A REDE *WORDNET* E A COMPILAÇÃO DE UM THESAURUS ELETRÔNICO

Bento Carlos Dias-da-Silva[1]
UNESP

Resumo

Este artigo discute a tarefa de compilação de um thesaurus eletrônico para o português do Brasil restrita a limitados recursos humanos, tecnológicos e lexicais. Depois de pontuar as principais etapas de desenvolvimento do projeto, apresenta o arcabouço teórico-metodológico adotado e aos principais problemas enfrentados no processo de montagem do modelo de representação lingüístico-computacional. Na seqüência, focaliza a caracterização do polissêmico termo thesaurus. *Por fim, com base na rede* WordNet, *apresenta a implementação do modelo: a programação do editor do thesaurus, isto é, uma ferramenta de autoria, criada para auxiliar o lingüista no processo de montagem da base do thesaurus. Para concluir, arrola os principais resultados e alguns desdobramentos futuros.*

Palavras-chave

estrutura do léxico; relações de sentido; thesaurus eletrônico; wordnet; tecnologia da linguagem humana.

Abstract

This paper focuses on the task of compiling a computerized Brazilian Portuguese thesaurus resorting to limited human, technological and lexical resourses. After presenting an overview of the ongoing project, it takes up the core theoretical and methodological issues, where it is sketched the hard problems the team had to face to devise a particular computational linguistic model. Next, it limits the polysemous term thesaurus, *for different specialists have used it to refer to very different objetcs. Finally, it addresses implementation issues, which amounts to describing a* WordNet*-based thesaurus editor, an authoring tool designed to help linguists feed the thesaurus database with the appropriate lexical information. The paper concludes with some results and an overview of the next stages of the work.*

Keywords

synonymy; lexicon; wordnet; computerized thesaurus; human language technology.

1 Programa de Pós-Graduação em Lingüística e Língua Portuguesa - Faculdade de Ciências e Letras - Universidade Estadual Paulista (UNESP) - Campus de Araraquara - CEP 14800-901, C.P. 174, Araraquara, SP, Brasil. bento@fclar.unesp.br

Apresentação

> *It is the fate of those who dwell at the lower employments of life, to be rather driven by the fear of evil, than attracted by the prospect of good; to be exposed to censure, without hope of praise; to be disgraced by miscarriage, or punished for neglect, where success would have been without applause, and diligence without reward. Among these unhappy mortals is the writer of dictionaries...*
>
> (Preface to the Dictionary, Samuel Johnson, 1775)

Dentro da temática *Dictionaries and computational approaches to the lexicon*,[2] este artigo apresenta uma parte do equacionamento do processo de compilação do *Thesaurus Eletrônico para o Português do Brasil* (TEP). Enquanto produto, deverá resultar em uma obra de referência em meio digital composta de um estoque de aproximadamente 40 mil palavras, organizado em termos das relações de sinonímia e antonímia.

Tomemos, como ponto de partida, as palavras do lexicógrafo inglês Samuel Johnson, que há muito já apresentava uma perspectiva pouco animadora para o trabalho do dicionarista. Conforme registra a epígrafe, a tarefa de compilação de dicionários é descrita como uma atividade complexa e laboriosa, fadada ao insucesso, alvo de críticas severas e, sobretudo, trabalho menor, se comparado à investigação lingüística.

A compilação da base do TEP (doravante BT), uma base de dados lexicais compostas de milhares de palavras organizadas em conjuntos de palavras sinônimas, colocou-nos diante de dificuldades adicionais, entre elas o tratamento lingüístico-computacional da sinonímia e antonímia, a manipulação de uma considerável massa de dados - milhares

2 Temática de uma das mesas-redondas do Congresso Internacional Sobre Polissemia e Indeterminação Semântica, realizado na Universidade Federal de Santa Catarina (UFSC), de 22 a 25 de agosto de 2001.

de palavras e expressões, distribuídas em milhares de conjuntos, interligados por meio de uma complexa rede de relações - e, sobretudo, a necessária, mas difícil, realização de pesquisa interdisciplinar, que exigiu o trabalho cooperativo e sincronizado entre lingüistas e cientistas da computação.

Em linhas gerais, o projeto, iniciado em janeiro de 1999 e concluído em outubro de 2001, estruturou-se segundo as seguintes etapas principais:

- Análise da forma e do conteúdo dos mais variados tipos de dicionários do português e inglês, publicados em papel ou em meio digital, com vistas à delimitação do objeto *thesaurus* e à montagem do *corpus de referência*, isto é, o conjunto de obras das quais foi extraída a informação léxico-semântica relevante para a compilação da BT;
- Estabelecimento dos critérios de filtragem para a compilação dos conjuntos de palavras sinônimas e dos conjuntos de palavras antônimas;
- Especificação da forma e do conteúdo da BT, isto é, o modelo subjacente de representação da informação léxico-semântica, e do sistema de armazenamento e gerenciamento dessa informação;
- Implementação computacional do editor para a compilação da BT;
- Inserção dos conjuntos na BT.[3]

Na seqüência, o artigo apresenta: (1) Questões metodológicas centrais; (2) A especificação do objeto *thesaurus*; (3) A delimitação do *corpus de referência*; (4) O modelo lingüístico-computacional de representação e armazenamento das relações de sinonímia e antonímia; (5) A arquitetura da BT; (6) O editor da BT; (7) Resultados e possíveis desdobramentos.

