Ref. DIAS-DA-SILVA, B.C.; MORAES, Helio Roberto de; OLIVEIRA, Mirna Fernanda de; HASEGAWA, Ricardo; AMORIM, D. A. ; PACHOALINO, C. Construção de um thesaurus eletrônico para o português do Brasil. **Processamento Computacional do Português Escrito e Falado**, ATIBAIA, v. 4, p. 1-10, 2000

Construção de um Thesaurus Eletrônico para o Português do Brasil

Bento Carlos DIAS-DA-SILVA
Nilc, Unesp
CEP 14800-901, C.P. 174, Araraquara, Brasil
bento@fclar.unesp.br

Mirna Fernanda de OLIVEIRA
Nilc, Unesp
CEP 14800-901, C.P. 174, Araraquara, Brasil
m.oliveira@techs.com.br

Ricardo HASEGAWA
Nilc, Usp
CEP 13560-970, C.P. 668, São Carlos, Brasil
rh@nilc.icmsc.sc.usp.br

Helio Roberto de MORAES
Nilc, Unesp
CEP 14800-901, C.P. 174, Araraquara, Brasil
helio@sartre.fclar.unesp.br

Daniela AMORIM
Nilc, Unesp
CEP 14800-901, C.P. 174, Araraquara, Brasil
dddamorim@uol.com.br

Christie PASCHOALINO
Nilc, Unesp
CEP 14800-901, C.P. 174, Araraquara, Brasil
christie_gp@hotmail.com

Ana Cláudia NASCIMENTO
Nilc, Usp
CEP 13560-970, C.P. 668, São Carlos, Brasil
ana@nilc.icmsc.sc.usp.br

Abstract

This paper examines a number of issues involved in the task of compiling a *Core Brazilian Portuguese Thesaurus* (*TeP*), resorting to limited human, technological and lexical resourses. After presenting an overview of the ongoing project, we take up the core theoretical and methodological issues, where we sketch the hard problems we had to face to devise a particular computational linguistic model. Next, we focus on the characterization of what is understood by the term *thesaurus* itself, for it has been used by different specialists to refer to very different objetcs. Finally, we address implementation issues, which amounts to describing the thesaurus editor, a specific authoring tool designed to help linguists feed the thesaurus database with the appropriate lexical information. The paper concludes with some expressive results and a brief overview of the next stages of the work.

Resumo

Este artigo aborda uma série de questões envolvidas na tarefa de compilação de um *Thesaurus Eletrônico Básico para o Português do Brasil* (*TeP*), restrita a limitados recursos humanos, tecnológicos e lexicais. Depois de delinearmos, na introdução, um breve panorama do projeto, apresentamos o equacionamento global do empreendimento, esboçando o arcabouço teórico-metodológico adotado e arrolando os principais problemas que tivemos de enfrentar no processo de montagem de um modelo de representação lingüístico-computacional adequado e eficiente para a realização da tarefa. Na seqüência, focalizamos a caracterização do termo *thesaurus*, posto que esse termo tem sido empregado por diferentes especialistas para denotar objetos bastante diversos. Por fim, apresentamos uma solução para a questão de implementação do modelo, descrevendo o editor do thesaurus, uma ferramenta de autoria específica, projetada para auxiliar o lingüista no processo de montagem da base do thesaurus. Na conclusão, enumeramos os resultados alcançados até o estágio atual de desenvolvimento do projeto e pontuamos as etapas subseqüentes.

1. Introdução

> *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)[1]

Este trabalho apresenta uma parcela do equacionamento do processo de compilação de um *Thesaurus Eletrônico Básico para o Português do Brasil* (*TeP*),[2] que, acoplado a outras ferramentas computacionais de auxílio à expressão escrita, deverá somar-se a outras obras de referência em meio digital como dicionários e gramáticas. Essa ferramenta deverá oferecer ao usuário da língua portuguesa a oportunidade ímpar de escolher palavras sinônimas e antônimas que ele, por motivos de estilo, de precisão, de correção ou de aprendizagem, deseja substituir.

Além das dificuldades apontadas por Samuel Johnson - uma atividade complexa e extremamente laboriosa, fadada ao insucesso, alvo de críticas severas e considerada trabalho menor, se comparado à investigação lingüística -, a construção de uma ferramenta dessa natureza colocou-nos diante de dificuldades adicionais, pois a extração do conhecimento lingüístico que lhe dá sustentação pressupõe a análise de um dos mais complexos fenômenos lingüístico-conceituais: a sinonímia e a antonímia. Além disso, por envolver o tratamento computacional de uma considerável massa de dados - milhares de palavras e expressões, distribuídas em milhares de conjuntos e projetando uma complexa rede de relações -, apresenta uma outra complexidade, pois se reveste de um caráter necessariamente interdisciplinar, exigindo o difícil trabalho cooperativo e sincronizado entre lingüistas e cientistas da computação.

Ciente dessas dificuldades, a equipe composta de seis lingüistas e um cientista da computação, aceitando o desafio de realizar a tarefa em dois anos e contando apenas com os restritos recursos lingüísticos e computacionais disponíveis para o português brasileiro, procurou equacionar o problema com doses de criatividade e muita ousadia.

