Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Herança automática das relações de hiperonímia para a WordNet.Br

Carolina Evaristo Scarton
Sandra Maria Aluísio

NILC-TR-09-10
Dezembro, 2009

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Uma wordnet pode ser entendida como uma base de dados que sistematiza o conjunto dos verbos, substantivos, adjetivos e advérbios de um dado idioma em termos de uma rede de quatro relações: sinonímia, antonímia, hiponímia/hiperonímia e meronímia/holonímia [Cruse, 1986]. A Wordnet do Brasil (Wordnet.Br) é um projeto que começou em 2003 ([Dias-da Silva et al., 2002] e [Dias-da Silva, 2003]). Esta wordnet é construída a partir do alinhamento com a Wordnet de Princeton (Wordnet.Pr). A partir desta construção alinhada, é possível herdar as relações da Wordnet.Pr automaticamente para a Wordnet.Br. Atualmente, a Wordnet.Br finalizou o alinhamento de verbos. Neste relatório técnico, descrevemos a criação de uma ferramenta que herda automaticamente a relação de hiperonímia para os verbos da Wordnet.Br. Mais do que isso, apresentamos a construção de uma base de dados relacional para a Wordnet.Br que servirá como base para outros trabalhos futuros, tais como, o armazenamento de outros conjuntos alinhados (substantivos, adjetivos e advérbios) e a herança automática da relação de hiponímia.

Palavras-chave: wordnet, Wordnet.Br, relações semânticas, hiperonímia

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Dentre as atividades compreendidas pela área de Processamento de Língua Natural (PLN) se encontram a criação e disponibilização de recursos léxicos e gramaticais. Estes recursos são fundamentais para trabalhos de diversas áreas. A construção manual destes recursos é praticamente inviável dado a grande carga de trabalho e a quantidade de tempo necessária para realização destas tarefas. Por isso, há um grande esforço para a criação destes recursos apoiados em técnicas computacionais.

1.1 Contextualização, Motivação e Domínio de Aplicação

A principal motivação deste projeto é o trabalho de Iniciação Científica (IC) da aluna em questão com o título *Avaliação da Inteligibilidade de Textos em Português: uma aplicação na área de simplificação de textos para o público infantil* financiado pela Fapesp (Processo: 2008/54282-9). Este projeto de IC[1] consiste da criação de uma ferramenta web que avalia a inteligibilidade de textos de acordo com métricas psicolinguísticas [Scarton et al., 2009]. Esta ferramenta (batizada de Coh-Metrix-PORT[2]) é baseada na ferramenta Coh-Metrix [McNamara et al., 2002] que foi desenvolvida por pesquisadores da Universidade de Memphis para avaliar a inteligibilidade de textos em inglês.

Os fatores importantes em uma análise do processo de compreensão de um texto são, segundo [Leffa, 1996]: o texto, o leitor e as circunstâncias em que se dá o encontro. Entre os fatores relativos ao texto destacam-se a legibilidade (apresentação gráfica do texto) e a inteligi-

[1] O projeto de IC está inserido no escopo do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), aprovado no âmbito do Edital Microsoft-Fapesp (proc. nro. 2007/54565-8)

[2] Para acessar o Coh-Metrix-PORT: http://caravelas.icmc.usp.br/coh

bilidade, ou seja, capacidade de um texto de ser lido, que está relacionada ao uso de palavras freqüentes para atingir uma gama maior de leitores, e estruturas sintáticas menos complexas para facilitar a sua leitura [Max, 2006]. É bem sabido que sentenças longas, com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença, além do uso de palavras de baixa frequência aumentam a complexidade de um texto para leitores com problemas de leitura como, por exemplo, analfabetos funcionais, afásicos e dislexos [Klebanov et al., 2004] e [Siddharthan, 2002]. Para o projeto PorSimples é muito importante analisar a inteligibilidade de textos, pois esta análise poderá servir como um ”termômetro”para ser usado na ferramenta SIMPLIFICA, por exemplo, exibindo ao autor do texto, métricas que indicam o quão inteligível está o texto.

Dentre as métricas utilizadas na versão livre do Coh-Metrix (chamada de Coh-Metrix 2.0[3]), já implementamos para a versão inicial do Coh-Metrix-PORT 34 métricas. Estas métricas são relativamente simples dado a falta de recursos linguísticos existentes para o português do Brasil. No processo de avaliação da inteligibilidade textual, muitas variáveis são levadas em consideração e uma delas é o quão abstrata é uma palavra. Uma maneira de medir o grau de abstração de uma palavra é através dos níveis de hiperonímia que esta palavra possui. Quanto mais níveis de hiperonímia uma palavra possui mais concreta é esta palavra[4]. Para a próxima versão do Coh-Metrix-PORT, entre outras coisas, pretendemos implementar uma métrica que seja capaz de avaliar a concretude de uma palavra utilizando os níveis de hiperonímia[5] desta palavra.