---

3 Este projeto é parte do Programa PADCT-III-CDT/ MCT, Finep-Itautec-Philco S.A [Processo RC: 3.1.3-0012/98 - Convênio: 8.8.98.059.00].

Questões de metodologia

Como qualquer projeto que tem por objeto o processamento computacional de entidades e processos lingüísticos, o trabalho de construção da BT exigiu planejamento cuidadoso, sobretudo porque envolveu (i) a montagem de equipe interdisciplinar, (ii) a resolução de problemas lingüísticos e computacionais e (iii) o equacionamento de etapas que levassem em consideração a especificidade de cada fase de desenvolvimento do trabalho (Dias-da-Silva, 1998b).

Aceitando o desafio de realizar a tarefa em pouco mais de dois anos, mesmo contando com restritos recursos lingüísticos e computacionais disponíveis para o português brasileiro, montamos, no NILC, Núcleo Interinstitucional de Lingüística Computacional, formado por docentes e alunos da USP-São Carlos, UNESP-Araraquara e UFSCAr-São Carlos, a equipe de desenvolviemento do projeto, composta de quatro lingüistas e um cientista da computação.

Uma vez montada a equipe, passamos para o equacionamento global das atividades. Aplicamos a metodologia desenvolvida por Dias-da-Silva (1998a), que prevê a fatoração das atividades em três fases solidárias, cíclicas e complementares: Fase Lingüística, Fase de Representação e Fase de Implementação. Na primeira, foram especificadas a concepção de thesaurus, a seleção e análise dos principais dicionários e obras de referência de língua portuguesa, a seleção dos critérios operacionais para a extração do conhecimento léxico-semântico contido implicitamente nessas obras (pois nelas estão incrustados meticuloso trabalho centenário e valiosas intuições de lexicógrafos) e a montagem das estratégias heurísticas para garantir a consistência interna de cada um dos verbetes e de cada um dos conjuntos neles contidos, bem com a consistência do inter-relacionamento entre os conjuntos, elementos constitutivos da BT. Na segunda fase, foram projetadas as diretrizes norteadoras da implementação da BT e do editor, isto é, a arquitetura conceitual. Nessa fase é que foram especificados a estrutura do verbete, o modelo lingüístico-computacional de representação das relações

de sinonímia e antonímia e o modo de armazenamento e gerenciamento dos dados da BT. Na última fase, foi implementado o editor e, por meio dele, foram inseridos todos os conjuntos na BT.

O termo *thesaurus*

Nosso ponto de partida foi, necessariamente, a delimitação do termo *thesaurus*, uma vez que a consulta às mais variadas obras de referência (periódicos, dicionários, enciclopédias, manuais de lexicografia, entre outras) mostrou que o termo *thesaurus*, ou seu outro equivalente em língua portuguesa *tesauro*, aplica-se a objetos bastante diferentes, não havendo, portanto, consistência no seu emprego. Não partimos para uma análise etimológica ou exaustivas pesquisas de *corpora* na busca de uma explicação para o fato. Buscamos, sim, delimitar o objeto que pretendíamos construir: um *thesaurus*, um *tesauro* ou um *tesouro*? Uma leitura cuidadosa das principais obras de referência, sobretudo as de língua inglesa, permitiu concluir que o termo *thesaurus* aplica-se, hoje, a, pelo menos, seis objetos distintos.

O emprego clássico do termo *thesaurus* refere-se ao **objeto 1**: um inventário, que pretende ser exaustivo, do vocabulário de uma determinada língua, ou de um determinado ramo do conhecimento, um *tesouro* vocabular (GELC, 1998; Ferreira, 1999). Um outro emprego, não menos clássico, denota o **objeto 2**: um dicionário organizado em função de conceitos lexicalizados, isto é, um dicionário onomasiológico (Weiszflog, 1998), um *tesauro* ou *tesouro*, ou ainda um *dicionário analógico*, cujo precursor foi o *Thesaurus de Roget*.[4] Ao utilizarmos os programas de busca na Internet para pesquisar o termo *thesaurus*, encontramos o emprego que atualiza o jargão corrente nos domínios da Infor-

4 O Thesaurus de Roget, publicado em 1852, levou 12 anos para ser concluído. O Dicionário analógico da língua portuguesa (Azevedo, 1983) pode ser considerado a versão do dicionário de Roget para o português brasileiro.

