagem e gerenciamento computacional da base de dados lexicais do TeP.

Os conjuntos de sinônimos

A busca de resolução do problema que acabamos de explicitar na seção anterior foi motivada pela construção da rede *WordNet*, descrita em Miller e Fellbaum. (1991). Desse empreendimento, utilizamos três constructos básicos O *synset*, isto é, um conjunto de sinônimos ou quase-sinônimos (inglês *synonym set*), elemento formal que possibilita a representação computacional dessa relação. A "matriz lexical", outro elemento formal que especifica uma correspondência biunívoca entre sentido e *synset*. A idéia de matriz lexical parte da hipótese de que, dado um *synset* bem-formado, o falante é capaz de inferir, a partir das unidades lexicais que o compõem, o sentido expresso pelo conjunto. Trata-se do princípio psicolingüístico de ativação de conceitos, na mente do falante, por meio da interpretação do conjunto de formas lexicais relacionadas pela sinonímia. Dessa forma, não há necessidade de se explicitar o valor semântico de cada conjunto de sinônimos por meio de um rótulo conceitual ou de uma definição. Por fim, a "indexação rotulada", que formaliza a relação de antonímia por meio de indexadores rotulados especificados entre pares de *synsets* que apontam para sentidos opostos.

Do ponto de vista formal, a rede *WordNet* pode ser entendida como uma base relacional de dados que sistematiza uma parcela do léxico de uma língua – substantivos, verbos, adjetivos e advérbios – em termos de uma rede de quatro relações: sinonímia, antonímia, hiponímia e meronímia (LYONS, 1979; CRUSE, 1986).

O construto básico dessa base, o *synset*, é responsável pela estruturação da rede. É importante salientar que o *synset* não define um conceito, mas fornece informação suficiente para que os locutores identifiquem o conceito por ele evocado. Vale também observar que a noção de sinonímia adotada é aquela que preconiza que dois termos são sinônimos se existir algum contexto em que ambos puderem ser intersubstituíveis, sem que haja alteração substancial do significado, posto que, em última instância, são os locutores que decidem o grau de sinonímia existente entre as expressões de uma língua (CRUSE, 1986, ILARI; GERALDI, 1985).

Do ponto de vista computacional, os *synsets* são conjuntos munidos de dois tipos de ponteiros que representam dois tipos de relações entre os conjuntos: ponteiros que especificam relações léxico-semânticas (sinonímia e antonímia), relações entre formas, e ponteiros que especificam relações conceptuais (hiponímia e meronímia), relações entre conceitos atualizados por formas.

Do ponto de vista da implementação, a rede *WordNet* é composta de: (a) arquivos preparados por lexicógrafos (ALs), (b) um programa que converte esses arquivos em uma base de dados (DB), (c) rotinas de busca e (d) interfaces para a apresentação da informação a partir da base de dados. Nos ALs, substantivos, verbos, adjetivos e advérbios estão sistematizados em conjuntos de sinônimos; a relação de antonímia, quando pertinente, é especificada entre pares desses conjuntos. O programa que converte os ALs na DB é também responsável pela codificação dessas relações. As diferentes

interfaces de acesso à DB utilizam uma biblioteca comum de rotinas criadas para exibir os diversos tipos de relação.

Devido à adoção do modelo de representação proposto para a rede *WordNet*, a laboriosa tarefa de construção da base do TeP, em termos operacionais, ficou reduzida à especificação de conjuntos (os sinônimos) e de relações entre conjuntos (os antônimos). O esquema, a seguir, ilustra a estrutura típica de um verbete:

Entrada n (categoria X)
Acepção n.1
Conjunto de Sinônimos
Conjunto de Antônimos

Acepção n.m
Conjunto de Sinônimos
Conjunto de Antônimos

Nesse esquema, n é o número de identificação da entrada, X representa uma das quatro categorias gramaticais, substantivo, verbo, adjetivo ou advérbio, e n.1... n.m são os números de identificação das acepções da entrada n.

Criando conjuntos de sinônimos

Esta seção tem por objetivo exemplificar o processo de seleção e filtragem da informação lexical para a base do TeP. Tomamos como fonte de informações Weiszflog (1998), um dos componentes do *corpus* de referência, que será apresentado na seção 3. Ressaltamos que, embora a base do TeP seja composta de substantivos, adjetivos, verbos e advérbios, neste artigo vamos focalizar nossa atenção na categoria verbal.