Para que o leitor se situe no contexto deste projeto, é importante ressaltar que o processo global de seu desenvolvimento foi decomposto em oito etapas: (1) análise da forma e do conteúdo de obras de referência disponíveis (os mais variados tipos de dicionários do português e inglês, publicados em papel ou disponíveis em meio digital), com vistas à delimitação do objeto *thesaurus* e, sobretudo, à utilização dessas obras como fontes de conhecimento lexical; (2) seleção das obras de referência, agora enquanto fontes legítimas de conhecimento lexical, e estabelecimento de critérios de filtragem da informação lexical delas extraída; (3) especificação do conteúdo e da forma da base do thesaurus; (4) implementação de um editor para a construção dessa base; (5) inserção dos dados na base do thesaurus por lingüistas; (6) aplicação de testes de consistência global da base e de sua completude relativa às fontes de conhecimento lexical selecionadas e ao léxico do ReGra (Nunes *et al*., 1996); (7) conversão da base do thesaurus no *TeP*; (8) análise de questões referentes à apresentação e disseminação do *TeP*, bem como ao seu modo de integração a outros aplicativos.

Nas seções subseqüentes, após a apresentação da metodologia adotada, a delimitação do termo *thesaurus* e a especificação das fontes de conhecimento lexical que selecionamos (Seção 2), esboçamos o modelo lingüístico-computacional de representação adotado (Seção 3) e descrevemos o editor que projetamos para construir a base do thesaurus (Seção 4). Na conclusão, sumarizamos os principais resultados alcançados e pontuamos as tarefas que deverão ser objeto de trabalho futuro.

---

[1] Johnson (*apud* Levin & Pinker, 1991).

[2] Este projeto é parte do projeto maior *Revisor Gramatical e Ferramentas de Auxílio à Escrita* - Programa PADCT-III-CDT/ MCT, Finep-Itautec-Philco S.A [Processo RC: 3.1.3-0012/98 - Convênio: 8.8.98.059.00].

2. Questões de metodologia:

2.1. Equacionamento global do projeto

Como em qualquer projeto cujo objeto é o processamento computacional de entidades e processos lingüísticos, o complexo trabalho de construção de um thesaurus eletrônico, com as características mencionadas, exigiu cuidadoso planejamento, que envolveu (i) a montagem de uma equipe interdisciplinar de profissionais, composta de cientistas da linguagem e da computação, (ii) o equacionamento das principais etapas de desenvolvimento desse trabalho interdisciplinar e (iii) a resolução dos problemas lingüísticos e computacionais específicos ao empreendimento. Dias-da-Silva (1998a,b), transpondo para o âmbito dos estudos do processamento automático das línguas naturais o processo de construção de um sistema de conhecimento, propõe uma metodologia de trabalho que, aplicada ao desenvolvimento do thesaurus eletrônico, permitiu que os problemas fossem equacionados de forma modular:

- Fase Lingüística – Nesta fase, delimitamos o objeto de investigação, selecionamos e analisamos os principais dicionários e obras de referência similares de língua portuguesa, estabelecemos os critérios operacionais para a extração do conhecimento léxico-semântico contido implicitamente nessas obras (Briscoe & Boguraev, 1989; Cruse 1986; Kilgarriff, 1993; Landau, 1996) e montamos as estratégias heurísticas para garantir a consistência dos dados que formam a base do thesaurus (Hartmann, 1983; Saint-Dizier & Viegas, 1995);
- Fase Representacional – Nesta fase, projetamos a arquitetura conceitual que serve de planta para a implementação do thesaurus, tarefa que consistiu em determinar a estrutura dos verbetes e a estrutura global do aplicativo, levando-se em conta as funcionalidades previstas para a ferramenta (Miller & Fellbaum, 1991; McCreary, 1996);
- Fase de Implementação – Nesta fase, implementamos o editor e, por meio dele, estamos alimentado a base do thesaurus. Compete ainda a esta fase não só a implementação do produto final e a especificação do modo de sua integração a outros aplicativos, mas principalmente a realização de testes de verificação da precisão e consistência dos dados incorporados ao thesaurus, a implementação de suas funcionalidades e a escolha de modos de apresentação da ferramenta.

2.2. Os termos *thesaurus, tesauro* e *tesouro*

O ponto de partida deste projeto foi a delimitação do objeto de investigação, uma vez que a consulta às mais variadas obras de referência (periódicos, dicionários, enciclopédias, manuais de lexicografia, entre outras) revelou 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 se trata, porém, de mergulharmos, aqui, em uma análise etimológica, ou em exaustivas pesquisas de *corpora*, para buscar esclarecimento dos fatos. O breve estudo, a seguir, além de nos auxiliar na delimitação de nosso objeto, ilustra também os problemas adicionais que os dicionários, no confronto das informações neles apresentadas, colocam para o consulente e para nós pesquisadores, que deles fazemos os mais diversos usos. Além disso, especificado o objeto, como denominá-lo: *thesaurus*, *tesauro* ou *tesouro*?

Uma leitura, nas entrelinhas de importantes obras de referência, sobretudo as de língua inglesa, permite concluir que o termo *thesaurus* aplica-se a, pelo menos, seis objetos distintos, que passaremos a enumerar.