mática e Documentação e refere-se ao **objeto 3**:[5] "Vocabulário controlado e dinâmico de *descritores* [palavra ou expressão utilizada em indexação e tesauro para representar, sem ambigüidade, um determinado conceito], relacionados semântica e genericamente, que cobre de forma extensiva um ramo específico de conhecimento" (Ferreira, op. cit.). Nessa acepção, todas as três formas estão abonadas: *thesaurus*, *tesauro* e *tesouro*. O quarto emprego, também motivado pelo advento da Informática, refere-se ao **objeto 4**: "Arquivo contendo sinônimos que são exibidos como alternativas para uma palavra escrita de forma incorreta, durante uma verificação de ortografia" (Weiszflog, op. cit.). Já o emprego do termo *thesaurus* para fazer referência ao **objeto 5**, isto é, um dicionário de sinônimos e antônimos, aproxima-se da concepção que previamos para o TEP, este definido como o **objeto 6**: um dicionário de sinônimos e anônimos, armazenado na memória do computador, para ser utilizado no processamento manual ou automático de textos (Flexner, 1997; Neufeldt, 1997).

Duas considerações justificam a escolha do termo *thesaurus* para denominar o TEP. Por um lado, cinco dos seis objetos (exclui-se o objeto 1) possuem um traço comum: são dicionários particulares, estruturados segundo critérios específicos: relações conceptuais, relações léxico-semânticas, campos semântico-nocionais e sistema de termos. Além disso, utilizam, com diferentes graus de prioridade, duas relações léxico-semânticas gerais: a semelhança e a oposição de sentidos, que no limite são as relações de sinonímia e antonímia. O que os diferencia são o propósito, a funcionalidade e o meio de disseminação. Essas considerações são suficientes para descartarmos o termo *tesouro*. Por

5 Nesse sentido, o termo data da década de 50, quando H.P.Luhn, funcionário da da IBM, propôs um procedimento computacional capaz de gerar uma lista de termos técnicos para indexar textos científicos (Cf. http://www.gslis.utexas.edu/~ssoy/organizing/l391d2c.htm - Acesso ago de 2001).

outro lado, o termo *thesaurus*, considerado variante do termo *tesauro*, para denotar o objeto 6, está abonado em Weiszflog (1998), o que corrobora nossa escolha, que ganha suporte adicional em Ferreira (1999), que restringe a aplicação do termo *tesauro* ao objeto 3.

Constituição do *corpus de referência*

Feitos os estudos preliminares, levantamos os principais problemas de natureza conjuntural e teórico-metodológica:

- a formação de equipe interdisciplinar, diante da carência de especialistas voltados para a lexicografia computacional e a semântica lexical computacional (Saint-Dizier e Viegas, 1995).
- a seleção dos dicionários para a extração do conhecimento lexical, diante da inexistência de obras do português do Brasil computacionalmente tratáveis;
- o estabelecimento de critérios de filtragem para minimizar as inconsistências, lacunas e imprecisões detectadas nos dicionários;
- a seleção das entradas e a especificação formal, e computacionalmente tratável, das relações de sinonímia e antonímia;
- a construção de um editor eficiente e amigável, para agilizar a inserção (semi-) automática dos dados, e equipado com recursos para que a consistência do grande volume de dados e da intricada rede de relações que se estabelecem entre eles pudesse ser testada;

Sem a pretensão de apontar solução para todos esses problemas, selecionamos, a seguir, algumas das dificuldades que nos motivaram a utilizar, como fontes de conhecimento lexical, um conjunto de obras que, apesar de suas limitações, são consideradas referências para muitas questões lexicográficas do português do Brasil e mostraram-se adequadas para os propósitos do projeto.

Embora as dificuldades postas pelo tratamento computacional de fenômenos da linguagem sejam consideráveis, as maiores estão, sem dúvida, na própria descrição e análise do léxico (Cruse, 1986). Como já ad-

vertia Lyons (1976), investigar o léxico é enfrentar questões de natureza fonético-fonológica, passando pelas questões morfossintáticas, culminando com as complicadíssimas questões semânticas e pragmáticas.