A extração de informações léxico-semânticas a partir de verbetes de dicionários exigiu a observância de dois princípios. O primeiro refere-se ao cuidado que precisamos tomar quando analisamos as definições utilizadas nos verbetes durante o procedimento de extração da informação léxico-semântica pertinente para o *thesaurus*. A análise dos verbetes das obras de referência demonstrou que é freqüente a substituição de sinônimos por paráfrases. Por exemplo, no verbete "prolongar", a primeira acepção diz: "dar maior comprimento". Essa paráfrase é o mesmo que "encompridar", cuja definição, no mesmo dicionário, é "tornar mais comprido".

O segundo refere-se à importância de se considerar o componente aspectual do significado de cada vocábulo ou expressão, pois o aspecto é parte integrante do seu significado, não podendo ser ignorado. Por exemplo: "cochichar" é definido como "falar em voz baixa". Embora "cochichar" seja definido como "falar", não podemos dizer que "cochichar" seja sinônimo de "falar", pois não se trata de sinonímia, mas de troponímia, isto é, uma relação de sentido definida por "*x* é *y* de um certo modo" (MILLER;

FELLBAUM, 1991), ou seja, "cochichar" é o mesmo que "falar" de um certo modo. Mas essa restrição deve ser observada com cautela, pois há casos em que não estamos diante da troponímia. Por exemplo, não há a relação aspectual de troponímia entre "labutar" e "trabalhar com intensidade", porque "labutar" não é o mesmo que "trabalhar" de um certo modo.

Feitas essas considerações, tomemos um exemplo concreto, o verbo "lembrar", para ilustrar o procedimento de filtragem. Partimos do seguinte verbete do dicionário:

lembrar

v. 1. Tr. dir. Trazer à memória; recordar. 2. Tr. ind. Vir à idéia, tornar-se recordado. 3. Pron. Recordar-se, ter lembrança de. 4. Tr. dir. Fazer vir à memória por analogia ou semelhança. 5. Tr. dir. Advertir, notar. 6. Tr. dir. Sugerir. 7. Tr. dir. Recomendar.

Ao examinarmos a informação do verbete, identificamos quatro acepções, representadas em termos dos seguintes conjuntos de sinônimos:

{lembrar, recordar}
{lembrar, advertir, notar}
{lembrar, sugerir}
{lembrar, recomendar}

Observamos que a acepção 3 apresenta uma forma pronominal, com o sentido de "processo", o que nos autoriza construir o conjunto:

{lembrar-se, recordar-se}

Terminada essa montagem preliminar dos conjuntos, passamos a verificar a consistência da informação extraída do verbete "lembrar". Para isso, o próximo passo consiste em consultar, preferencialmente nesta ordem, os seguintes verbetes "recordar", "recordar-se", "advertir", "notar", "sugerir" e "recomendar", processo fundamental para a ampliação dos conjuntos de sinônimos.

Tomemos, então, o verbete "recordar":

re.cor.dar

v. 1. Tr. dir. Trazer à memória. 2. Pron. Lembrar-se. 3. Tr. dir. Fazer lembrar; ter analogia ou semelhança com; parecer. 4. Tr. ind. Lembrar.

Essa consulta confirma os dois conjuntos existentes, {lembrar, recordar} e {lembrar-se, recordar-se}, e permite construir um novo conjunto: {recordar, parecer}.

Esse procedimento deve prosseguir até esgotarmos todos os verbetes "atingíveis" a partir do verbete "lembrar". Terminado esse procedimento, retomamos a ordem alfabética.

Suponhamos, agora, que estamos consultando o verbete "esquecer":

es.que.cer

v. 1. Tr. dir. Deixar sair da memória; perder a memória de; tirar da lembrança; olvidar. 2. Pron. Perder a lembrança ou a memória; olvidar-se. 3. Tr. dir. Não fazer caso de, pôr em esquecimento. 4. Tr. ind. e intr. Escapar da memória, ficar em esquecimento: Esqueceu-lhe o final do discurso. Seu prestígio foi momentâneo, passou e esqueceu. 5. Tr. dir. Descurar-se de: Não esquecia as suas tarefas. 6. Pron. Perder a ciência ou a habilidade adquiridas: Já me esqueci do latim. 7. Pron. Descuidar-se: Meu secretário esqueceu-se de tudo. 8. Intr. Ficar dormente ou tolhido, perder a sensibilidade: Naquela má posição a perna esqueceu.

Filtrando a informação desse verbete, obtemos os seguintes conjuntos:

{esquecer, olvidar},
{esquecer-se, olvidar-se},
{esquecer-se, descuidar-se, descurar-se}.

