A construção da base da Wordnet.Br: conquistas e desafios

Bento Carlos Dias-da-Silva[1,2]

[1]Centro de Estudos Lingüísticos e Computacionais da Linguagem (CELiC)
Faculdade de Ciências e Letras (FCL) – Universidade Estadual Paulista (UNESP/Ar.)
Caixa Postal 174 – 14.800-901 – Araraquara – SP – Brazil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brazil

{bento@fclar.unesp.br}

Abstract. *This paper presents the state of the art of the development of the Wordnet.Br lexical database. It starts off with the overall planning of the project, describes its challenges and main achievments, and concludes with an overview of the ongoing work.*

Resumo*. Este trabalho apresenta o estado da arte do desenvolvimento da base lexical da Wordnet.Br. Parte da exposição do planejamento geral do projeto, descreve, na seqüência, seus desafios e suas principais conquistas e conclui com a indicação do trabalho em andamento.*

1. Introdução

Compartilhando conquistas e partilhando desafios, este artigo visa a divulgar o estágio atual do projeto de construção da base de uma *wordnet* para o português brasileiro, reportando-se aos resultados das investigações iniciais que ocorreram entre janeiro de 2002 e junho de 2003 e ao projeto "Montagem da Base Wordnet para o Português do Brasil", desenvolvido entre junho de 2003 a maio de 2004.[1]

Esse percurso é apresentado em três seções: na seção 2, discute-se a fase inicial do projeto, em que se esboça o plano global de pesquisa e desenvolvimento (*O Plano de P&D da Base da Wordnet.Br*) que norteia a compilação da base, incluindo nessa exposição os resultados iniciais; na seção 3, aborda-se a "fase CNPq", em que se constituiu a equipe de P&D e, com ela, a construção parcial das bases de verbos, adjetivos e advérbios (*Fundamentos, Atividades e Resultados*); na seção 4, sumarizam-se a importância científico e tecnológica do empreendimento e o trabalho em andamento (*Conclusões e Perspectivas*).

2. O Plano de P&D da Base da Wordnet.Br

Partindo de um recurso lexical construído em um empreendimento anterior, o *TeP*[2], o problema em torno do qual este artigo se articula é a tarefa de desenvolvimento de uma

[1] Essa fase do projeto contou com auxílio do CNPq: Processo 552057/01-0, "Chamada-CNPq 09/2001 - Conteúdos Digitais - Edital SocInfo/ProTeM-01/2001".

[2] Trata-se do ***T****hesaurua* ***e****letrônico do* ***P****ortuguês*: um dicionário eletrônico de sinônimos e antônimos formado por cerca de 44 mil unidades lexicais [Dias-Da-Silva, 2003a; Dias-Da-Silva e Moraes 2003].

base relacional de dados lexicais para o português do Brasil nos moldes da *WordNet de Princeton* [Fellbaum 1999], doravante referida por *WordNet 1.5*. e descrita na Seção 3.1.[3]

O sucesso desse empreendimento norte-americano pode ser aferido por suas extensões, por sua sua relevância científica e pelo seu potencial tecnológico. Além de estimular a construção de wordnets para o português europeu [Marrafa 2001] e para oito línguas da União Européia, no âmbito do amplo projeto *EuroWordNet* [Vossen 1998], a associação internacional *Global Wordnet Association* [GWA 2004] registra wordnets em desenvolvimento para mais de 35 línguas diferentes. Do ponto de vista científico-tecnológico, a implementação de wordnets para diferentes línguas contribui, de um lado, para a produção de obras de referência, tanto lingüística como lexicográfica, para ensino e pesquisa da linguagem humana e, de outro, para o aprimoramento do desempenho qualitativo de sistemas de processamento automático de língua natural e de informação como, por exemplo, os sistemas de tradução automática, sumarizadores e motores de busca na Internet [Harabagiu e Moldovan 2000].

2.1 O Plano de Ação

A iniciativa de construção da base da Wordnet.Br, partindo das listas de conjuntos de sinônimos da base do TeP, aqui referidos como *synsets brutos*,[4] e fundamentando-se nas metodologias desenvolvidas para a construção da WordNet 1.5 e de redes particulares da EuroWordNet, previu o plano de ação descrito na Figura 1.