Na tarefa de compilar a BT, uma questão crucial foi a delimitação das unidades lexicais que deveriam figurar como entradas, problema para o qual ainda não dispomos de uma solução satisfatória. Termos como *palavra*, *vocábulo*, *lexema*, *lexia*, entre outros, surgiram na tentativa de se delimitar essas unidades. Não é fato novo afirmarmos que a noção de palavra seja de difícil especificação formal. Evidências, entretanto, apontam para sua realidade psíquica: indivíduos de sociedades ágrafas são capazes de ditar um texto palavra por palavra e a fala holofrástica da criança, por exemplo. Teóricos apontam a questão da delimitação.[6] Nesse processo, dizem, surge uma dúvida: a seqüência de formas sob análise constitui uma unidade do léxico ou é uma combinatória de lexias, produto da sintaxe? Parte dessa dúvida é alimentada pela tradição gramatical que, fixando-se na forma ortográfica do lexema, que nem sempre revela seu estado de lexicalização, considera lexias como *dor de cabeça*, *à toa* e *contanto que*, por exemplo, como locuções. Não estamos diante de construções sintáticas, porque os elementos componentes encontram-se há muito tempo soldados. Não se pode, por exemplo, dizer, no sentido de "enxaqueca", **dor persistente de cabeça*, **dor da cabeça*, **dor das cabeças*, fato que permite reconhecer que essas unidades sofreram um processo de lexicalização e não são mais resultantes de operações sintáticas, o que garante a essas e outras tantas lexias figurem como entradas.

6 O termo lexia é, aqui, empregado no sentido de forma atualizada de um lexema. Este, por sua vez, refere-se à unidade básica e abstrata do sistema léxico de uma língua. Dubois et al. (1978) ajudam-nos oportunamente a esclarecer uma outra oposição de termos, encontrada na linguagem comum, e que se prestam a equívocos: vocábulo/palavra: o primeiro refere-se a um tipo de unidade da fala (type) e o segundo, às diferentes ocorrências (tokens) desse tipo de unidade.

Outra questão bem conhecida do lexicógrafo é a variedade de tipos de significado, cuja discriminação coloca para o lingüista inúmeras dificuldades. Pelo menos sete tipos já foram alvos de estudos: conceptual (*sentido*), conotativo, estilístico, afetivo, refletido (*reflected*), de colocação e temático.

Por fim, há os problemas inerentes ao trabalho prático do dicionarista diante da imensidão do léxico. Considerando-se que o léxico é um sistema aberto e em expansão, sua descrição exaustiva é tarefa impossível. Qualquer dicionário será necessariamente incompleto, refletindo os inevitáveis recortes feitos pelo dicionarista para registrar um determinado estado de língua. Conforme alertam Dubois et al. (1978), a esse se somam os problemas gerados (a) pela hesitação entre a impossível exaustividade e os limites materiais e práticos; (b) pelo viés na seleção do conteúdo, que, em geral, varia segundo a decisão de cada lexicógrafo, ao privilegiar o registro de determinados empregos técnicos e metafóricos, por exemplo, em detrimento de outros; (c) pela dificuldade de distinguir entre o vocabulário geral e o de língua especial, uma vez que esses matizes são de difícil observação; (d) pela não transparência dos critérios que sancionam a passagem de um neologismo para o léxico consagrado da língua.

Diante do exposto, e de trabalhos de lexicografia computacional, a reutilização de recursos disponíveis foi a saída que encontramos para montar o *corpus de referência* (Briscoe e Boguraev, 1989). Por um lado, essa estratégia possibilitou a agilização dos trabalhos, ao reduzir grande parte das atividades à extração e filtragem dos conhecimentos lexicográfico e lingüístico, direta ou indiretamente, contidos nas obras de referência analisadas (Moraes e Dias-da-Silva, 2000). Por outro lado, e decorrência dessa estratégia, a adoção do critério de abonação, além de ter minimizado a necessidade de investigação custosa e pontual de cada verbete a ser construído, minimizou também a necessidade das laboriosas pesquisas em *corpora* e garantiu a conformidade da BT aos padrões de expressão da norma escrita, característica que deverá ser contemplada, e até mesmo desejada, no TEP.

Das obras analisadas, selecionamos sete: Nascentes (1981), Azevedo (1983), Borba (1990), Weiszflog (1998), Fernandes (1997), Ferreira (1999), Barbosa (1999). Apesar de serem alvos constantes de críticas,[7] três observações justificam a utilização dessas obras como *corpus de referência*. Primeiro, essas obras são fontes, em estado bruto, de conhecimento léxico-semântico e, parte delas, possui uma tradição centenária. Segundo, elas utilzam em profusão, na especificação das acepções das entradas, o emprego de palavras sinônimas e antônimas. Terceiro, grande parte delas está disponível em meio digital.

Questões de representação: os *synsets*

Uma questão que se coloca freqüentemente nas discussões sobre a sinonímia é o fato das línguas naturais não apresentarem sinônimos perfeitos. Como conseqüência, observa Lutz (1994), um thesaurus limitado a registrar sinônimos exatos estaria reduzido a uma lista de poucas palavras e seria, portanto, de pouca utilidade. O usuário de um thesaurus, entretanto, não busca uma correspondência precisa para efetuar a substituição pretendida. Ele já dispõe de uma palavra que pode ser usada naquele contexto, mas, por razões de estilo, precisão, correção ou, até mesmo, aprendizagem, conforme assinalam Ilari e Geraldi (1985), julga necessário fazê-la, por necessidade de encontrar termos alternativos e mais eficientes para a expressão do pensamento. Logo, o usuário busca palavras de sentidos semelhantes. É, portanto, nesse sentido que o termo sinônimo é empregado nos diversos tipos de obras de referência arroladas anteriormente, prática que também adotamos neste projeto.