Apesar do verbete apresentar "descurar-se" e "descuidar-se" em acepções diferentes, a inserção de "descuidar-se" e "descurar-se" em um mesmo conjunto justifica-se por duas constatações: esse mesmo dicionário apresenta "descurar-se" como sinônimo de "descuidar-se" no verbete "descurar".

Note-se que, em nenhum dos verbetes transcritos, foram mencionados antônimos. Mas a oposição de sentido entre "lembrar/esquecer" é evidente. Esse fato é, entretanto, registrado por meio de paráfrases. Com efeito, no verbete "lembrar", lemos "trazer à memória" e, no verbete "esquecer", várias paráfrases são apresentadas: "deixar sair da memória; perder a memória de; tirar da lembrança". Isso nos autoriza estabelecer entre os conjuntos {lembrar, recordar} e {esquecer, olvidar} a relação de antonímia.

Ressaltamos que o procedimento de seleção e filtragem do verbete "lembrar" aqui descrito é apenas um recorte. Nossos "percursos" por todas as obras do nosso *corpus* de referência permitiram a montagem do conjunto {lembrar, amentar2, recordar, relembrar, rememorar, ver}.

A próxima seção apresenta os dicionários que compõem o *corpus* de referência.

Seleção e filtragem de informações léxico-semânticas

O corpus de referência

A compilação de dicionários, em geral, baseia-se em *corpus*, que são utilizados durante o procedimento de montagem dos verbetes e da complicadíssima discriminação dos diferentes sentidos que neles devem ser contemplados. Desnecessário dizer

que esse é um trabalho que demanda tempo e uma grande equipe. Uma alternativa para essa prática, proposta no âmbito da lexicografia computacional, é a reutilização de dicionários já existentes como *corpus* de referência (BRISCOE; BOGURAEV, 1989). Das obras disponíveis e analisadas, escolhemos Weiszflog (1998), Ferreira (1999), Barbosa (1999), Fernandes (1997) e Borba (1990).

Alguns fatos justificam a escolha. Essas obras são inegavelmente fontes de conhecimento lexical, seguem uma tradição centenária para a compilação dos verbetes, privilegiam o emprego de sinônimos e antônimos na especificação das diferentes acepções das entradas, foram elaboradas por significativo corpo de lexicógrafos e parte delas está disponível em meio digital, o que agiliza a extração da sinonímia e antonímia durante a montagem da base do TeP.

A escolha dessas obras como *corpus* de referência (C. R), porém, não está livre de problemas, uma vez que a maioria delas apresenta incoerências, lacunas e imprecisões. Ferreira (1999), por exemplo, foi severamente criticado por Cláudio Abramo, em matéria publicada no caderno "Mais" da *Folha de São Paulo* (23/01/2000). Como veremos nas seções seguintes, a reutilização de dicionários já prontos como ponto de partida para a extração das informações pertinentes exigiu grande cuidado para que fossem transportadas para a base do TeP informações seguras.

Origem dos problemas

Não podemos negar que os dicionários sejam importantes fontes de informação lexical. Sua utilização para fins de extração de informação de natureza lingüística, entretanto, requer cautela. Parte do cuidado está no fato de observarmos que, em geral, dicionários são produtos comerciais. Em decorrência disso, acabam por seguir um padrão tradicionalmente aceito por consumidores (KILGARRIFF, 1997), e caracterizam-se por imprecisões que mesmo um usuário não especializado em práticas lexicográficas consegue detectar (ABRAMO, 2000).

As tarefas complexas que o lexicógrafo tem de enfrentar em seu cotidiano concentram-se nas seguintes: delimitar o número de acepções que consegue isolar para cada palavra selecionada para figurar como entrada e, uma a uma, defini-las e exemplificá-las com abonações. Como resume Kilgarriff (1997, p.102), para realizar essa tarefa, delimita um *corpus* e nele mergulha para garimpar seus lexemas. Nessa tarefa, utiliza-se dos seguintes procedimentos:

- reunir as concordâncias para o lexema a ser descrito;
- separar essas concordâncias em grupos, de tal forma que os membros de cada grupo compartilhem o maior número de traços morfossintáticos e semânticos;
- avaliar, para cada grupo, os traços que mantêm seus elementos unidos;
- codificar as descrições na metalinguagem da lexicografia.