Para a identificação dos níveis de hiperonímia que uma palavra possui, o Coh-Metrix utiliza a Wordnet de Princeton [Miller et al., 1990] (aqui chamada de Wordnet.Pr) que é uma base lexical cujo design é inspirado em teorias psicolinguísticas da representação lexical. Como precisamos de uma wordnet para o português do Brasil que possua a relação semântica de hiperonímia para adaptar a métrica que conta hiperônimos de palavras, a herança automática da relação de hiperonímia para a Wordnet.Br é um trabalho fundamental para o projeto de IC em

[3]Para acessar o Coh-Metrix 2.0: http://cohmetrix.memphis.edu/cohmetrixpr/index.html

[4]Informação retirada da documentação do Coh-Metrix 2.0. Para acessar: http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm

[5]Hiperonímia é uma relação semântica que relaciona dois conjuntos com a relação *supertipo de*. Mais detalhes sobre relações semânticas serão apresentados na Seção 2.2 do Capítulo 2

questão.

Uma wordnet se diferencia de um dicionário comum, basicamente, porque é dividida em quatro classes: substantivos, verbos, adjetivos e advérbios. Esta divisão se faz necessária pois este recurso trabalha com o conceito de *synset* (synonym set) que são conjuntos de palavras sinônimas. São estes *synsets* que serão relacionados através de relações semânticas (este conceito será melhor explicado na Seção 2.2 do Capítulo 2). Assim sendo, uma palavra de uma determinada classe não pode estar relacionada a uma palavra de uma classe distinta, ou seja, um verbo não pode estar relacionado com um substantivo, um substantivo não pode estar relacionado com um adjetivo e assim por diante.

Após o trabalho da Wordnet.Pr, vários pesquisadores iniciaram trabalhos para criar wordnets para outros idiomas. É o caso da EuroWordNet [Vossen, 2002] e da MultiWordNet [Bentivogli et al., 2002], iniciativas para a criação de bases multilinguais que integram wordnets de diversos idiomas. Para o português do Brasil, o professor doutor Bento Carlos Dias da Silva, da UNESP de Araraquara, e pesquisador do NILC (Núcleo Interinstitucional de Linguística Computacional) iniciou um projeto para criação da Wordnet.Br através da Wordnet.Pr ([Dias-da Silva et al., 2002], [Dias-da Silva, 2003], [Dias-da Silva, 2005] e [Dias-da Silva et al., 2008]).

A construção da base de relações da WordNet.Br é feita por meio do alinhamento com a WordNet.Pr [Fellbaum, 1998]. O linguista começa o procedimento selecionando uma palavra do português. Então é realizada uma busca em um dicionário bilíngue (Português do Brasil - Inglês) e a palavra selecionada é relacionada com sua versão em inglês. Assim relações de hiperonímia podem ser herdadas automaticamente. Por exemplo, na WordNet.Pr consta que *try* é hiperônimo de *risk*, no procedimento descrito anteriormente *try* é relacionado com ”tentar”e *risk* com ”arriscar”, de modo que na WordNet.Br constará ”tentar”como hiperônimo de ”arriscar”. Este processo é representado na Figura 1.1.

Atualmente, este projeto já concluiu a etapa de alinhamento de verbos faltando a validação destes alinhamentos e a herança automática de relações semânticas da Wordnet.Pr. Para a validação dos alinhamentos é necessário que uma ferramenta computacional de edição auxilie