Um thesaurus eletrônico, além de conformar-se a essa expectativa do usuário, deve também viabilizar acesso imediato aos milhares de palavras sinônimas e antônimas, sistematicamente rela-

7 Cabe aqui lembrar as críticas severas que Cláudio Abramo lançou contra a nova versão do dicionário Aurélio. (Caderno Mais, da Folha de São Paulo, 23/01/2000).

cionados às milhares de entradas. Essa funcionalidade adicional cria, para o lingüista computacional, a necessidade de buscar formas de representação dessas relações de sentido que sejam computacionalmente tratáveis.

Buscamos solução para essas questões na metodologia empregada no desenvolvimento da rede *WordNet*, desenvolvida em Princeton (Miller e Fellbaum, 1991; Fellbaum, 1998). Dela, utilizamos três noções fundamentais: (i) o "método diferencial", que pressupõe o princípio de ativação de conceitos por meio de um conjunto de formas lexicais relacionadas pela relação de sinonímia; o que eleimina a necessidade de especificação de um valor semântico (um rótulo conceitual) para cada acepção da entrada, (ii) a noção constitutiva básica de *synset* (conjunto de palavras sinônimas) e (iii) a noção de "matriz lexical", que postula a existência de uma correspondência biunívoca entre sentido e *synset*. Há, porém, diferenças. Os objetivos dos dois aplicativos são bastante distintos. Como explicitamos anteriormente, o TEP deverá ser um objeto do tipo 6; já a rede *WordNet* é um modelo computacional construído para simular a estrutura dos conceitos lexicalizados na língua inglesa, a simulação de um léxico mental (Dias-da-Silva e Oliveira, 2001).

Devido à adoção desse modelo de representação, a tarefa de construção da BT ficou reduzida à compilação dos conjuntos de palavras sinônimas e das relações entre conjuntos de palavras antônimas. O seguinte esquema ilustra a estrutura de um verbete-tipo:

Entrada n (categoria sintática 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.

A base do thesaurus

Do ponto de vista lógico, a BT configura-se conforme ilustra a Figura 1: uma estrutura de dados composta por duas listas: *Lista de Entradas* (LE), as entradas da BT, ordenadas alfabeticamente, e *Lista de Conjuntos* (LC), os *synsets*. Cada elemento do *synset* é necessariamente um elemento da LE. Cada entrada da LE, além de conter a representação ortográfica do lexema, contém uma *Especificação de Acepções* (EA). Cada EA é realizada por três indexações: a primeira aponta para um determinado conjunto de sinônimos, a segunda aponta para um determinado conjunto de antônimos, ambos pertencentes à LC, e a terceira aponta para uma determinada entrada da LE, sinalizando que esses conjuntos fazem parte dessa acepção dessa entrada.

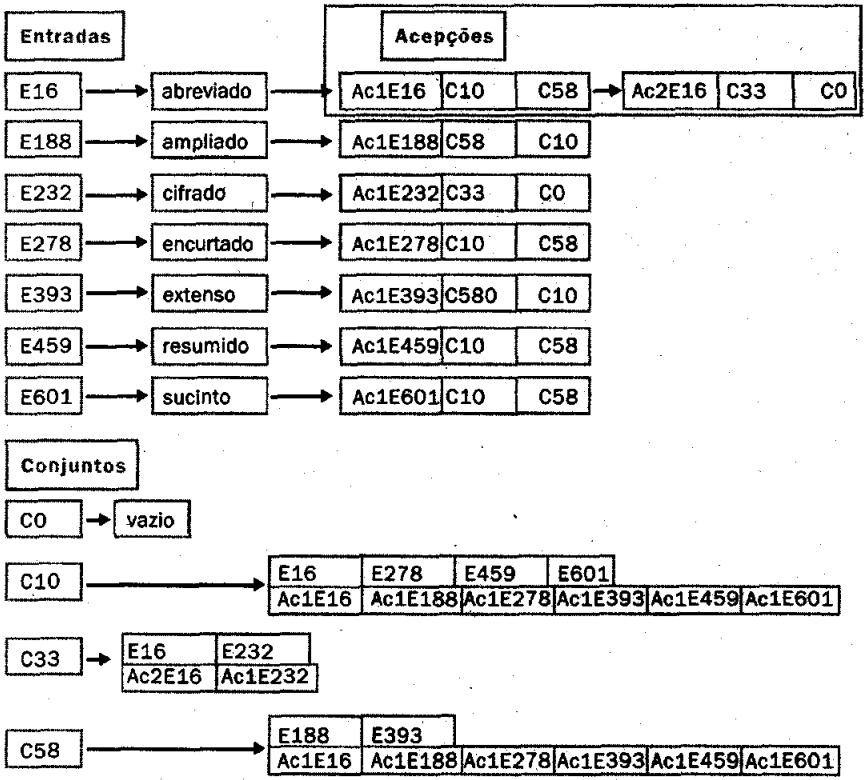

FIGURA 1 Exemplo de armazenamento da entrada **abreviado** na BT. Os índices Exx, Cxx e AcxxEyy são apenas para facilitar a ilustração, pois a estrutura interna utiliza ponteiros (endereços de memória) para indexar esses campos.