O maior problema é a delimitação precisa do sentido de cada grupo de concordância, posto que, em geral, não há uma separação nítida entre os sentidos que veiculam. Para definir quais grupos serão "convertidos" em acepções do verbete, os lexicógrafos buscam os sentidos mais *freqüentes* no uso, e menos *previsíveis* a partir de outros sentidos (KILGARRIFF, 1997). Porém, como cada dicionário segue estratégias de trabalho e padrões de excelência próprios, oriundos de decisões e escolhas muitas vezes *ad hoc*, a comparação entre verbetes de dicionários diferentes, e mesmo entre verbetes de um mesmo dicionário, apresenta diferenças consideráveis.

Embora o processo descrito acima não seja diretamente parte da rotina de compilação das entradas do TeP, sua compreensão é importante na análise dos tipos de distinções de acepções, e, portanto, de sentidos que podem ser encontrados nos dicionários de referência para a montagem da base de dados do TeP. Kilgarriff (1993), trabalhando com entradas do *Longman Dictionary of Contemporary English* (SUMMERS, 1995), encontrou quatro[5] categorias diferentes de distinção de sentido, das quais três podem ser aplicadas a verbos. Essas categorias também são válidas para as obras de referência do TeP:

- *Metáfora Generalizante* – distinção entre um sentido específico, que é a palavra certa em um determinado contexto, e um sentido mais geral, que pode ser atribuído a uma série de situações; por exemplo, o verbete *marretar* (WEISZFLOG, 1998) apresenta: *1. Bater com marreta em,* sentido mais específico (...). *3. Espancar* (...), sentido menos específico.
- *Informação Pressuposta (Must-be-there)* – se existe uma situação em que um sentido de um lexema pode ser aplicado, então é uma conseqüência lógica que outro sentido também possa ser aplicado para outro aspecto da mesma realidade, como no verbete *casar* (WEISZFLOG, 1998): *v. 1. Tr. dir. Ligar pelo casamento, promover o casamento de. 2. Tr. dir. Realizar o casamento de. 3. Tr. dir. e pron. Aliar (-se), ligar(-se).(...).* Nesse verbete, temos a ação sendo tomada do ponto de vista de quem realiza o casamento (acepções 1 e 2) e de quem se casa (acepção 3).
- *Mudança de Domínio* – essa distinção pode ser observada entre duas situações de uso de um lexema, cujos sentidos são de tal forma distantes entre si, que o lexicógrafo decide por estabelecer duas acepções diferentes, ainda que alguém possa argumentar que se trata de uma adaptação de sentido possível do mesmo lexema, dada uma situação ou entidade diferente que deve ser descrita. Weiszflog (1998), no verbete *levar,* registra: *v. 1. Tr. dir. Conduzir algo consigo de*

[5] A quarta distinção, *Tipo,* é aplicada com maior freqüência a substantivos. Isso pode ser reflexo da organização semântica específica dessa categoria, que segue uma hierarquização em que os conceitos lexicalizados se organizam em níveis, partindo dos mais abstratos para os mais específicos (MILLER; FELLBAUM, 1991).

um lugar para outro. 2. Tr. dir. Afastar, retirar. 3. Tr. dir. Arrastar, puxar. 4. Tr. dir. Conduzir, guiar.(...). Nesse verbete podemos observar que as acepções 1 e 4 possuem uma intersecção de sentido, mas a especificidade de cada uma justifica a separação, pois a acepção 1 pode perfeitamente ser substituída por *carregar*, o que não ocorre na acepção 4.

Essas distinções são recorrentes nas obras de referência escolhidas para a montagem da base de dados do TeP, e, muitas vezes, várias classificações são aplicáveis a uma mesma distinção, como veremos na seção 3.3. Para a compilação do TeP esse é um ponto importante: quanto mais distinções de sentido forem identificadas, mais segura será a informação oferecida pelo TeP. Portanto, essas distinções de sentido são um referencial para a análise dos problemas que surgem na difícil tarefa de delimitar o número de acepções, e, para cada acepção, especificar o seu valor semântico.

Tipos de problemas

Essa seção pretende apresentar os problemas mais recorrentes encontrados na tarefa de extração de informação lingüística dos verbetes dos dicionários. Há três tipos centrais de problemas, que Kilgarriff (1993) denomina: necessidade, consistência e centralidade.