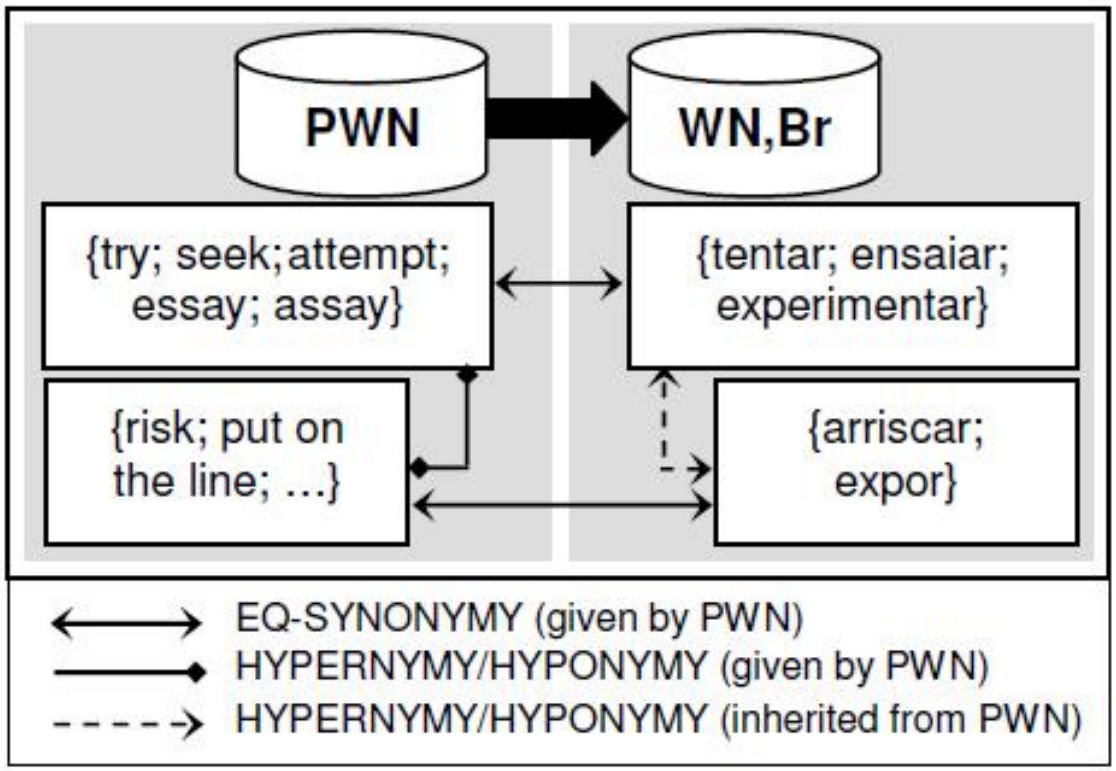

Figura 1.1: Exemplo de herança automática da relação de hiperonímia para a Wordnet.Br através da Wordnet.Pr [Di Felipo and Dias-da Silva, 2007]

o linguista identificar, automaticamente, possíveis inconsistências na base. Já para a herança automática das relações semânticas, pretendemos identificar as relações existentes nos *synsets* da Wordnet.Pr e estendê-las para os *synsets* da Wordnet.Br com eles relacionados e é esta tarefa que será abordada neste relatório técnico. O trabalho de [Dias-da Silva, 2003] será melhor explicado na Seção 2.4 do Capítulo 2.

1.2 Objetivos do Trabalho

Este projeto visa estender o trabalho de [Dias-da Silva et al., 2008] herdando automaticamente as relações semânticas presentes na Wordnet.Pr para a Wordnet.Br. Mais especificamente, este trabalho se concentra na herança automática da relação semântica de hiperonímia para verbos.

Outro ponto que merece ser citado é que, com este trabalho, será possível dar continuidade ao trabalho de [Scarton et al., 2009] que consiste da criação/adaptação de métricas que avaliam

inteligibilidade de textos.

Além disso, a disponibilização de uma wordnet para o português do Brasil, mesmo que em fase inicial, suprirá uma grande lacuna existente tanto na área de Linguística quanto na área de PLN.

1.3 Organização do Relatório Técnico

Este relatório técnico está organizado da seguinte forma: no Capítulo 2 é feita uma revisão bibliográfica relacionando este projeto com os principais trabalhos da área e também explicitando alguns conceitos básicos de linguística para o entendimento deste trabalho. As atividades realizadas, os resultados obtidos, as dificuldades e as limitações se encontram no Capítulo 3. Por fim, o Capítulo 4 finaliza este relatório técnico com uma análise geral deste trabalho e com uma discussão sobre trabalhos futuros.

2 *Revisão Bibliográfica*

2.1 Considerações Iniciais

Neste capítulo é apresentada a revisão bibliográfica sobre a Wordnet.Br e as demais wordnets e também aqui se encontram os conceitos necessários para o entendimento deste trabalho. Na Seção 2.2 é apresentada uma explicação sobre os conceitos básicos para o entendimento deste projeto. Na Seção 2.3 são apresentados os trabalhos relacionados com este projeto. O trabalho da Wordnet.Br é apresentado na Seção 2.4. Já na Seção 2.5 é feita uma análise crítica sobre os trabalhos relacionados e sobre o trabalho da Wordnet.Br.

2.2 Conceitos necessários

Linguística Computacional (LC) e Processamento de Língua (ou Linguagem) Natural (PLN) são denominações de áreas de pesquisa hoje usadas indistintamente. Todas remetem a uma área interdisciplinar que envolve a Ciência da Computação e a Linguística, e que trata da modelagem computacional (de aspectos) das línguas naturais especialmente a língua escrita visando seu tratamento em programas de computador. Até há algum tempo era comum se fazer a distinção entre LC e PLN, porém até hoje é difícil concluir o que as tornava distintas. Uma visão bastante aceita até então é que enquanto na LC se modelam computacionalmente teorias linguísticas conhecidas, em PLN não há compromisso com teorias linguísticas, podendo ser usados modelos empíricos, como os estatísticos ou baseados em regras. Em outras palavras, o que as distinguiria seria uma perspectiva mais linguística em LC e mais computacional em PLN. Por isso, para nós cientistas da computação, muitas vezes se faz necessária uma definição formal da parte linguística desta área.