Assim, cada conjunto da LC é formado por uma LE e uma LA. A lista de entradas contém entradas que estão relacionadas entre si pela sinonímia e pode pertencer a várias acepções. Por poder pertencer a várias acepções, o conjunto formado pelas entradas exige uma lista de acepções para agilizar a pesquisa de entradas na detecção do tipo de relação, sinonímia ou antonímia, que cada conjunto contrai com as entradas que o contém, e no gerenciamento desse conjunto pelo editor.

O editor

O editor é uma interface computacional gráfica para a montagem da BT. Sua implementação foi possível graças ao modelo de representação formal acima descrito, uma vez que, no contexto do modelo, as relações de sinonímia e antonímia passam a ter uma existência computacional: a sinonímia é especificada pela relação de pertença, que se estabelece entre formas da língua e o *synset* que as contém; a antonímia é especificada pela relação rotulada entre pares de *synsets*.

Com as funções básicas de agilizar a entrada de dados, armazenar e gerenciar os conjuntos, permitindo a criação e manipulação de cada verbete pelo lingüista, o editor oferece ferramentas que possibilitam ao lingüista editar e gerenciar vários tipos de informação durante o processo de montagem da base: salvar dados, desfazer operações, editar campos, visualizar o conteúdo da base, listar entradas e verbetes, imprimir partes da base e extrair dados estatísticos (número de entradas e conjuntos, proporção número de entradas e número de conjuntos inseridos, número de entradas e verbetes gerados automaticamente).

Além de apresentar uma barra de ferramentas e uma barra de menus, o editor apresenta três quadros redimensionáveis: dois superiores e um inferior. O quadro superior à esquerda apresenta, em camadas, quatro tipos de listas: Todas as Entradas, Entradas Automáticas, Atualização Automática e Novas Entradas, todas em ordem alfabética, para facilitar a edição. O quadro superior à direita exibe o verbete completo, estruturado em forma de árvore, referente à entrada selecionada no quadro à esquerda. Excetuando-se o item Acepção, que aparece na estrutura do verbete, todos os demais itens desse quadro podem ser renomeados. Além disso, um duplo clique no item raiz, ou nos itens Acepção x, a árvore expande-se ou contrai-se; um duplo clique nos itens terminais possibilita a navegação pela base: ao clicarmos, por exemplo, a palavra **extenso** (antônimo de **abreviado**), o editor nos remete para o verbete correspondente. Finalmente, o quadro inferior exibe todos os conjuntos que contêm pelo menos uma ocorrência da entrada selecionada.

O processo de inserção das expressões lexicais na base pode ser acompanhado através de um exemplo ilustrativo. Após extrair e filtrar as informações do *corpus de referência*, a tarefa do lingüista é inserir as expressões selecionadas nos campos apropriados do Assistente do editor, que pode ser acionado a partir de um botão na barra de ferramentas. O resultado desse procedimento é ilustrado na Figura 2.

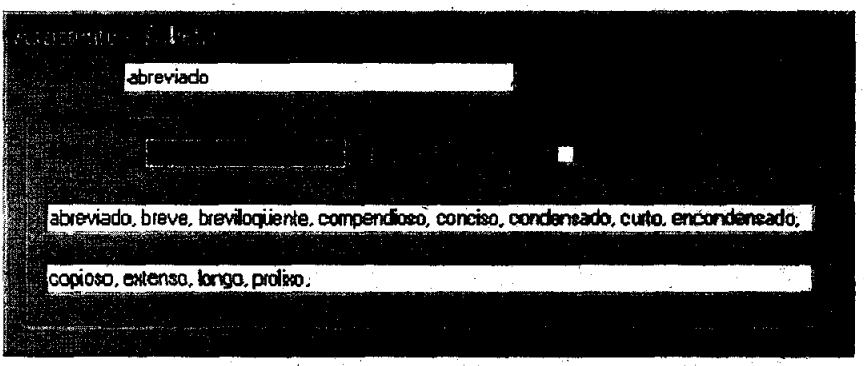

FIGURA 2 Estado do Assistente do editor durante a montagem do verbete do adjetivo **abreviado.**