Um 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). Outro emprego, não menos clássico, reveste-se do sentido de dicionário organizado em função de conceitos lexicalizados (Crystal, 1997) e refere-se ao **objeto 2**: um dicionário onomasiológico, um *tesauro* (Weiszflog, 1998) ou *tesouro* (Ferreira, *op. cit*.), ou ainda um *dicionário analógico* (Azevedo, 1983), cujo precursor foi o *Thesaurus de Roget* (Roget, 1953).[3] Ao utilizarmos os programas de busca da Internet para pesquisar o termo *thesaurus*,

[3] O *Thesaurus* de Roget, publicado em 1852, levou 12 anos para ser concluído. O *Dicionário analógico da língua portuguesa* de Azevedo (1983) foi uma tentativa única de se construir uma versão do dicionário de Roget para o português brasileiro.

encontramos um emprego particular, que atualiza o jargão corrente nos domínios da Informática e Documentação e refere-se ao **objeto 3**: "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.*).[4] Nessa acepção, todas as formas estão abonadas: *thesaurus* (Flexner, 1997; Neufeldt, 1997; Ferreira, *op. cit.*), *tesauro* (Ferreira, *op. cit.*), *tesouro* (Weiszflog, *op. cit.*). Um 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** - um dicionário organizado em termos de sinônimos e antônimos (Flexner, *op. cit.*; Neufeldt, *op. cit.*), revelou-se o mais apropriado para denotar o objeto que estamos construindo, 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, *op. cit.*; Weiszflog, *op. cit.*).

Duas observações importantes, decorrência desse levantamento, são suficientes para justificar a escolha do termo *thesaurus*. Por um lado, observamos que todos esse objetos, excetuando-se o objeto 1, possuem um traço comum: todos são dicionários particulares, estruturados segundo critérios específicos de natureza relacional como, por exemplo, relações conceptuais, relações léxico-semânticas, campos semântico-nocionais, sistemas terminológicos, entre outros. Em particular, com diferentes graus de prioridade, utilizam duas relações léxico-semânticas específicas: a semelhança e a oposição de sentidos, que no limite são as relações de sinonímia e antonímia. O que os diferencia são seu propósito, sua funcionalidade e o meio de sua disseminação. Com base nessas considerações descartamos o termo *tesouro*. Por outro lado, o termo *thesaurus*, considerado variante do termo *tesauro*, para fazer referência ao objeto 6, está abonado em Weiszflog (*op. cit.*), o que reduziu nossa decisão a uma simples escolha, que, apesar disso, encontra uma justificativa adicional pelo fato de Ferreira (*op. cit.*) restringir a aplicação do termo *tesauro* ao objeto 3.

2.3. Fontes de informação lexical

Feitos os estudos preliminares, problemas de natureza teórico-metodológica e conjunturais foram detectados:

- a seleção das entradas e a determinação das relações de sentido diante da complexidade da estrutura do léxico;
- a especificação formal dessas relações que refletisse motivação lingüística e, ao mesmo tempo, fosse computacionalmente tratável;
- a construção de um editor eficiente e amigável para agilizar a entrada automática ou semi-automática dos dados e permitir testes de consistência do grande volume desses dados e da intricada rede de relações que se estabelecem entre eles;
- a seleção de obras de referência que servissem de fontes para a extração do conhecimento lexical diante da inexistência de dicionários do português computacionalmente tratáveis;
- o estabelecimento de critérios de filtragem com o objetivo de minimizar as inconsistências, lacunas e imprecisões detectadas nas principais obras de referência analisadas;[5]
- a formação de uma equipe de trabalho interdisciplinar diante da carência de especialistas voltados para a lexicografia computacional.

Sem a pretensão de apontarmos solução para todos esses problemas, abordamos, a seguir, algumas das dificuldades que nos motivaram a utilizar, como fontes de conhecimento lexical, um conjunto de obras que, apesar de suas limitações, mostraram-se adequadas para nossos propósitos e são consideradas referências para muitas questões lexicográficas da língua portuguesa.

---

[4] Nesse sentido, o termo data da década de 50, quando H.P.Luhn, trabalhando para a IBM, propôs um processo computacional capaz de gerar uma lista de termos técnicos para indexar textos científicos (cf. http://www.britannica.com/bcom/eb/article/0/0,5716,109618+16+106477,00.html).

[5] Caberia, aqui, remeter o leitor à crítica severa que Cláudio Abramo fez contra Ferreira (*op. cit.*). O texto completo encontra-se publicado no Caderno *Mais*, Folha de São Paulo, 23/01/2000.

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. Como adverte Lyons (1979), 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 um thesaurus, como em outras obras dessa natureza, uma questão crucial é a delimitação das unidades lexicais que deverão 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, uma vez que há evidências que apontam para a sua realidade psíquica: indivíduos de sociedades ágrafas são capazes de ditar um texto palavra por palavra e a fala holofrástica da criança, por exemplo. Biderman (1978) aponta com precisão a questão da delimitação.[6] Nesse processo, diz ela, somos assaltados por uma dúvida: a seqüência de formas que estamos analisando constitui uma unidade do léxico, ou é uma combinatória de lexias, engendrada pela sintaxe? Criticamos com Biderman a persistência da gramática tradicional, que "iludida" pela forma ortográfica de um lexema, posto que nem sempre revela seu estado de lexicalização, considera lexias como *dor de cabeça*, *à toa* e *contanto que*, por exemplo, como "locuções". Observe-se que não estamos diante de construções sintáticas, posto que os elementos componentes encontram-se há muito soldados. Com efeito, não se pode dizer, no sentido de "enxaqueca", **dor persistente de cabeça*, **dor da cabeça*, **dor das cabeças*. Esse fato nos leva a reconhecer que essas unidades, lexias complexas, sofreram um processo de lexicalização e não são mais resultantes de operações sintáticas, o que garante a essas e outras tantas lexias o estatuto de legítimos lexemas e, portanto, passíveis de figurar como entradas.