Na compilação das entradas do TeP, é tarefa essencial refletir se um determinado traço semântico ou gramatical é condição necessária para um lexema em um determinado sentido. A importância de observarmos essa questão decorre da noção de sinonímia adotada, isto é, dois termos são sinônimos se existir um contexto em que os dois possam ser substituídos sem que haja prejuízo da significação (LYONS, 1979; ILARI; GERALDI, 1985). Isso limita sobremaneira o número de especificações gramaticais que um lexema deve ter para poder ser inserido em uma determinada acepção.

A verificação da consistência das entradas dos dicionários para a compilação do TeP implica basicamente a observação da simetria, uma característica importante da sinonímia, nem sempre observada pelos dicionários, que preconiza que: se A é sinônimo de B, B é obrigatoriamente sinônimo de A (MILLER; FELLBAUM, 1991).

O problema de centralidade de cada acepção de um verbete refere-se ao limite possível de variação de sentido admitido pela acepção. Esse limite corresponderia a uma suposta linha divisória de separação entre duas acepções. Esse problema é recorrente na construção do TeP, dado que não se pode considerar a sinônima uma relação transitiva: A é sinônimo de B, B é sinônimo de C, mas C não é sinônimo de A (LYONS, 1979).

Além de problemas como esses, cuidados, para evitar a transposição de im-

precisões para o TeP, devem ser observados. Na entrada do verbo "delimitar", por exemplo, Weiszflog (1998), equivocadamente, indica o verbo "extremar" como seu sinônimo. Note-se, porém, que "extremar", no sentido de "tornar extremo", não é sinônimo de "delimitar", que significa "demarcar". Trata-se de uma imprecisão ortográfica, posto que o sinônimo pretendido para o verbo "delimitar" é o verbo "estremar", grafado com "s". Weiszflog (1998) é o único dicionário do *corpus* de referência que registra a forma "reqüestar" com o trema. Borba (1990, p. 683) registra "espanadar" ao invés de "espadanar".

Além dessas imperfeições de natureza ortográfica, há problemas no tratamento da homonímia e polissemia. Em outras palavras, os dicionários consultados nem sempre são consensuais quanto ao número de entradas que devem abrir para uma mesma forma (o problema da homonímia) ou ao número de acepções que registram para uma mesma entrada (o problema da polissemia). Para "apontar", por exemplo, Weiszflog (1998) apresenta três entradas, Ferreira (1999) duas, e Borba (1990) apenas uma.

Essas arestas são, em geral, aparadas durante o processo de compilação da base de dados lexicais do TeP. Porém, os casos que envolvem as três classes de problemas acima listados e os tipos de distinção de sentidos descritos na seção 3.2 apresentam dificuldades bem maiores para os compiladores do TeP.

São esses casos que pretendemos apresentar na seção seguinte.

A filtragem da informação para o TeP

Os tipos de problemas exemplificados a seguir foram encontrados durante todo o processo de compilação dos verbos.

O primeiro tipo refere-se à metáfora generalizante. Os lexemas "acarar, encarar, arrostar" possuem o sentido de "ficar face a face", e também possuem o sentido de "enfrentar"; portanto, esses dois sentidos poderiam ser incluídos em um mesmo conjunto na base de dados do TeP; por exemplo, poderíamos criar o conjunto {acarar, encarar, arrostar, confrontar, enfrentar}. Isso ocorre porque lexemas muito específicos, como "acarar", passaram a denotar um sentido menos específico. Esse fato poderia gerar um problema para o usuário do TeP, pois ele não teria como definir o sentido de "ficar face a face".

O inverso da metáfora generalizante também ocorre, ou seja, um sentido menos específico também denotar um sentido mais especificado. Por exemplo: os dicionários Ferreira (1999) e Weiszflog (1998) nos permitem, a partir da entrada "abastardar", sugerir o seguinte conjunto de sinônimos {abastardar, alterar, corromper, decompor}. Porém, a pesquisa, a partir desses elementos, mostra que "alterar" nem sempre é registrado como sinônimo de "abastardar". Esses dois exemplos demonstram como a identificação da metáfora generalizante é útil na resolução de problemas de centralidade de significado.

Os casos relacionados à metáfora generalizante, que geram dois conjuntos com elementos comuns e podem eventualmente confundir o usuário quanto ao valor semântico de cada conjunto, podem ser sanados facilmente com a inserção de frases-tipo para cada acepção, um refinamento previsto para o futuro. Por essa razão, a inserção das duas acepções, mesmo que semelhantes, é imprescindível para o TeP, e, portanto, com relação ao exemplo citado, foram criados os dois conjuntos: {acarar, encarar, arrostar} e {acarar, encarar, arrostar, confrontar, enfrentar}.