Neste trabalho abordamos relações semânticas entre *synsets* (conceitos). Os conceitos ou *synsets* (*synonym sets*) são, basicamente, um conjunto de palavras que possuem o mesmo significado (sinônimos). Em outras palavras um *synset* denota um único significado. Por exemplo, o *synset* {esconder, fingir} composto pelas palavras *esconder* e *fingir* possui o sentido de *"fazer acreditar"*. O sentido que descreve o *synset* é chamado de *glosa*.

Relações semânticas relacionam dois ou mais *synsets* de acordo com uma característica semântica. Dentre todas as relações semânticas existentes, destacamos [Snow and Patel, 2008] e [Snow et al., 2005]:

- *sinonímia*: mesmo significado (X é sinônimo de Y, se X possui o mesmo significado de Y).

  Exemplo: *esconder* é **sinônimo** de *fingir* com o sentido de *"fazer acreditar"*.

- *antonímia*: significado oposto (X é antônimo de Y, se X possui significado oposto a Y).

  Exemplo: *descontrolar* (com o sentido de *"manejar mal ou incompetentemente"*) é **antônimo** de *controlar*.

- *hiperonímia*: é um supertipo de (X é hiperônimo de Y, se X é um supertipo de Y).

  Exemplo: *animal* é **hiperônimo** de *cachorro*.

- *hiponímia*: é um subtipo de (X é hipônimo de Y, se X é um subtipo de Y).

  Exemplo: *vermelho* é **hipônimo** de *cor*.

- *holonímia*: é o todo de (X é holônimo de Y, se X é o todo que inclui Y).

  Exemplo: *carro* é **holônimo** de *roda*.

- *meronímia*: é uma parte de (X é merônimo de Y, se X é parte de Y).

  Exemplo: *dedo* é **merônimo** de *mão*.

Em geral uma wordnet possui uma parte das relações semânticas descritas anteriormente ou todas elas. A wordnet mais completa é a de Princeton que abordaremos na Seção 2.3.1.

Na Figura 2.1 (retirada do trabalho de [Vossen, 2002]) há um exemplo das relações envolvendo o *synset* {car; auto; automobile; machine; motorcar} da Wordnet.Pr (versão 1.5). Um *synset* hiperônimo é {motor vehicle; automotive vehicle}. Já como hipônimos temos {cruiser; squad car; patrol car; police car; prowl car} e {cab; taxi; hack; taxicab}. Por fim, os merônimos são {bumper}, {car door}, {car mirror} e {car window}.

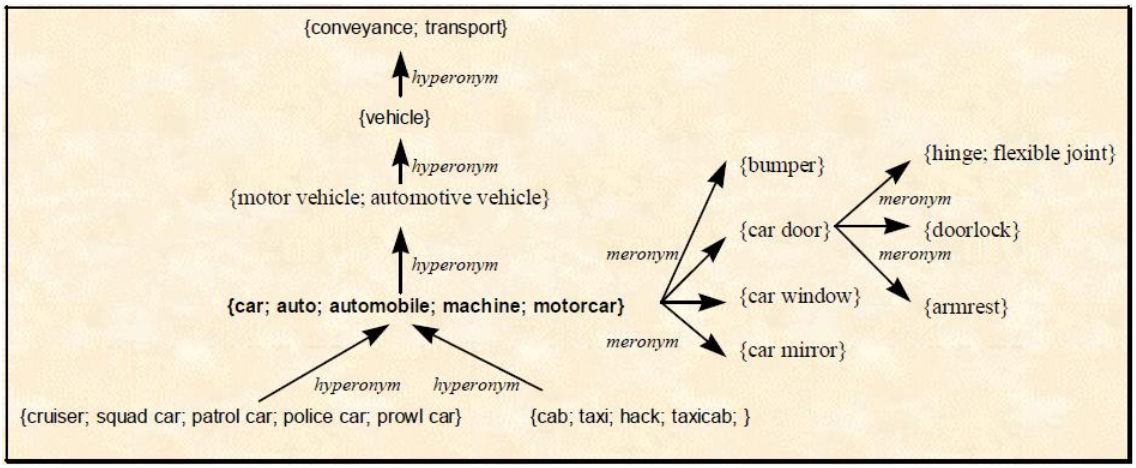

Figura 2.1: Synset da palavra *carro* em seu primeiro sentido na versão 1.5 da Wordnet.Pr [Vossen, 2002]

No caso de nosso trabalho, visamos extrair, automaticamente, a relação semântica de *hiperonímia* para os verbos da Wordnet.Br e, por isso, ilustramos na Figura 2.2 a relação de hiperonímia do verbo *paralize* extraída da Wordnet.Pr. Na Wordnet.Br, o verbo ”paralizar”apresenta a relação de hiperonímia apresentada na Figura 2.3.