Outra questão, também complexa, velha conhecida do lexicógrafo, é a variedade de tipos de significado, cuja discriminação coloca para o lingüista inúmeras dificuldades (Kilgarriff, *op.cit.*). Leech (1974: 26), por exemplo, distingue sete tipos de significado: conceptual (s*entido*), conotativo, estilístico, afetivo, refletido (*refleted*), de colocação e temático.

Por fim, existem 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, torna-se praticamente impossível sua descrição exaustiva. Assim, qualquer dicionário será necessariamente incompleto, refletindo os inevitáveis recortes feitos pelo dicionarista, que se esforça por descrever e registrar parte do patrimônio lexical de um estado de língua (Hartmann, *op. cit.*). Somam-se a esse os problemas apontados por Dubois *et al.*(1978): (a) a hesitação entre a impossível exaustividade e os limites materiais e práticos; (b) o 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) a dificuldade de distinguir entre o vocabulário geral e o de língua especial, uma vez que essas nuanças são de difícil observação; (d) a não transparência dos critérios que sancionam a passagem de um neologismo para o léxico consagrado da língua.

Diante do exposto, a reutilização de recursos disponíveis foi a saída que encontramos. Por um lado, essa estratégia possibilitou a agilização dos trabalhos, ao reduzir grande parte do trabalho à extração e filtragem dos conhecimentos lexicográfico e lingüístico, direta ou indiretamente, contidos nas obras de referência analisadas, que, inegavelmente são frutos de uma tradição centenária de pesquisa lexicográfica sobre o português brasileiro. Por outro lado, e decorrência dessa estratégia, a adoção do critério de abonação, além de minimizar a necessidade de investigação custosa e pontual de cada verbete a ser construído, elimina também as laboriosas pesquisas em *corpora* e garante a conformidade do thesaurus com os padrões de expressão da norma escrita sancionada pela tradição lexicográfica brasileira.

---

[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. Conforme define Biderman (*op. cit.*: 130): "Os lexemas se manifestam, no discurso [=fala], através de formas ora fixas, ora variáveis. [...] Assim, em português o lexema CANTAR pode manifestar-se discursivamente como *cantei*, *cantavam*, *cantas*, *cantando*, etc... A essas formas que aparecem no discurso, daremos o nome de *lexia*". Dubois *et al.* (1978: 360-361) 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.

Das obras analisadas, selecionamos sete, como fontes de conhecimento lexical: Nascentes (1981), Azevedo (1983), Borba (1990), Weiszflog (1998), Fernandes (1997), Ferreira (1999), Barbosa (1999). Além dessas fontes, dispomos ainda do *corpus* do NILC, com aproximadamente 30 milhões de palavras,[7] e do léxico do ReGra. O primeiro, enquanto fonte de informação lexical, deverá servir de parâmetro para decidirmos a inclusão ou não de determinadas entradas. O segundo deverá viabilizar a realização de testes de compatibilização entre ambas as bases de dados lexicais, o thesaurus e o próprio léxico do ReGra, o que contribuirá para o refinamento de ambos.[8]

3. Questões de representação: o conjunto de sinônimos

Uma questão que se coloca freqüentemente nas discussões sobre a sinonímia é o fato de as línguas naturais não apresentarem sinônimos perfeitos. Como conseqüência, diz Lutz (1994), um thesaurus limitado a registrar sinônimos exatos seria reduzido a uma lista de poucas palavras e, 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 já apontadas, julga necessário fazê-la, por necessidade de encontrar termos alternativos e mais eficientes para expressar suas idéias. Logo, o usuário busca palavras que mantenham entre si uma relação de semelhança de sentido (Cruse, *op. cit*.; Miller & Fellbaum, *op. cit*.; Ilari & Geraldi, 1985).Constata-se, com efeito, que é 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 com essa expectativa do usuário, sugerindo-lhe conjuntos de sinônimos e antônimos internamente coerentes e adequados ao estado de língua, deve também viabilizar acesso rápido aos milhares de sinônimos e antônimos sistematicamente relacionados às milhares de entradas que compõem a obra. Para isso, a questão de sua representação formal é crucial.

Encontramos uma solução para essa questão na metodologia empregada no desenvolvimento da rede *WordNet* (Miller *et. al*., 1990). 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, eliminando a necessidade de especificação do valor semântico, isto é, um rótulo conceitual para cada acepção de uma entrada, (ii) a noção constitutiva básica de *synset*, isto é, um conjunto de sinônimos e (iii) a noção de "matriz lexical", que postula uma correspondência biunívoca entre sentido e *synset*.

Há, porém, diferenças. Os objetivos dos dois aplicativos são bastante distintos. Como explicitamos anteriormente, o *TeP* pretende ser um objeto do tipo 6. Já a rede *WordNet* é um modelo computacional construído para responder à pergunta: "Qual é a natureza e organização dos conceitos lexicalizados que as palavras podem expressar?" (Miller e Fellbaum, *op. cit*.). Em outras palavras, o *WordNet* pretende ser um modelo da estrutura dos conceitos lexicalizados na língua inglesa.