O problema de inserção do lexema "alterar" no conjunto {abastardar, corromper, decompor}, é que ele seria o único elemento do conjunto que não carregaria um traço disfórico. O fato é que "abastardar", "corromper" e "decompor" significam "alterar de um certo modo". Portanto, "alterar" não foi inserido no conjunto, pois esse lexema não se relaciona por sinonímia com os demais elementos do conjunto, e sim, por outra relação de sentido: a troponímia (MILLER; FELLBAUM, 1991, p.216).

O tipo de problema decorrente da interpretação de informação pressuposta pode ser percebido neste exemplo selecionado no verbete "visualizar". Para ele, Borba (1990) apresenta uma única acepção: "perceber pela visão, conceber (sem ver) uma imagem mental de". O primeiro segmento da definição ("perceber pela visão") é claramente sinônimo de "ver", fato confirmado pelo exemplo: "Assusteime ao visualizar à minha frente à imagem de dois homens de clã". Observe que nesse exemplo podemos substituir "visualizar" por "ver", sem nenhum prejuízo para o sentido da frase. Não é possível, porém, a substituição por "imaginar", que é sinônimo do segundo segmento da definição: "conceber (sem ver) uma imagem mental de". Observe o exemplo: "podemos talvez alimentar a esperança de visualizar/imaginar todas as novas dimensões da realidade". Com clareza, Borba (1990) identificou os dois sentidos e os abonou com exemplos claros. O problema está no fato de ter mantido os dois sentidos diferentes em uma mesma acepção, talvez por julgar que o primeiro sentido ("perceber pela visão") fosse suficientemente pressuposto a partir do sentido do lexema "visão", explícito no radical do verbo. Os outros dicionários do nosso *corpus* de referência apresentam apenas a acepção de "imaginar". Como foram claramente identificados dois sentidos distintos, inseriram-se dois conjuntos diferentes na base de dados do TeP: {ver, visualizar, enxergar,...}, e, {ver, visualizar, imaginar}.

Foram encontradas situações inversas. Há entradas para as quais os dicionários apresentam duas acepções distintas, mas só é possível a identificação de um único sentido. Borba (1990, p.1330), por exemplo, apresenta os lexemas "forçar", "obrigar" e "impelir" para definir uma das acepções de "urgir", abonada pelo exemplo: "Urgiam-nos de todos os lados para que caminhássemos". Observe-se que todos os itens lexicais sinônimos, sugeridos no dicionário, são intersubstituíveis no co-texto da frase.

O problema é que Weiszflog (1998), que também registra esses mesmos itens lexicais como sinônimos (na acepção cinco da mesma entrada), registra, na acep-

ção sete, o mesmo exemplo encontrado em Borba (1990, p.1330), cujo sentido é definido pelos itens lexicais: "empurrar" e "compelir".

A pesquisa, em cada um dos verbetes, nos revelou que podemos sugerir o conjunto: {urgir, compelir, forçar, obrigar, impelir,...}, mas a entrada de "empurrar" nos dicionários não possui o sentido do conjunto, e nenhum outro sentido que remeta a "urgir". Sendo assim, apesar de Weiszflog (1998) discriminar duas acepções diferentes, só conseguimos estabelecer um sentido.

Nesse exemplo, identificamos dois tipos de problemas: (i) problema da centralidade, pois o problema central é definirmos se "empurrar" deve ou não constar naquela acepção; (ii) problema da consistência, pois Weiszflog (1998) estabelece duas acepções para apenas um sentido. Como solução para o TeP, inserimos apenas uma acepção, como no caso do conjunto {urgir, compelir, forçar, obrigar, impelir,...}. O lexema "empurrar" não foi inserido no conjunto por não termos identificado qualquer contexto de ocorrência em nosso *corpus* de referência.

O terceiro problema, o da necessidade, abarca uma mudança de domínio no uso do verbo "exalar", que significa "emitir ou lançar de si emanações odoríficas ou fétidas". De acordo com essa definição, esse verbo deveria ser igualmente inserido em conjuntos relacionados na base de dados do TeP por antonímia: {feder, catingar} e {recender (exalar cheiro bom)}, o que geraria uma incoerência. Para solucionar o problema, considera-se, então, que esse verbo exige complemento específico nesse sentido. Fato semelhante ocorre com o verbo "cheirar". Observem-se os exemplos: "o cadáver já está cheirando" e "o assado já está cheirando". A criação do conjunto "neutro" {cheirar, exalar, trescalar,...}, com o sentido de "exalar cheiro forte" (bom ou ruim), parece ser a solução procurada.