2.3 Trabalhos Relacionados

Uma wordnet pode ser entendida como uma base de dados que sistematiza o conjunto dos verbos, substantivos, adjetivos e advérbios de um dado idioma em termos de uma rede de quatro relações: sinonímia, antonímia, hiponímia/hiperonímia e meronímia/holonímia [Cruse, 1986]. Nesta seção são apresentados os trabalhos com wordnet para diversos idiomas. Na Seção 2.3.1 é apresentada a Wordnet.Pr, que se trata da wordnet para o inglês. A EuroWordNet é apresentada na Seção 2.3.2. O trabalho da MultiWordNet é apresentado na Seção 2.3.3. Por fim, outros trabalhos relacionados são apresentados na Seção 2.3.4.

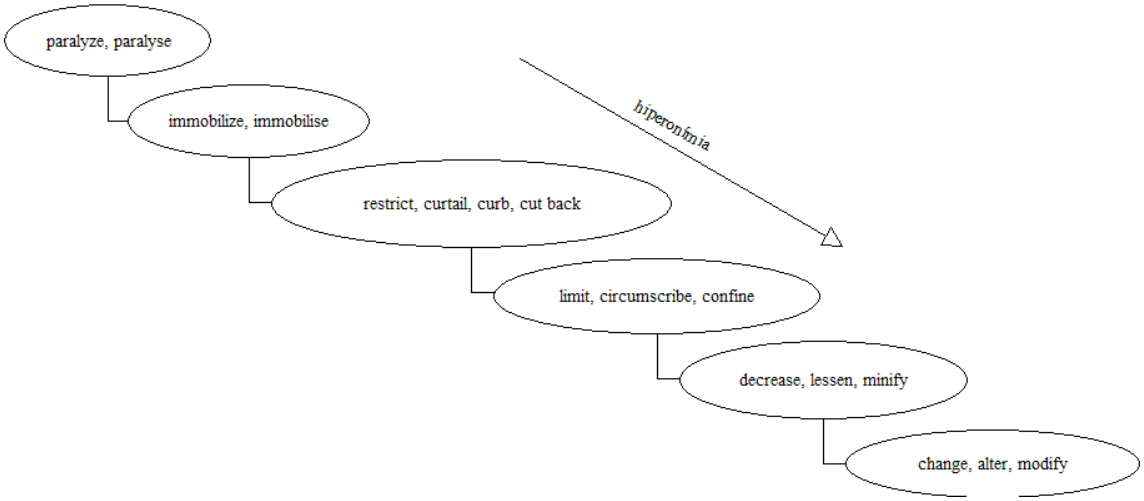

Figura 2.2: Exemplo de Relação de Hiperonímia para o verbo *paralize* na Wordnet.Pr

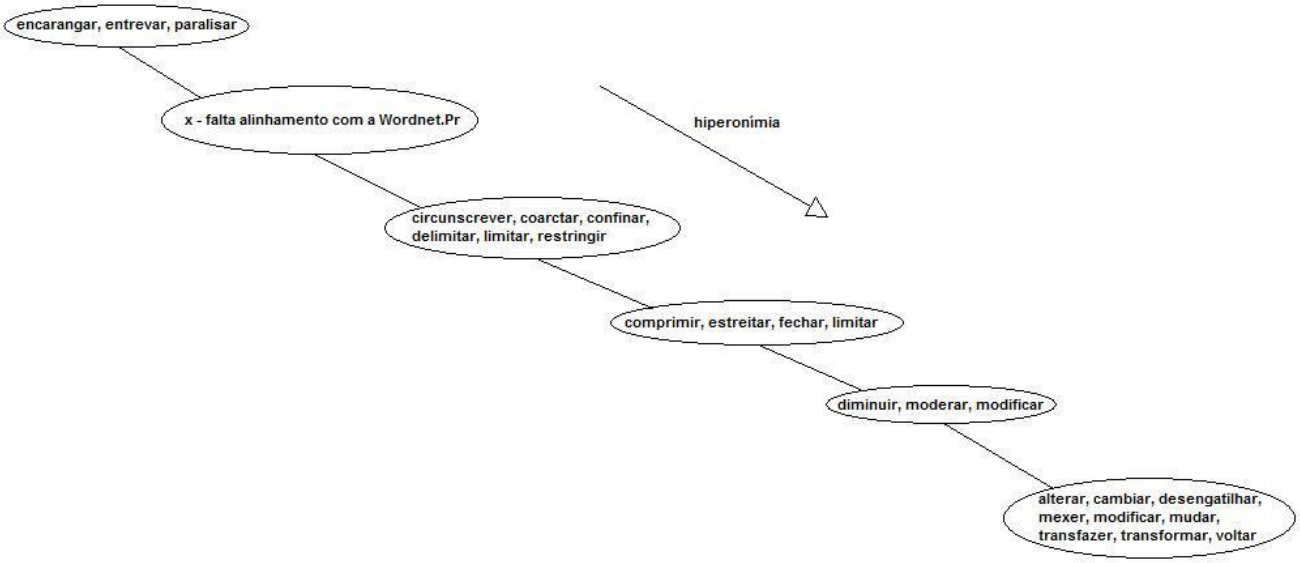

Figura 2.3: Exemplo de Relação de Hiperonímia para o verbo *paralizar* na Wordnet.Br

2.3.1 Wordnet.Pr

A Wordnet de Princeton (aqui chamada de *Wordnet.Pr*) é a pioneira entre as wordnets. Seu projeto foi uma proposta para uma combinação efetiva entre a tradicional informação lexicográfica e os avanços computacionais. A idéia inicial era fornecer uma ajuda para realizar buscas em dicionários de forma conceitual, e não, alfabética. Como esta tarefa se mostrou não trivial, nasceu a Wordnet.Pr [Miller et al., 1990].