Do ponto de vista lógico, a base do thesaurus estrutura-se como mostra a Fig.1. Sua estrutura é composta por duas listas principais: uma lista que contém todas as **Entradas** (LE), ordenadas alfabeticamente, e uma lista que contém uma coleção de **Conjuntos** (LC), os *synsets*, em que cada conjunto é formado a partir das entradas das acepções a que cada conjunto pertence. Cada entrada da LE, além de conter a representação ortográfica do lexema, contém uma lista de **Acepções** (LA). Cada acepção é formada pela dupla de sinônimos e antônimos que apontam para seus respectivos conjuntos na LC e por um campo que indica a qual entrada cada acepção pertence.

---

[7] 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/.

[8] No léxico do ReGra, a soma de todos os lexemas relevantes para o *TeP* (substantivos, verbos, adjetivos e advérbios) é 55.478.

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 precisa de uma lista de acepções para facilitar a pesquisa das 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.

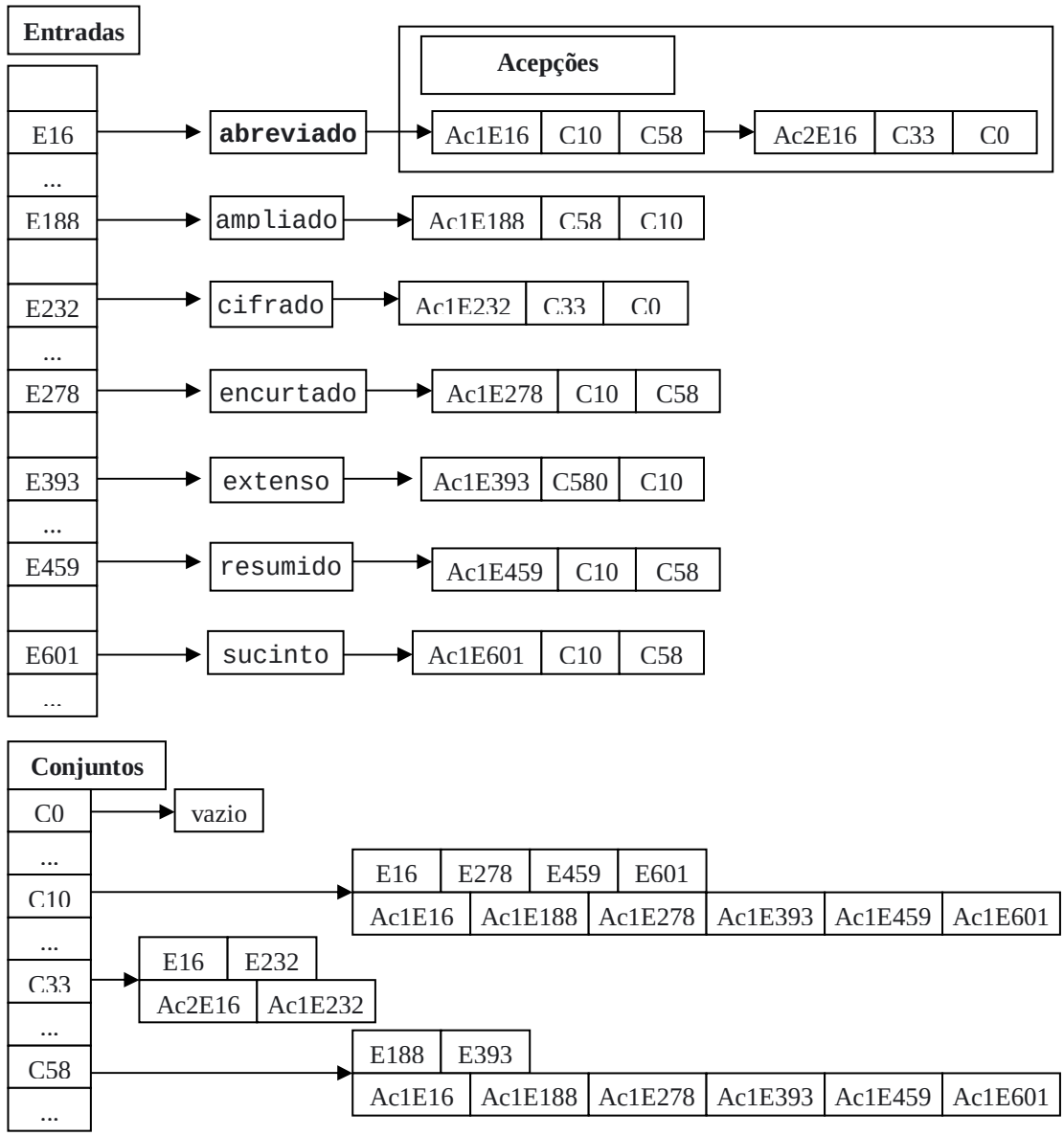

Figura 1: Exemplo de armazenamento da entrada **abreviado** na base do thesaurus. Os índices Exx, Cxx e AcxxEyy são apenas para facilitar a ilustração, pois a estrutura interna utiliza ponteiros (endereços de memória) para indexar esses campos.