Atividades	
a.	Gerar, para cada unidade lexical (verbo, substantivo, adjetivo e advérbio) que constitui cada synset bruto da base do TeP, sua respectiva *concordância*, a partir do *córpus de referência*.
b.	Analisar a boa-formação gráfica e léxico-semântica de cada synset, isto é, avaliar se as unidades lexicais que o compõem estão grafadas corretamente e até que ponto elas *lexicalizam* um mesmo conceito, a partir das concordâncias e da consulta aos dicionários referidos em d abaixo.
c.	Especificar uma *glosa* para cada synset analisado.
d.	Especificar, para cada unidade lexical, pelo menos uma frase-exemplo, selecionada, nesta ordem de preferência: 1. entre as concordâncias ou 2. entre frases extraídas de textos em português brasileiro coletados diretamente na internet ou 3. entre as abonações registradas em um dos três dicionários Weiszflog (1998), Ferreira (1999) e Houaiss (2001).
e.	Especificar entre os synsets as relações léxico-conceituais descritas da Figura 2.
f.	Desenhar e implementar o *editor* para a montagem e gerenciamento da base da Wordnet.Br.
g.	Investigar a viabilidade técnica de se associarem os synsets da base da Wordnet.Br a synsets equivalentes da base da WordNet 1.5, com vistas à implementação de uma base bilíngüe inglês-português brasileiro.

Figura 1. O Plano de Ação

Relações de Léxico-Conceituais		
Tipo de Relação	**Classe Lexical Relevante**	**Exemplo**
Antonímia	Substantivo Verbo Adjetivo	homem/mulher entrar/sair bonito/feio
Hiponímia/Hiperonímia	Substantivo	rosa/flor
Troponímia	Verbo	caminhar/mover
Meronímia (parte-todo)	Substantivo	cabeça/nariz
Acarretamento	Verbo	comprar/pagar
Causa	Verbo	matar/morrer

Figura 2. Relações léxico-conceituais de uma *wordnet*

[3] Embora a WordNet de Princeton já se encontre na versão 2.0, a referência deste trabalho é a versão 1.5, sem qualquer prejuízo das conclusões e generalizações apresentadas.

[4] A noção de *synset*, bem como tantas outras noções (*concordância*, *córpus de referência*, *lexicalização*, *glosa*, *editor*, entre outras), serão oportunamente especificadas no decorrer da exposição.

2.2 A Quantificação da Equipe

Para realizar esse Plano de Ação, além da montagem de uma infraestrutura de recursos materiais (espaço, equipamentos de informática, com acesso à internet, e software), planejou-se a constituição de uma equipe de especialistas conforme a descrição da Figura 3.

Plano de Ação	Equipe
Montagem das concordâncias	8 lingüistas
Investigação da boa-formação dos synsets	
Especificação das glosas	
Coleta e seleção das frases-exemplo	
Especificação das relações léxico-conceituais	
Desenho e implementação do editor	2 cientistas da computação
Estudo da co-indexação entre a base da Wordnet.Br e a base da WordNet 1.5	

Figura 3. Constituição da Equipe de P&D

2.3 Desafios e Primeiras Conquistas

Durante as investigações iniciais, entre janeiro de 2002 e junho de 2003, sem a constituição da equipe ideal acima prevista,[5] além da necessária compilação do conhecimento teórico e aplicado necessário para a construção da base da Wordnet.Br, cinco atividades, com diferentes graus de profundidade e sistematicidade, foram realizadas:

- A conclusão positiva a respeito da possibilidade de relacionamento entre os synsets da base da Wordnet.Br e os synsets equivalentes da WordNet 5.1;
- A preparação das concordâncias para aproximadamente mil unidades lexicais;
- O refinamento de cerca de 700 verbos;
- A especificação de frases-exemplo para cerca de 500 verbos e 500 adjetivos.
- O planejamento da estrutura conceitual do editor da base da Wordnet.Br e teste do protótipo.

Nesse período, três dificuldades se fizeram sentir. As duas primeiras dizem respeito, respectivamente, à dificuldade do gerenciamento de 44 mil unidades lexicais (11.000 verbos, 15.000 substantivos, 16.000 adjetivos e 1.000 advérbios), distribuídas em cerca de 20 mil synsets, e à complexiadade da análise da consistência léxico-conceitual desses synsets brutos. A terceira refere-se à contratação e à formação do pessoal qualificado para a realização das tarefaso. Como as atividades neste campo de pesquisa exigem não só conhecimentos lingüísticos como também conhecimentos de informática e de ciências da computação [Dias-Da-Silva 1998, 2003b], foram investidos mais de quatro meses para a formação de competências em seções exploratórias de aplicação dos métodos e técnicas de análise léxico-semântica e de manipulação de software.