Segundo seus criadores [Miller et al., 1990], a principal relação semântica da Wordnet.Pr

(o que deve se estender para qualquer wordnet) é a sinonímia, afinal é baseada nestas relações que os *synsets* são criados. Além disso, é por causa da sinonímia que a Wordnet.Pr (e também as demais wordnets) é dividida em: substantivos, verbos, adjetivos e advérbios. Como os conceitos são representados por *synsets* e os sinônimos devem ser substituíveis, então, palavras de diferentes classes sintáticas não poderiam ser sinônimas. Esta sinonímia que ocorre nas wordnets é chamada de sinonímia fraca, pois, uma expressão é sinônima de outra em um contexto X se a substituição de uma pela outra não altera o significado de X[1].

A Wordnet.Pr possui uma inteface web[2] e uma aplicação desktop[3]. A Figura 2.4 apresenta a tela de entrada da ferramenta online da Wordnet.Pr

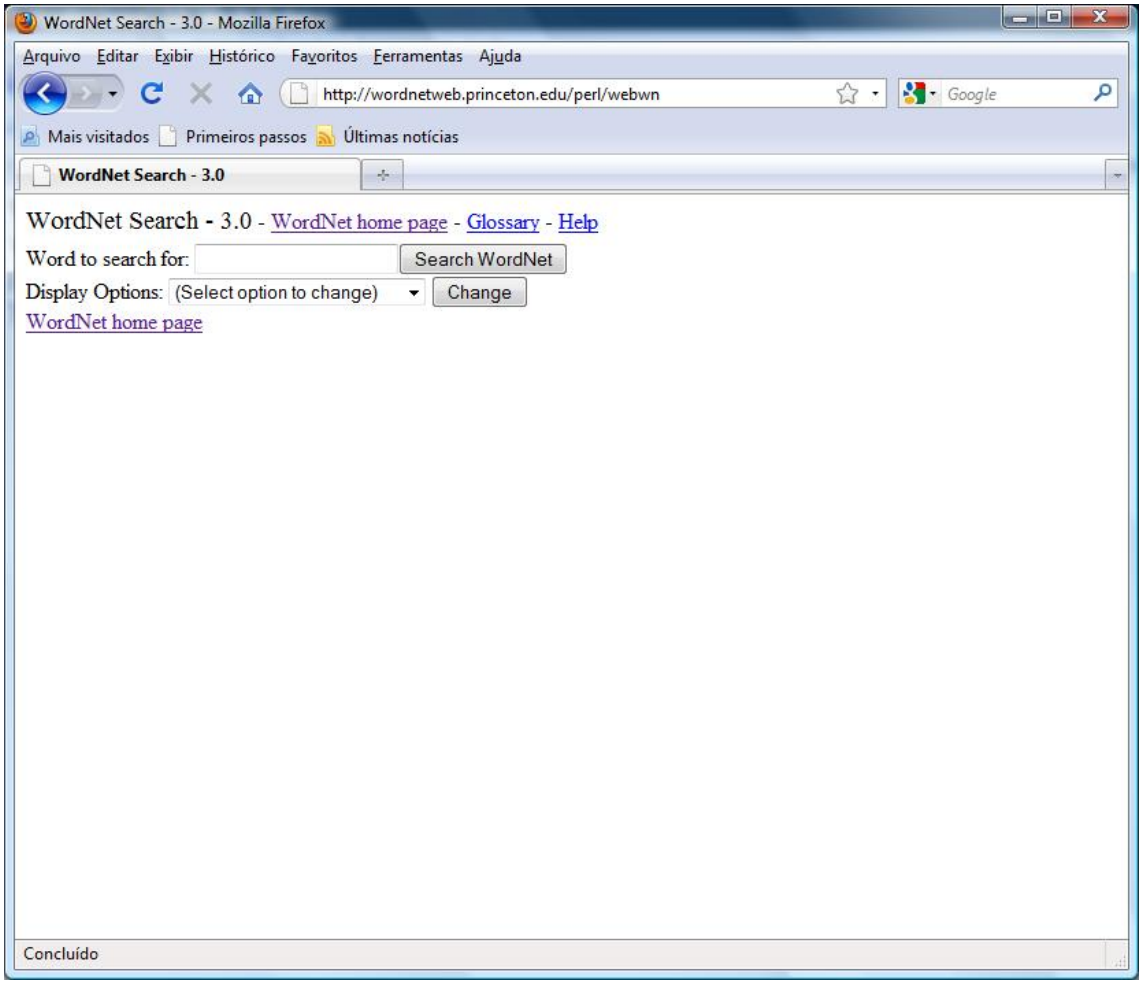

Figura 2.4: Tela de entrada da Wordnet.Pr (versão web)

[1]Mais informação pode ser encontrada em [Cruse, 1986], capítulo 12
[2]Para acessar a Wordnet.Pr: http://wordnetweb.princeton.edu/perl/webwn
[3]Para baixar a Wordnet.Pr: http://wordnet.princeton.edu/wordnet/download/

Atualmente está na versão 3.0 e, apesar de seu download e acesso a web estarem disponíveis a todos, sua estrutura interna e seu banco de dados não são liberados.