4. Questões de implementação: o editor do thesaurus

Esta seção esboça o editor para dar suporte à montagem da base do thesaurus. Conforme a metodologia empregada, optamos pela utilização de dois tipos de representação em fases distintas. Na fase lingüística do processo, os lingüistas utilizam o editor para a montagem da base do thesaurus. Através dele, são criados e editados os conjuntos de sinônimos e antônimos. Arquivos no formato texto armazenam as informações dessa fase.Na fase de implementação, pretendemos utilizar um Sistema de Gerenciamento de Base de Dados, cuja função será unir as coleções de conjuntos de cada uma das categorias gramáticas e verificar as ambigüidades e inconsistências que, por ventura, possam

ocorrer entre os conjuntos, uma vez que o editor não processa todos eles ao mesmo tempo, pois, por razões metodológicas, cada lingüista trabalha uma única classe gramatical individualmente.
Com a função básica de agilizar a entrada de dados, permitindo a criação e manipulação de cada verbete, o editor apresenta para o lingüista uma interface gráfica e ferramentas que permitem a ele editar e gerenciar vários tipos de informações durante o processo de montagem da base: salvando os dados, desfazendo uma operação, editando campos, visualizando a base, listando entradas e verbetes, imprimindo partes da base e extraindo dados estatísticos como, por exemplo, o número de entradas e conjuntos, a proporção entre o número de entradas e o número de conjuntos inseridos e o número de entradas e verbetes gerados automaticamente, entre outras.
As principais funcionalidades do editor poderão ser melhor apreciadas através de um exemplo. Mostraremos como o lingüista montaria uma parte do verbete do adjetivo **abreviado**.
Após extrair e filtrar as informações provenientes das fontes apresentadas em 2.3. acima, sua tarefa é inserir os dados 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 Fig.2.[9]

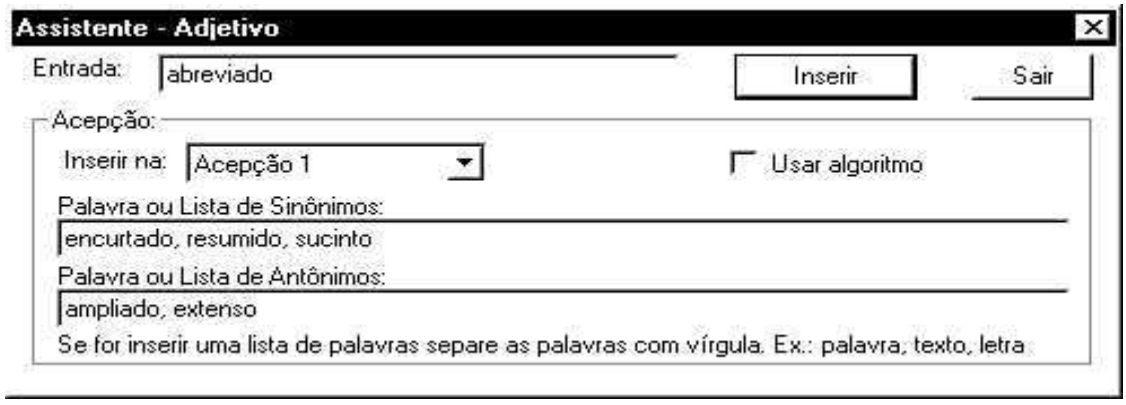

Figura 2: Montagem de parte do verbete do adjetivo **abreviado** no Assistente do editor.

Após o lingüista digitar **abreviado** no campo Entrada, o Assistente verifica se essa entrada já existe na base do thesaurus. Em caso afirmativo, os campos Palavra ou Lista de Sinônimos e/ou Palavra ou Lista de Antônimos são preenchidos com as informações recuperadas da base, permitindo que esses campos sejam editados. Em caso negativo, ele 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. Esse Assistente desempenha papel essencial no processo, uma vez que ele utiliza filtros 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 Fig.3.
Além de apresentar uma barra de ferramentas e uma barra de menus, o editor apresenta três frames redimensionáveis: dois superiores e um inferior. O frame superior à esquerda apresenta, em camadas, quatro tipos de listas: Todas as Entradas, Entradas Automáticas, Atualização Automática e Novas Entradas, todas em ordem alfabética. O frame superior à direita exibe o verbete completo, estruturado em forma de árvore, referente à entrada selecionada no frame à sua esquerda. Com exceção do item Acepção, que aparece na estrutura do verbete, todos os demais itens desse frame podem ser renomeados. Além disso, ao darmos um duplo clique no item raiz ou nos itens Acepção x, a árvore pode ser expandida ou contraída e ao darmos um duplo clique nos itens terminais é possível navegar pela base do thesaurus: ao clicarmos, por exemplo, a palavra **extenso**, enquanto antônimo de **abreviado**, o editor nos remete para o verbete correspondente.. Finalmente, o frame inferior exibe todos os conjuntos que contêm uma ocorrência da entrada selecionada.
Além dos recursos de edição, o editor, como pontuamos acima, oferece duas outras funcionalidades: a visualização de todos os conjuntos de sinônimos e antônimos contidos na base e a estatística da base.

[9] A discussão dos critérios de filtragem, imprescindíveis para garantir a precisão e consistência do thesaurus, estão discutidos em Moraes & Dias-da-Silva (2000).

Essa visualização tem por objetivo auxiliar o usuário a verificar os relacionamentos e os possíveis erros cometidos durante a inserção dos dados.