[5] A equipe contou apenas com três lingüistas e um cientista da computação.

3. Fundamentos, Atividades e Resultados

Nesta seção, apresentam-se os fundamentos, atividades e resultados da fase do projeto em que se constituiu a equipe de P&D, composta de 7 lingüistas e dois cientistas da computação, que realizou a construção parcial das bases de verbos, adjetivos e advérbios, no período de junho de 2003 a maio de 2004 [Dias-da-Silva, Rocha e Nunes 2004].

3.1 A fase CNPq

Nesta fase, cumpriu-se o Plano de Ação da Figura 4, recortado do plano apresentado na Figura 1.

Atividades
a'. Gerar, para cada unidade lexical que constitui cada synset bruto da base do TeP especificado em d' abaixo, sua respectiva concordância, a partir do *córpus de referência*.
b'. Analisar a boa-formação gráfica e léxico-semântica de cada synset, isto é, avaliar se as unidades lexicais que o compõem estão grafadas corretamente e até que ponto elas *lexicalizam* um mesmo conceito, a partir das concordâncias e da consulta aos dicionários referidos em d' abaixo.
d'. Especificar, para as unidades lexicais de cerca de 3.600 synsets brutos de verbos, 130 synsets brutos de adjetivos e 560 synsets brutos de advérbios, pelo menos uma frase-exemplo, selecionada, nesta ordem de preferência: 1. entre as concordâncias ou 2. entre frases extraídas de textos em português brasileiro coletados diretamente na internet ou 3. entre as abonações registradas em um dos três dicionários Weiszflog (1998), Ferreira (1999) e Houaiss (2001).
f'. Implementação parcial do editor para a montagem da base da Wordnet.Br.

Figura 4. O Plano de Ação CNPq

3.2 As Wordnets e a Wordnet.Br

Visando a emular o léxico mental, as *wordnets* (“redes de palavras”) são bases relacionais de dados, no sentido computacional do termo, formadas por unidades lexicais de uma língua natural, isto é, por palavras e expressões compostas de mais uma palavra.

A wordnet norte-americana [Miller e Fellbaum 1991; Fellbaum 1999], também conhecida como "a WordNet de Princeton" e disponível para acesso na internet [Wordnet 2004] assemelha-se à um *thesaurus eletrônico* "léxico-conceitual", isto é, à um dicionário eletrônico de sinônimos e antônimos enriquecido com a especificação das outras relações de natureza lógico-conceitual.

Nela, a estruturação das unidades sinônimas, distribuídas entre as quatro categorias lexicais especificadas na base de dados da rede (94.000 substantivos; 10.000 verbos; 20.000 adjetivos; 4.500 advérbios), materializa-se em "conjuntos de sinônimos", isto é, em *synsets* (forma abreviada do termo inglês *synonym sets*), que representam os conceitos lexicalizados pelas unidades lexicais do inglês que os compõem, dado que essa rede foi construída para o inglês norte-americano. Observe-se que a relação de sinonímia, propriedade léxico-semântica das unidades lexicais, é modelada pela relação matemática de pertença a um conjunto: "A unidade U1 é sinônima da unidade U2 se e somente se ambas pertencerem ao mesmo synset."

Graficamente, a coleção de synsets materializa-se nos pontos ou nós que formam a rede. Já as relações rotuladas constituem os arcos que ligam os diferentes nós. Computacionalmente, os arcos são implementados como ponteiros. Para auxiliar o usuário na identificação do conceito lexicalizado no synset, a rede registra-se, para cada

synset, uma glosa, isto é, uma definição informal desse conceito. Por fim, para ilustrar o contexto de uso de uma unidade lexical, associa-se a ela uma frase-exemplo.

Já a *EuroWordNet* pode ser visualizada como uma megawordnet, em que wordnets de diferentes línguas interligam-se por meio de diversas relações de correspondência. Nesse empreendimento multilíngüe, uma das principais tarefas é estabelecer a relação de equivalência léxico-conceitual os synsets de cada uma das redes locais e os synsets da *WordNet de Princeton*. É por meio desse tipo de relação de correspondência que os synsets de línguas diferentes que lexicalizam um mesmo conceito são interligados. Essa conexão é intermediada por um "registro", isto é, por um synset (com seu respectivo número de identificação e respectiva glosa) da wordnet norte-americana. A lista de todos os registros é denominada *Inter-Lingual-Index* ("Índice Interlingual"), ou simplesmente ILI.