2.3.2 EuroWordNet

EuroWordNet é uma base de dados multilingual composta de várias wordnets de idiomas Europeus que é estruturada nos mesmos moldes da WordNet.Pr [Vossen, 2002]. Cada uma das wordnets da EuroWordNet é uma rede semelhante das relações entre os significados de palavras em um idioma específico. As relações semânticas são, portanto, consideradas relações internas do idioma. Além destas relações, cada *synset* é também ligado com o *synset* mais próximo da Wordnet.Pr (versão 1.5).

Para armazenar as várias wordnets em uma única base de dados multilingual, os *synsets* da Wordnet.Pr funcionam como um índice inter-lingual (ILI) [Vossen, 2004]. Assim é possível acessar um *synset* de uma wordnet através do *synset* de outra wordnet, desde que ambos *synsets* estejam ligados ao mesmo conceito da Wordnet.Pr.

Uma base de dados multilingual é útil para recuperação de informação entre os idiomas, transferência de informação de uma wordnet para outra e a comparação de diferentes wordnets. A arquitetura do banco de dados da EuroWordNet permite manter as características específicas de cada idioma e também manter a relação entre as diversas wordnets através do índice da Wordnet.Pr. Primeiramente a EuroWordNet compreendeu quatro wordnets dos idiomas: holandês, italiano, espanhol e inglês. Como extensão do projeto, incluíram os idiomas: alemão, estoniano, francês e tcheco. Todas as wordnets são limitadas a substantivos e verbos [4].

2.3.3 MultiWordNet

MultiWordNet também é composta de várias wordnets [Bentivogli et al., 2002]. Porém, é diferente da EuroWordNet na metodologia adotada para sua elaboração. Enquanto a EuroWordNet consiste de construir wordnets independentes para, numa fase posterior, tentar encontrar

[4] Informações retiradas de um documento datado de 2004 [Vossen, 2004], nenhuma referência recente foi encontrada

correspondências entre elas, a MultiWordNet consiste em construir wordnets de linguagens distintas mantendo, na medida do possível, as relações semânticas disponíveis na Wordnet.Pr. Isto é feito criando um novo *synset* correspondendo com um *synset* da Wordnet.Pr e herdando as relações semânticas que há entre os *synsets* da Wordnet.Pr para os novos *synsets*.

Outra vantagem importante deste modelo é que procedimentos automáticos podem ser concebidos para acelerar tanto a construção de *synsets* correspondentes quanto a detecção de divergências entre a Wordnet.Pr e a WordNet sendo construída. A MultiWordNet corresponde a wordnets de sete idiomas: inglês, português, italiano, espanhol, hebreu, latim e romeno [5].

2.3.4 Outros Trabalhos

Nesta subseção são apresentados outros trabalhos relacionados. Estes trabalhos se espelham nos trabalhos citados em 2.3.1, 2.3.2 e 2.3.3 e merecem ser citados dada a compatibilidade com nosso trabalho:

- *Russian Wordnet* Wordnet.Ru [Balkova et al., 2004]: é a wordnet para o idioma russo. Este projeto prevê o alinhamento da base da wordnet russa com a Wordnet.Pr, porém com o foco de criar uma base de dados multilingual Russo-Inglês e não pretendendo herdar relações da Wordnet.Pr automaticamente.

- *Automatic Building of Wordnets* [Barbu and Mititelu, 2007]: este trabalho descreve como criar wordnets a partir de outra wordnet já existente, automaticamente. A tarefa é dividida em duas fases: a primeira consiste da geração de *synsets* para a nova wordnet através do mapeamento da wordnet fonte, seguindo uma série de heurísticas. Na segunda fase, as relações semânticas que podem ser automaticamente herdadas da wordnet fonte são identificadas e, então, extraídas. Neste trabalho também é apresentado um caso de uso em que a wordnet do idioma Romeno foi gerada, automaticamente, a partir da Wordnet.Pr.

[5] Para acessar a MultiWordNet: http://multiwordnet.itc.it/english/home.php

2.4 Wordnet.Br

A citação retirada o trabalho de [Dias-da Silva, 2005] descreve sucintamente a estrutura de uma wordnet:

> ”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 identificação informal desse conceito. Por fim, para ilustrar o contexto de uso de uma unidade lexical, associa-se a ela uma frase-exemplo.”