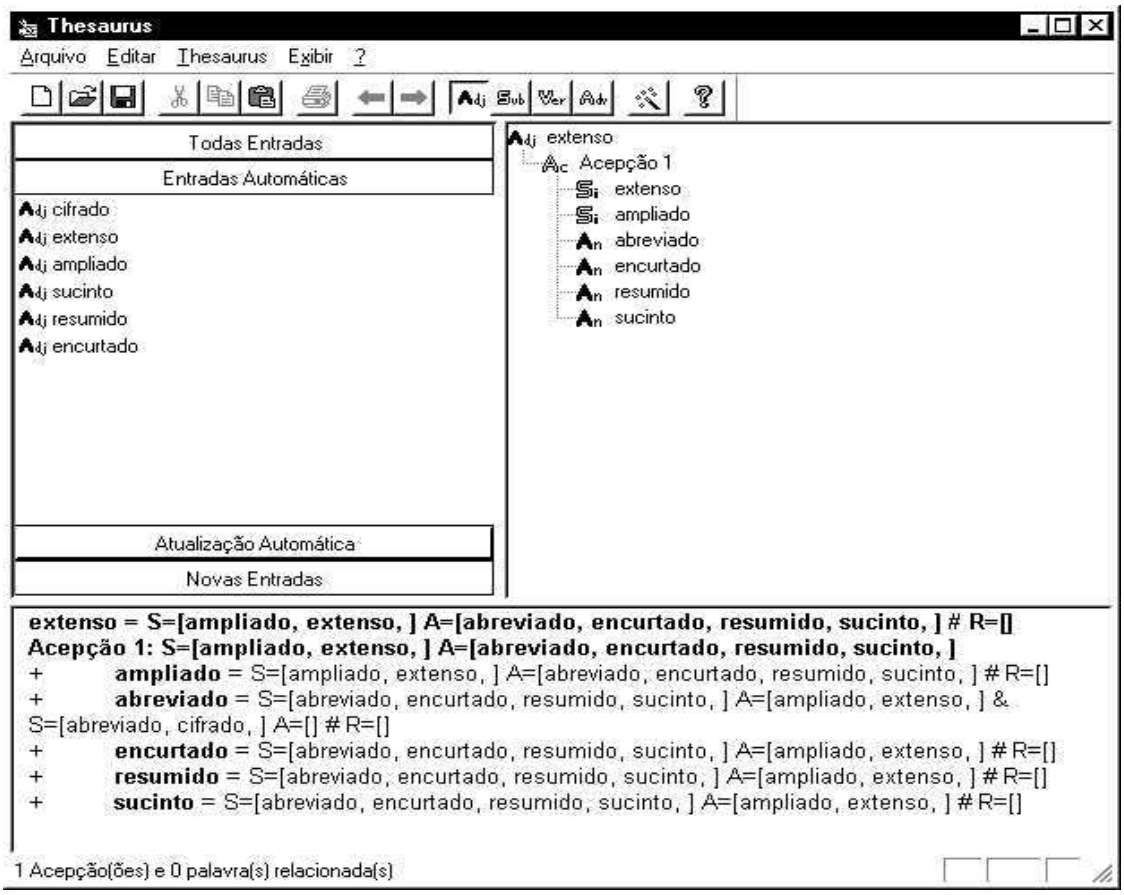

Figura 3: Ilustração do editor da base do thesaurus e a sua situação após clicarmos a palavra **extenso**, um dos antônimos de **abreviado**. A janela Entradas Automáticas mostra as entradas dos verbetes que foram criados automaticamente pelo editor.

5. Conclusão

Têm sido inúmeros os ganhos que vimos acumulando durante o desenvolvimento do *TeP*. Do ponto de vista do difícil trabalho interdisciplinar, há que se ressaltar a salutar troca de experiências e o profícuo exercício de construção do necessário diálogo cooperativo e colaborativo entre lingüistas e informatas. Do ponto de vista da pesquisa lingüística, o empreendimento tem contribuído para aguçar o senso de análise de fenômenos lingüísticos ligados à semântica lexical e à lexicalização, bem como selecionar, testar e refinar práticas lexicográficas.

Uma das grandes vantagens da metodologia descrita em 2.1. é o fato de permitir que cada membro da equipe de desenvolvimento do thesaurus, sem perder a visão global do processo, possa concentrar-se na solução dos problemas específicos de seu domínio. Com efeito, a análise das relações de sentido relevantes para o thesaurus, os critérios de seleção e filtragem das informações lexicais extraídas das fontes selecionadas e a montagem dos conjuntos de sinônimos e antônimos, bem como a escolha e aplicação de métodos de abordagem são atividades próprias da fase lingüística. Na fase de representação, a discussão de questões referentes à escolha ou à proposição de sistemas de representação das entradas e de estratégias de codificação dos elementos trabalhados no domínio anterior possibilitou um rico intercâmbio de informações e conhecimentos entre especialistas das duas áreas envolvidas. Na fase de implementação, além de buscar soluções para questões que envolvem a criação de programas, e que dizem respeito à montagem global do sistema computacional em que o programa deverá ser alojado, os especialistas da computação tiveram a oportunidade de apreciar com maior profundidade os resistentes problemas postos pela linguagem humana, que parece resistir a qualquer tentativa de ser reduzida a um código de máquina.