3.3 A Sinonímia e a Matriz Lexical

Três noções são essenciais para a compreensão de como as wordnet se estruturam: a noção de sinonímia fraca, a noção de matriz lexical e as relações léxico-conceituais que se estabelecem entre synsets

A sinonímia fraca é assim definida: duas expressões são sinônimas num contexto lingüístico C (em que C é uma frase, por exemplo) se a substituição de uma pela outra em C não alterar o significado de C.[6] Essa relação de sentido é responsável pela estruturação básica de uma wordnet e, como já se disse, materializa-se nos synsets. A Figura 5 ilustra o Synset3723={fiscalizar, patrulhar, policiar, rondar} da base da Wordnet.Br, que evoca o sentido de "examinar cuidadosamente".

Synset3723:
{*fiscalizar*: O censor israelense pode [fiscalizar os atos da administração da justiça, ainda que não possa se envolver nas decisões da magistratura.,
patrulhar: A esquerda deve [patrulhar as escolhas de FHC.,
policiar: A inflação é ínfima e isso acontece não porque haja um órgão estatal encarregado de [policiar os preços ou porque as pessoas tenham um espírito patriótico e zelem para que os preços permaneçam estáveis.,
rondar: A esquadra M]rondava os portos da ilha. }

Figura 5. O Synset 3723 que evoca o sentido "examinar cuidadosamente "

O teste que demonstra a boa-formação do Synset3723 é o fato de todos os verbos que o compõem poderem ser intersubstituíveis em cada uma das frases-exemplo selecionadas. A Figura 6 exemplifica esse teste com o esqueleto da frase `[A esquerda deve _____ as escolhas de FHC]`, extraído do próprio synset. A verificação da relação de simetria fornece um teste complementar de consistência do synset e de adequação da frase selecionada: *fiscalizar* é *patrulhar*, *patulhar* é *policiar*, *policiar* é *rondar*, *rondar* é *fiscalizar*.

[A esquerda deve _____ as escolhas de FHC]
A esquerda deve fiscalizar as escolhas de FHC.
A esquerda deve patrulhar as escolhas de FHC.
A esquerda deve policiar as escolhas de FHC.
A esquerda deve rondar as escolhas de FHC.

[6] Lyons (1977), Cruse (1985), Ilari e Geraldi (2000) e Handke (1995) fornecem os subsídios essenciais para a compreensão da noção de sinonímia no sentido geral.

Figura 6. Exemplo de um esqueleto da frase empregado para a avaliação da consistência semântica do synset.

Já a noção de matriz lexical possibilita a construção de synsets de modo independente da especificação explícita do conceito por ele lexicalizado. Esse modelo tem como fundamento: (a) a adoção do modelo relacional de representação do significado lexical, também conhecido com método diferencial de representação, que parte do princípio de que a ativação de um conceito lexicalizado na mente do falante realiza-se por meio da ativação do conjunto de formas lexicais que o lexicaliza; (b) a noção de matriz lexical, que formaliza a correspondência biunívoca que se estabelece entre a forma e o significado das unidades lexicais de uma língua. Conforme ilustra a Figura 7, a matriz lexical é visualizada por meio de um plano cartesiano em que, no eixo das abscissas estão representadas as formas lexicais (F1,..., Fn) e, nos eixos das ordenadas, estão representados os synsets (S1,..., Sn), que representam os conceitos lexicalizados.[7]

SYNSETS (conceitos lexicalizados)	**FORMAS LEXICAIS**								
	F1 *carecer*	**F2** *demandar*	**F3** *necessitar*	**F4** *pedir*	**F5** *precisar*	**F6** *querer*	**F7** *reclamar*	**F8** *requerer*	**F9** *faltar*
S1{carecer; demandar; necessitar; pedir; precisar; querer; reclamar; requerer}	**S1*F1**	**S1*F2**	**S1*F3**	**S1*F4**	**S1*F5**	**S1*F6**	**S1*F7**	**S1*F8**	
S2{carecer; faltar}	**S2*F1**								**S2*F9**
S3 {carecer; necessitar; precisar}	**S3*F1**		**S3*F3**		**S3*F5**				

Figura 7. Exemplo da matriz lexical que representa os conceitos lexicalizados pelas formas *carecer*, *demandar*, *necessitar*, *pedir*, *precisar*, *querer*, *reclamar*, *requerer*, *faltar*

Enquanto a relação de sinonímia se estabelece entre unidades lexicais, as demais relações de significado necessárias para a constituição de uma wordnet (e já exemplificadas na Figura 2) são de natureza lógico-conceitual e estabelecem-se entre synsets.