A próxima seção apresenta o Editor do *Thesaurus* e suas principais funcionalidades. Esse aplicativo é uma ferramenta computacional através da qual os compiladores entram e verificam os conjuntos da base do TeP. As principais características da arquitetura computacional do editor são descritas em Dias da Silva et al. (2000). Restringimos a descrição às suas principais características, que, associadas à observância dos problemas e estratégias mencionados, contribuem para evitar que inconsistências semelhantes sejam inseridas no TeP.

O Editor do *Thesaurus*

Enquanto Editor do Thesaurus, a ferramenta de autoria é uma interface computacional gráfica para a montagem da base do TeP. Sua implementação foi possível graças ao modelo de representação formal que descrevemos na seção 2, uma vez que, no contexto desse modelo, as relações de sinonímia e antonímia passam a ter uma "existência" computacional. Com efeito, a relação de sinonímia é especificada pela relação de pertença que se estabelece entre formas da língua e o *synset* que as contém. Já a

relação de antonímia é convencionalmente especificada como uma relação entre pares de *synsets*.

Durante o processo de montagem, os recursos implementados no Editor possibilitaram construir, visualizar e editar os conjuntos de sinônimos e antônimos e verificar as estatísticas referentes ao número de verbetes, entradas e conjuntos contidos na base e a proporção nº de entradas/nº de conjuntos, para cada uma das categorias gramaticais especificadas.

Merece também destaque a geração automática de verbetes, pois, com esse recurso, o sinônimo digitado em um conjunto, que representa uma determinada acepção, caso ele ainda não tenha sido inserido como entrada, é automaticamente transformado em tal pelo Editor. Este se encarrega também de transportar para essa nova entrada tanto o conjunto em que esse sinônimo foi inserido como o conjunto de antônimos, associado a esse conjunto, se houver. Esse algoritmo é, portanto, responsável pela construção automática desse novo verbete.

Assim, a verificação dos sinônimos é também agilizada. Durante a coleta e seleção de sinônimos e antônimos, quando consultamos o verbete "recordar", por exemplo, nos dicionários", já sabemos que existe, na base de dados lexicais do TeP, a entrada recordar, pertencente ao conjunto {lembrar, recordar}, pois o Editor informa ao compilador todos os verbetes que constam da base. Para o compilador, a tarefa passa a ser, então, a identificação e inclusão de novos sinônimos e antônimos nas acepções apropriadas do verbete recordar, objetivando a complementação, se possível exaustiva, desse verbete.

É importante esclarecer que, embora descritos separadamente, os procedimentos de extração das informações lexicais e de inserção dos dados no Editor são realizados simultaneamente, pois a interface gráfica do Editor permite ao lingüista total controle visual do verbete que está em processo de montagem.

Considerações finais

Para finalizar, é importante ressaltar as dificuldades que tivemos de enfrentar para o desenvolvimento de um trabalho que envolve a manipulação de grande massa de dados: a falta de um *corpus* digital disponível, cuja existência poderia auxiliar o lingüista a contextualizar acepções pouco usuais e minimizar as inadequações apresentadas pelos dicionários, tarefa que consumiu tempo e exigiu cautela durante o procedimento de seleção e filtragem de informações.

Em termos quantitativos, a base de dados lexicais do TeP conta com mais de 19 mil conjuntos, responsáveis pela indexação de 44 mil entradas, assim distribuídas: 17 mil substantivos, 15 mil adjetivos, 11 mil verbos e mil advérbios.

Os ganhos com a montagem do TeP são também significativos. Ressaltamos o im-

portante relacionamento interdisciplinar entre lingüistas e cientistas da computação, essencial para o projeto e para a formação interdisciplinar de pesquisadores.

A base de dados lexicais do TeP, criada segundo o modelo da rede *WordNet*, constitui o ponto de partida para a construção da rede *WordNet* para o português do Brasil[6] (DIAS-DA-SILVA; OLIVEIRA; MORAES, 2002). Para esse salto qualitativo, está previsto o desenvolvimento de três etapas: (i) associar, para cada lexema que constitui cada conjunto de sinônimos, uma frase-tipo, extraída de *corpus*; (ii) atribuir, para cada conjunto, uma glosa, isto é, uma glosa e rótulo conceituall; (iii) especificar as relações de hiponímia, meronímia, causa, acarretamento e troponímia.