Do ponto de vista da reutilização de materiais, os estudos realizados sobre as possibilidades de extração do conhecimento lexical, implícita ou explicitamente contido nos mais variados tipos de dicionários analisados, e o estabelecimento de critérios de filtragem das informações deles extraídas demonstram a viabilidade de sua utilização no processo de montagem de bases de dados lexicais.

Enquanto produto, acreditamos que o *TeP* deverá ser um rico recurso de auxílio à expressão escrita, oferecendo ao usuário do português a oportunidade de consultar e escolher *on line* vocábulos que ele, por motivos já apontados neste trabalho, decide substituir.

Em termos quantitativos, os esforços da equipe têm sido recompensados com os resultados.No estágio atual, a base do thesaurus conta com cerca de 20000 entradas, assim distribuídas: 6000 verbos, 2000 substantivos e 12000 adjetivos.

Trabalhos futuros deverão incluir a execução das etapas 6, 7 e 8, mencionadas na introdução deste trabalho, e os necessários refinamentos, ajustes e ampliação da base do thesaurus, bem como a sua integração à base de dados lexicais, atualmente em desenvolvimento no âmbito do NILC.

Referências

Azevedo, F.F.S. (1983). *Dicionário analógico da língua portuguesa*. Thesaurus, Brasília.

Barbosa, O. (1999). *Grande dicionário de sinônimos e antônimos*. Ediouro, Rio de Janeiro.

Biderman, M.T.C. (1978). *Teoria lingüística*. Livros Técnicos e Científicos, Rio de Janeiro.

Borba, F.S. (coord.*)* (1990) *Dicionário gramatical de verbos do português contemporâneo do Brasil*. Fundação Editora Unesp, São Paulo.

Briscoe, E.J. & B. Boguraev, (eds) (1989) Computational lexicography for natural language processing. Longman, London/New York.

Cruse, D.A. (1986). *Lexical semantics*. Cambridge University Press, New York.

Crystal, D. (1997). *The Cambridge encyclopedia of the English language*. Cambridge University Press, Cambridge.

Dias-da-Silva, B.C. (1998a). Os Domínios Lingüístico e Tecnológico do Estudo do Processamento Automático das Línguas Naturais. *Estudos Lingüísticos*, 26, pp. 612-617.

_____ (1998b). Bridging the gap between linguistic theory and natural language processing In: *Proceedings of the 16th International Congress of Linguists*. Elsevier-Pergamon, Oxford, Paper 0425, 10 p.

Dubois, J. et al. (1978). *Dicionário de lingüística*. Trad. Izidoro Blikstein. Cultrix, São Paulo.

Fernandes, F. (1977). *Dicionário de sinônimos e antônimos da língua portuguesa*. Globo, São Paulo.

Ferreira, A.B.H. (1999). *Dicionário Aurélio eletrônico século XXI* (v. 3.0). Lexikon Informática Ltda, São Paulo..

Flexner, S.B. (ed.) (1997). *Random house Webster´s unabridged electronic dictionary* (v. 2.0). Random House Inc., New York.

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

Hartmann, R.R.K. (1983). *Lexicography: principles and practice*. Academic Press, London.

Ilari, R. & Geraldi, J.W. (1985). *Semântica*. Ática, São Paulo.

Kilgarriff, A. (1992). Dictionary Word Sense Distinctions: A Linguistic Evaluation. *Computers and the Humanities*, 26 (5-6), pp.365-387.

Landau, S. I. (1996). *Dictionaries: the art and craft of lexicography*. Scribners, Cambridge.

Leech, G. (1974). *Semantics*. Penguin Books, Middlessex.

Levin, B. & Pinker, S. (eds) (1991). Lexical and conceptual semantics. *Cognition*, 41 (1-3), pp. 1-7.

Lutz, W.D. (1994). *The Cambridge thesaurus of American English*. Cambridge University Press, Cambridge.

Lyons, J. (1979). *Introdução à lingüística geral*. Trad. Isaac Nicolau Salum. Editora Nacional/Edusp, São Paulo.

McCreary, D.R.(1996). Computational lexicology and lexicography. *Language*, 72, p. 4-49.

Miller, G. A. & Fellbaum, C. (1991). Semantic Networks of English. *Cognition*, 41(1-3), pp 197-229.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross & K. Miller. (1990). Five Papers on WordNet. CSL Report 43. Cognitive Science Laboratory, Princeton University.

Moraes, H.R. & Dias-da-Silva, B.C. (2000). A Questão da Representação Lingüístico-Computacional da Sinonímia e Antonímia na Compilação de um Thesaurus Eletrônico. In: *XII Congresso de Iniciação Científica*, IBILCE-UNESP, São José do Rio Preto. (no prelo)

Nascentes, A. (1981). *Dicionário de sinônimos*. Nova Fronteira, São Paulo.

Neufeldt, V. (ed.) (1997). *Webster´s new world dictionary & thesaurus* (v. 1.0). Macmillan Publishers, New York.

Nunes, M.G.V. et al. (1996). "A Construção de um Léxico da Língua Portuguesa do Brasil para Suporte à Correção Automática de Textos". Relatórios Técnicos do ICMSC, n. 42, São Carlos, 37 p.

Roget, P.M. (1953). *Roget´s thesaurus* (ed. original, 1852). Penguin Books, Middlessex.

Saint-Dizier, P., & Viegas, E. (1995). *Computational lexical semantics*. Cambridge University Press. Cambridge.

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