[7] Observe-se que a sinonímia e a polissemia estão representadas, respectivamente, nas linhas e nas colunas da matriz. Formas sinônimas compartilham a mesma linha da matriz; formas polissêmicas ocorrem em linhas distintas.

3.4 As Análises

A coleta e seleção das frases-exemplo foi realizada no córpus de referência do projeto, composto por três fontes digitalizadas de informação lexical, apresentadas na ordem de prioridade de pesquisa [Dias-Da-Silva, Oliveira e Moraes 2003]: (i) o Córpus do NILC [Corpus Nilc 2004], composto por textos escritos em português do Brasil, nos registros jornalístico, didático e epistolar; (ii) textos do português do Brasil localizados na Internet por meio do motor de busca *Google*; (iii) as abonações registradas nos dicionários mencionados neste trabalho. A Figura 8, além de outras, descreve a convenção adotada para indicar a fonte de origem das frases-exemplo, conforme exemplificadas no synset abaixo:

Synset3665
{
abafar: Os deputados foram acusados de tentar [abafar escândalos de corrupção que os atingem.,
acafelar: H]Acafelar as más intenções.,
acobertar: O poeta vai tocando fatos, costumes, homens e versos petrificados pelo hábito que torna cego, surdo e insensível ao escândalo que [acoberta. ,
amofumbar: no_sentence,
dissimular: [Dissimulava com um discurso talvez até convincente toda a frustração por não ter comemorado ainda uma vitória.,
emascarar: Se tenta I]emascarar ou disfarçar o problema. ,
empanar: O Paternalismo se presta à tarefa de I]empanar as mazelas do Sistema Capitalista, aceitando paliativos como solução de problemas sociais graves, que o próprio Sistema se reconhece incapaz de resolver. #Na etapa de difusão, o lavador de dinheiro tentar I]empanar ainda mais a trilha que liga os fundos à atividade criminosa. ,
encobertar: Dúvidas sobre o papel de Kennedy e alegações de que sua responsabilidade pelo acidente foi [encoberta continuam a persegui-lo.# Nada é [encoberto pelo homem, que não seja revelado por Deus. ,
encobrir: O governador há muito patrocina contra ele uma campanha de desgaste político, para [encobrir os erros de sua própria administração.,
mascarar: O marxismo ainda é a matriz de qualquer heresia que combata os dogmas antigos que [mascaram a opressão e justificam a desigualdade.,
ocultar: A relevância dos números, na verdade, [oculta um problema para o PT. ,
rebuçar: Quanto à justiça, Sócrates, longe de I]rebuçar sua opinião, patenteava-a por atos.,
reprimir: Nessa época, subir no high life ainda não exigia do sujeito que dissimulasse ou [reprimisse os sentimentos.,
sufocar: Com esta única lição renasce-me a mocidade que eu estupidamente me empenhava em [sufocar! ,
velar2: Tenho ainda a necessidade de I]velar a angústia.,
}

Símbolo	Função
¶{	identificar início de synset
}¶	identificar final de synset
¶$...$;	identificar os limites de unidade
,¶	identificar final de frase-exemplo
[	sinalizar unidade e identificar frase-exemplo extraída do Córpus do NILC
I]	sinalizar unidade identificar frase-exemplo extraída da Internet
A]	sinalizar unidade identificar frase-exemplo extraída do Aurélio
M]	sinalizar unidade identificar frase-exemplo extraída do Michaelis
H]	sinalizar unidade identificar frase-exemplo extraída do Houaiss
no_sentence	identificar a não ocorrência de frases-exemplo no córpus de referência
#	identificar limites entre duas frases-exemplo

Figura 8. Convenções notacionais: identificadores

3.5 Os Resultados

AlAs estatísticas da relação de sinonímia, relação responsável pela formação dos synsets, mostram que, no estágio atual de desenvolvimento, a base da Wordnet.Br contém cerca de 44 mil formas lexicais, agrupadas em 4.129 synsets de verbos (11 mil formas), 8.526 synsets de substantivos (17 mil formas), 6.648 synsets de adjetivos (15 mil formas) e 566 synsets de advérbios (mil formas). As estatísticas da relação de

antonímia registram os seguintes dados: 1.158 pares de synsets antônimos para a classe de verbos, 1.412 pares para a classe dos substantivos, 1.564 pares para a classe de adjetivos e 150 pares para a classe de advérbios.