Ao detectar que a palavra **abreviado** foi inserida no campo Entrada, o editor verifica se essa entrada já existe na base. Em caso afirmativo, os campos Palavra ou Lista de Sinônimos e/ou Palavra ou Lista de Antônimos do Assistente são automaticamente preenchidos com as expressões lexicais recuperadas da base, permitindo que esses campos sejam editados. Em caso negativo, o editor cria novos campos. Ao terminar de preencher os campos, o lingüista confirma a entrada dos dados, clicando o botão Inserir, o que completa a montagem do verbete. Mecanismos de gerenciamento e manutenção de dados estão incorporados ao editor para verificar a consistência dos dados, disparando o algoritmo de geração automática de verbetes e de atualização da estrutura interna da base. Como resultado dessas operações, o editor apresenta as informações ilustradas na Figura 3.

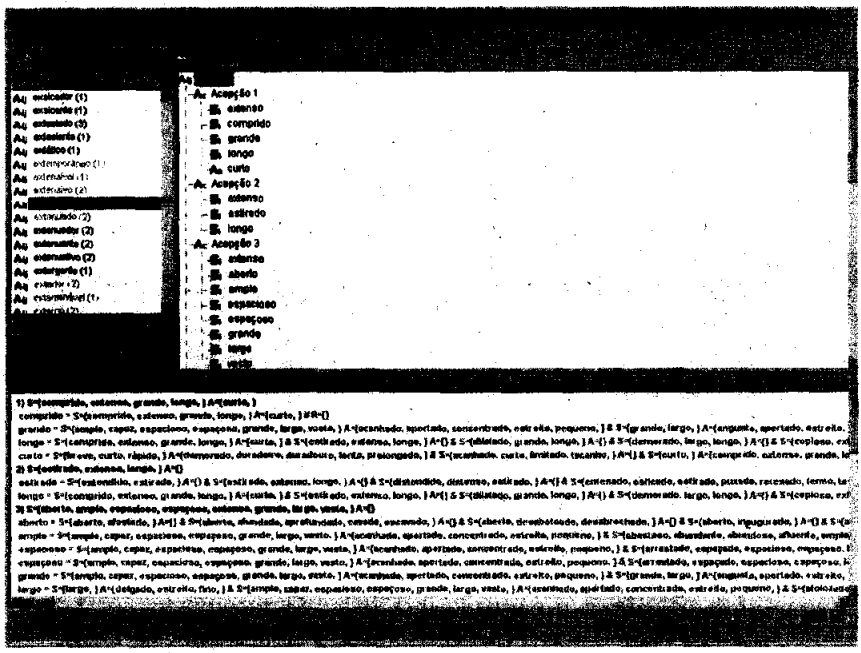

FIGURA 3 Estado do editor da BT após seleção da entrada **extenso**, um dos antônimos de **abreviado**. A janela inferior exibe três dos cinco conjuntos de sinônimos construídos para essa entrada.

Resultados e perspectivas

Do ponto de vista da pesquisa interdisciplinar, envolvendo lingüistas e cientistas da computação, a metodologia proposta para a compilação da base do thesaurus demonstrou ser eficiente, além de ter permitido que cada membro da equipe, sem perder a visão global do projeto, concentrasse na solução dos problemas específicos de seu domínio. Por exemplo, a especificação das relações de sentido relevantes para o thesaurus, o estabelecimento de critérios para a seleção e filtragem da informação léxico-semântica extraída do *corpus de referência* e a montagem dos conjuntos de sinônimos e antônimos foram as principais atividades desenvolvidas na fase que

denominamos lingüística. Na fase de representação, as discussões concentraram-se nas especificações da forma e funcionalidade do sistema de representação dos verbetes e na proposição de estratégias de codificação das relações e dos conjuntos delimitados no domínio anterior. Essas atividades possibilitaram um rico intercâmbio de informações e conhecimentos entre os especialistas das duas áreas envolvidas. Finalmente, na fase de implementação, além do trabalho cooperativo ter apontado soluções para as questões referentes à montagem de um ambiente computacional adequado para a montagem global do sistema em que a base do thesaurus deveria ser armazenada e manipulada, os especialistas da computação tiveram a oportunidade de apreciar, com maior profundidade, os resistentes problemas postos pela linguagem humana, que, como sabemos, vem resistindo às incontáveis tentativas de ser reduzida a um simples código de máquina.

Do ponto de vista da exclusivamente lingüístico, o empreendimento contribuiu para aguçar o senso de análise de fenômenos lingüísticos ligados à semântica lexical e à lexicalização de conceitos (Levin e Pinker, 1991).

Já do ponto de vista da tecnologia da linguagem humana, nossos estudos comprovaram a viabilidade de reutilização de fontes de informação lexical na etapa de extração do conhecimento léxico-semântico implícita ou explicitamente contido nos mais variados tipos de dicionários de língua porruguesa, parte do processo de construção de bases de dados lexicais robustas.