DIAS-DA-SILVA, B. C.; MORAES, H. R. de. Construction of a Brazilian Portuguese electronic thesaurus. *Alfa*, São Paulo, v.47, n.2, p.101-115, 2003

- *ABSTRACT: This paper examines the core problems involved in the linguistic task of compiling a Brazilian Portuguese Electronic Thesaurus. After presenting the natural language processing framework in which it is couched, it sets up the linguistic and computational representation for synonymy and antonymy, and describes the process of synonym and antonym mining from the lexical reference corpus, i.e. a set of four updated Brazilian Portuguese dictionaries. Next, it argues for the importance of reusing traditional published dictionaries in computational lexicon building, and, in the meantime, outlines the typology of the basic problems such a strategy poses for human compilers. Then, it is outlined the features of the thesaurus Editor, a specific authoring tool designed to help linguists feed the thesaurus database with the appropriate lexical information. Finally, it summarizes the thesaurus current lexical database statistics.*

- *KEY WORDS: Electronic thesaurus; synonymy; antonymy; WordNet.*

Referências bibliográficas

ABRAMO, C. Dicionários que horror. *Folha de São Paulo*, São Paulo, 23 de jan. 2000. Caderno Mais.

BARBOSA, O. *Grande dicionário de sinônimos e antônimos.* Rio de Janeiro: Ediouro, 1999.

BORBA, F.S. (Coord.) *Dicionário gramatical de verbos do português contemporâneo do Brasil.* São Paulo: Ed. Unesp, 1990.

BRISCOE, E. J.; B. BOGURAEV, (Eds.) *Computational lexicography for natural language processing.* London, New York: Longman, 1989.

CRUSE, D.A. *Lexical semantics.* New York: Cambridge University Press, 1986.

DIAS-DA-SILVA, B. C.; OLIVEIRA, M. F.; MORAES, H. R. Groundwork for the development of the

6 Projeto financiado pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Processo: nº 552057/01-0.

Brazilian Portuguese Wordnet. In: ADVANCES in natural language processing. Berlim: Springer-Verlag, 2002. p. 189-196.

DIAS-DA-SILVA, B. C. et al. Construção de um thesaurus eletrônico para o português do Brasil. In: ENCONTRO PARA PROCESSAMENTO DA LÍNGUA PORTUGUESA ESCRITA E FALADA, 5., 2000, Atibaia. *Anais*... São Carlos: Ed. ICMC/USP, 2000. p.1-11.

DIAS-DA-SILVA, B. C. Bridging the gap between linguistic theory and natural language processing. In: PROCEEDINGS OF THE 16th INTERNATIONAL CONGRESS OF LINGUISTS, 16., 1997, Paris. *Anais*...Oxford: Elsivier-Pergamon, 1998. paper 0425, CD-ROM 16.

FERNANDES, F. *Dicionário de sinônimos e antônimos da língua portuguesa*. São Paulo: Globo, 1997.

FERREIRA, A. B. H. *Dicionário Aurélio eletrônico século XXI*. Versão 3.0. São Paulo: Lexikon Informática, 1999. 1 CD-ROM.

FLEXNER, S.B. (Ed.) *Random house Webster's unabridged electronic dictionary*. Version 2.0. New York: Random House, 1997. 1 CD-ROM.

ILARI, R.; GERALDI, J. W. *Semântica*. São Paulo: Ática, 1985.

KILGARRIFF, A. Dictionary word sense distinctions: an enquiry into their nature. *Computers and the Humanities*, Amsterdam, v. 26, p. 365-387, 1993.

_____. I don't believe in word senses. *Computers and the Humanities*, Amsterdam, v. 31, p. 91-113, 1997.

LYONS, J. *Introdução à Lingüística teórica*. Tradução de Rosa Virgínia Mattos e Silva e Hélio Pimentel. São Paulo: Ed. Nacional, Ed. da Universidade de São Paulo, 1979.

MILLER, G. A.; FELLBAUM, C. Semantic networks of English. *Cognition*, Amsterdam, v.41, n.1-3, p.197-229, 1991.

SAINT-DIZIER, P.; VIEGAS, E. *Computational lexical semantics*. Cambridge: Cambridge University Press, 1995.

SUMMERS, D. (Ed.) *Longman dictionary of contemporary English*. Essex: Longman, 1995.

WEISZFLOG, W. (Ed.) *Michaelis português*: moderno dicionário da língua portuguesa. Versão 1.0. São Paulo: DTS Software Brasil, 1998. 1 CD-ROM.