O Quadro 1 registra as principais estatísticas do trabalho realizado, resultado da análise de mais de 20 mil unidades lexicais, componentes de mais de 4 mil synsets, da coleta e seleção de mais de 17 mil frases-exemplo extraídas do córpus de referência, a partir de análise de concordâncias geradas por concordanceadores, como o "WordSmith Tools", e da inserção, por meio de um editor especificamente construído para gerenciar dados e relações, dessas frases-exemplo na base Wordnet.Br.

Quadro 1. Estatísticas do trabalho realizado

IDENTIFICAÇÃO - VERBOS	**GLOSAS 0000-2930, GLOSAS 3007-3427, GLOSAS 3801-4128**
N° de synsets analisados	3.677
N° de formas verbais analisadas	19.003
N° de frases-exemplo selecionadas	16.008
N° de no_sentence/N° de DEVIANT	3.059/220
IDENTIFICAÇÃO - ADVÉRBIOS	**GLOSAS 001-566**
N° de synsets analisados	566
N° de formas adverbiais analisadas	1.495
N° de frases-exemplo selecionadas	1.128
N° de no_sentence/N° de DEVIANT	367/35
IDENTIFICAÇÃO - ADJETIVOS	**GLOSAS 1617-1755**
N° de synsets analisados	139
N° de formas adjetivais analisadas	363
N° de frases-exemplo selecionadas	248
N° de no_sentence/N° de DEVIANT	93/05
TOTAL DA TRÊS CLASSES (Arquivo com 39.279 linhas)	
N° de synsets analisados	4.382
N° de formas analisadas	20.861
N° de frases-exemplo selecionadas	17.384
N° de no_sentence/N° de DEVIANT	3.519/260

4. Conclusões e Perspectivas

Diante da importância cultural, científica e tecnológica do projeto de construção de uma wordnet, por questões apenas circunstanciais de ordem prática, antes de abordar a análise dos substantivos, dos demais adjetivos e advérbios, o trabalho atual concentra-se no estabelecimento da correspondência entre as bases de verbos do português e da WordNet 1.5. Como para aquela, para a base de verbos da Wordnet.Br, deverão ser especificadas, manualmente, as glosas (isto é, uma definição do significado do conceito implicitamente codificado em cada synset e que constitui a definição de um ontologia recortada por uma língua natural) e, semi-automaticamente, as relações lógico-conceituais e hierárquicas de troponímia, causa e acarretamento, específicas para os synsets de verbos.

Em termos gerais, destaca-se a relevância tecnológica (disponibilização livre de léxicos computacionais robustos e semanticamente hierarquizados e enriquecidos com informações relacionais e conceituais) e científica das wordnets (estudo das propriedades lexicais, dos padrões de lexicalização, da materialização de ontologias em léxicos computacionais e da própria constituição do léxico mental). Já em termos específicos, a relevância das wordnets para a tecnologia da informação está no fato de

seus resultados preverem um recurso lingüístico-computacional específico para auxiliar sistemas de processamento computacional de informações tanto na tarefa de busca de informação codificada em textos escritos do português quanto na tarefa de representação, indexação e sistematização computacionalmente tratáveis dessa informação [Saint-Dizier e Viegas 1995; Iwanska e Shapiro 2000]. Em termos operacionais, esse recurso poderá ser integrado aos mais diversos tipos de sistemas, potencializando suas taxas de precisão e recuperação da informação [Harabagiu e Moldovan 2002]: sistemas de recuperação de textos e de informação, desambiguização de palavras, sumarização de textos e motores de busca na Internet. Em particular, ele poderá integrar diferentes tipos de sistema: como, por exemplo, "extração de informação de bases textuais", "indexação semântica latente para ranqueamento", "mineração de textos e de estruturas ontológicas" e "encadeamento lexical". Por fim, agrega-se, a essas, a relevância cultural: a geração totalmente automática de dois tipos de dicionários digitais: um dicionário monolíngüe do português e um dicionário bilíngüe inglês-português, ambos com a possibilidade de disponibilização gratuita em rede e de atualização permanente dos seus verbetes

Referências

Corpus NILC (2004) *Córpus do Nilc*. Disponível em http://www.linguateca.pt/. Acesso em: 13 ago. 2004.