Em termos quantitativos, podemos afirmar que os esforços da equipe foram recompensados com os resultados. A base do TEP conta hoje com mais de 18 mil conjuntos, responsáveis pela geração de mais de 44 mil entradas, conforme ilustra a Tabela 1.

TABELA 1 Estatísticas aproximadas da base do thesaurus

CATEGORIA	Nº DE ENTRADAS	Nº DE CONJUNTOS	ENTRADAS/CONJUNTOS
Verbo	11.000	4.000	2,75
Substantivo	17.000	8.000	2,13
Adjetivo	15.000	6.000	2,50
Advérbio	1.000	500	2,64
TOTAL	**44.000**	**18.500**	**2,38**

Alguns desdobramentos futuros podem ser avançados: a aplicação de testes de consistência dos dados lexicais inseridos na base e verificação da completude da base em relação ao *corpus de referência* e a um "léxico de controle" previamente compilado; a conversão da base do thesaurus no aplicativo Thesaurus Eletrônico; a análise de questões referentes à apresentação, modos de consulta e disponibilização do aplicativo, bem como de possíveis modos de integração a outras ferramentas. Além de um possível refinamento dos conjuntos de sinônimos e da ampliação da informação da base com a adição de abonações (frases-tipo, extraídas de concordâncias), é possível também o acréscimo de indexadores conceituais, isto é, categorias ontológicas *ontos* propostas por Nirenburg (1992), por exemplo, para cada conjunto de sinônimos.

Observamos que há a possibilidade de utilização do *corpus*, com aproximadamente 30 milhões de palavras,[8] e o léxico do NILC. O primeiro, enquanto fonte de informação lexical, deverá servir de parâmetro para a inclusão ou não de determinadas entradas; o segundo deverá viabilizar a realização de testes de compatibilidade entre ambas as bases de dados lexicais, a BT e o léxico do NILC, o que deverá contribuir para o refinamento de ambos.[9]

8 A porção corrigida do corpus do NILC está acessível para consulta, através do IMS corpus query tools, da Universidade de Stuttgart: http://cgi.portugues.mct.pt/acesso/.

9 No léxico do NILC, a soma de todos os lexemas relevantes para o TEP (substantivos, verbos, adjetivos e advérbios) é 55.478.

Referências bibliográficas

AZEVEDO, F.F.S. *Dicionário analógico da língua portuguesa.* Brasília: Thesaurus, 1983.

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: Fundação Editora 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. Estrutura do léxico: modelo lingüístico-computacional de representação das relações semânticas. *Estudos Lingüísticos*, v.30, 2001. 1 CD.

DIAS-DA-SILVA, B.C. Bridging the gap between linguistic theory and natural language processing. In: B.Caron (org) *Proceedings of the 16th International Congress of Linguists.* Oxford: Elsevier, 1998a. 1 CD, Paper 0425.

_____ Os domínios lingüístico e tecnológico do estudo do processamento automático das línguas naturais. *Estudos Lingüísticos*, v.26, p.612-617, 1998b.

DUBOIS, J. et al. *Dicionário de lingüística.* Trad. Izidoro Blikstein. São Paulo: Cultrix, 1978.

FELLBAUM, C. (ed) *WordNet: an electronic lexical database.* Cambridge, Mass.: The MIT Press, 1998.

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 Ltda., 1999.

FLEXNER, S.B. (ed) *Random house Webster's unabridged electronic dictionary* (Version 2.0). New York: Random House Inc, 1997.

GELC (1998). *Grande Enciclopédia Larousse Cultural.* Nova Cultural, São Paulo.

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

LEVIN, B.; PINKER, S. (eds) Lexical and conceptual semantics. *Cognition*, v.41, n.1-3, p.1-229, 1991.

LUTZ, W.D. *The Cambridge thesaurus of American English.* Cambridge: Cambridge University Press, 1994.

LYONS, J. *Introdução à lingüística geral.* Supervisão de tradução de Isaac Nicolau Salum. São Paulo: Editora Nacional/Edusp, 1979.

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

MORAES, H.R.; DIAS-DA-SILVA, B.C. A questão da representação lingüístico-computacional da sinonímia e antonímia na compilação de um thesaurus eletrônico. *Revista de Iniciação Científica*, v.2, p.414-423, 2000. 10p.

NASCENTES, A. *Dicionário de sinônimos.* São Paulo: Nova Fronteira, 1981.

NEUFELDT, V. (ed) *Webster's New World Dictionary & Thesaurus* (Version 1.0). New York: Macmillan Publishers, 1997.

NIRENBURG, S. *Machine Translation.* San Mateo: Morgan Kaufmann Publishers, 1992.

ROGET, P.M. *Thesaurus.* Middlessex: Penguin Books, 1953. (ed. original, 1852)

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

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