Cruse, D.A. (1986) *Lexical semantics*. Cambridge, Mass: Cambridge University Press.

Dias-Da-Silva, B.C. (2003a) O TeP: construção de um thesaurus eletrônico para o português do brasil. *Boletim da ABRALIN*. Fortaleza: Imprensa Universitária, v.26, número especial, p.86-89.

_____ (2003b) Human language technology research and the development of the Brazilian Portuguese wordnet. In: Haji•ová, E., Kot•šovcová, A., Mírovský, J. (Ed.). *Proceedings of the 17th International Congress of Linguists*. Prague: Matfyzpress, MFF UK,12p. 1 cd.

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

Dias-Da-Silva, B.C., Moraes, H.R. (2003) A construção de thesaurus eletrônico para o português do Brasil. *Alfa*, v.47, n.2, p.101-115.

Dias-Da-Silva, B.C., Oliveira, M.F., Moraes, H.R (2002) Groundwork for the development of the Brazilian Portuguese Wordnet. In: E.M. Ranchhod; N.J. Mamede (Eds.) *Advances in natural language processing*. Berlin: SpringerVerlag, p.189-196.

_____ (2003) Reusability of Dictionaries in the Compilation of NLP Lexicons In: N.J. Mamede; J. Baptista; I. Trancoso; M.G.V. Nunes (Eds.) *Computational processing of the portuguese language*. Berlin : Springer-Verlag, 2003, p.78-85.

Dias-da-Silva, B.C.; Rocha, M.A.E; Nunes, M.G.V. (2004) Projeto Montagem da Base Wordnet para o Português do Brasil-Processo CNPq 552057/01-0. *Relatório Técnico*. Araraquara: FCL-Unesp, 52p.

Eurowordnet (2004) *EuroWordNet*. Disponível em http://www.illc.uva.nl/EuroWordNet/data/sampleData.html. Acesso em: 10 ago. 2004.

Fellbaum, C. (Ed.) (1999) *WordNet: an electronic lexical database*. 2. Ed. Cambridge (Mass.): MIT Press.

Ferreira, A.B. de H. (1999) *Dicionário Aurélio eletrônico século XXI*. (Versão 3.0). São Paulo: LexiKon Informática Ltda.

GWA (2004) *Global Wordnet Association*. Disponível em http://www.globalwordnet.org. Acesso em: 10 ago. 2004.

Harabaglu, S.; Moldovan, D.I. (2000) Enriching the wordnet taxonomy with contextual knowledge acquired from text. In: Iwanska, L.; Shapiro, S.C.. *Natural language processing and knowledge representation*. Menlo Park, Ca; Cambridge, Mass: AAA Press; The Mit Press, p.301-333.

Handke, J. (1995) *The structure of the lexicon: human versus machine*. Berlin: Mouton de Gruyter.

Houaiss, A. (2001) *Dicionário eletrônico Houaiss da língua portuguesa*. (Versão 1.0). Rio de Janeiro: FL Gama Design Ltda.

Ilari, R.; Geraldi, J. W. (2000) *Semântica*. São Paulo: Editora Ática.

Iwanska, L.; Shapiro, S.C. (2000) *Natural language processing and knowledge representation*. Menlo Park, Ca; Cambridge, Mass: AAA Press; The Mit Press.

Lyons, J. (1977) *Semantics*. Cambridge: Cambridge University Press.

Marrafa, P. (2001) *WordNet do Português: uma base de dados de conhecimento lingüístico*. Lisboa: Instituto Camões.

Miller, G. A.; Fellbaum, C. (1991) Semantic networks of English. *Cognition*, Amsterdam, v. 41., n.1-3, p. 197-229.

Saint-Dizier, P., Viegas, E. (Eds.) (1995) *Computational Lexical Semantics*. Cambridge: Cambridge University Press, 1995.

Vossen, P. (1998) EuroWordNet: a multilingual database with semantic networks. *Computers and the Humanities*, Dordrecht, v. 32., n.2 e 3.

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

Wordnet (2004) *WordNet de Princeton*. Disponível em http://www.cogsci.princeton.edu/cgi-bin/webwn. Acesso em: 10 ago. 2004.