Ainda segundo [Dias-da Silva, 2005], há três noções necessárias para o entendimento da estrutura de uma wordnet: a noção de sinonímia fraca, a noção de matriz lexical e as relações léxico-conceituais entre *synsets*. A primeira já descrevemos na Seção 2.3.1. A segunda possui dois fundamentos: a adoção do modelo relacional de representação do significado lexical e a noção de matriz lexical. A Figura 2.5 (retirada do trabalho de [Dias-da Silva, 2005]) apresenta uma matriz lexical em que na horizontal estão as formas lexicais (palavras) e na vertical estão os *synsets*. As células da matriz apresentam a relação, se houver, entre uma forma lexical e um *synset*.

A Wordnet.Br é a wordnet para o português do Brasil. A construção da base de relações da WordNet.Br [Dias-da Silva et al., 2008] e [Dias-da Silva, 2005] é feita por meio de um alinhamento com a WordNet.Pr [Fellbaum, 1998]. O linguista começa o procedimento selecionando uma palavra em português (inicialmente este processo foi feito para verbos). Então é realizada uma busca em um dicionário bilíngue (Português Brasil - Inglês) e a palavra selecionada é relacionada com sua versão em inglês. Assim como na MultiWordNet e na EuroWordNet, os verbos são relacionados utilizando o índice da Wordnet.Pr. Como exemplo para este processo, considere o verbo *arriscar*. Este é traduzido para *risk* e por fim é associado ao *synset* de índice ’02470374’ que compreende {risk, put on the line, lay on the line}.

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 2.5: Exemplo de matriz lexical que representa os conceitos lexicalizados pelas formas *carecer, demandar, necessitar, pedir, precisar, querer, reclamar, requerer e faltar* [Dias-da Silva, 2005]

A base da Wordnet.Br é separada em glosas e cada glosa corresponde a um arquivo no formato .doc (Microsoft Word). Como não havia um editor para esta tarefa, o ambiente do Microsoft Word foi o mais confortável para o trabalho dos linguistas. A Figura 2.6 apresenta um arquivo que contém uma glosa como exemplo.

Podemos observar na Figura 2.6 que a glosa em questão é a 'Glosa1A' cujo sentido é "derramar saliva da boca"e a chave é 'BABAR' (esta chave é escolhida pelo linguista e será a palavra que representará a glosa. Esta escolha é feita intuitivamente, procurando escolher a palavra mais conhecida dentre as palavras do *synset*). O ILI para esta glosa é '00100199' cujo *synset* na Wordnet.Pr é {drivel, drool, slabber, slaver, slobber, dribble}. Há, também, o sentido do *synset* da Wordnet.Pr (neste caso 'Sense 1'). Este sentido foi recuperado quando o linguista selecionou o verbo que na Wordnet.Pr que representava o conceito que ele estava procurando, ou seja, procurando a palavra *drivel* na base da Wordnet.Pr encontraremos o sentido 1 (Sense 1) como o sentido que se associa com a nossa 'Glosa1A'. Esse conceito de sentido (glosa) é necessário pois uma palavra pode pertencer a mais de um *synset* com sentidos diferentes. Por fim, no final da figura, há a hierarquia de hiperonímia da Wordnet.Pr para o *synset* de ILI

ANÁLISE DA GLOSA0001A

DADO O SYNSET:

S='Glosa1'
{ababalhar: x,
babar: Alguns doentes H]babam.,
escumar: O touro, bravio e enfurecido, M]escumava. ,
espumar: O cavalo está H]espumando.,}

A ANÁLISE RESULTA EM:

S='Glosa1A' = **"derramar saliva da boca"**
{babar: Anteontem, Bisol foi flagrado novamente num instante em que [babava paixão na camisa.,}

Chave: BABAR <verb.body>
ILI: 00100199

Sense 1
{00100199} <verb.body> drivel, drool, slabber, slaver, slobber, dribble -- (let saliva drivel from the mouth; "The baby drooled")
=> {00100016} <verb.body> salivate -- (produce saliva; "We salivated when he described the great meal")
=> {00009947} <verb.body> act involuntarily, act reflexively -- (act in an uncontrolled manner)
=> {00010141} <verb.body> act, behave, do9 -- (behave in a certain manner; show a certain behavior; conduct or comport oneself; "You should act like an adult"; "Don't behave like a fool"; "What makes her do this way?"; "The dog acts ferocious, but he is really afraid of people")

Figura 2.6: Exemplo de uma glosa da Wordnet.Br

'00100199'.

Outro atributo que pode aparecer em uma glosa da Wordnet.Br é a relação que há entre o *synset* da Wordnet.Br e o *synset* da Wordnet.Pr. No caso da Figura 2.6, não aparece nenhuma relação, então, esta glosa será tratada, a priori, com a relação **EQ_SYNONYM** que significa que há uma relação de sinonímia entre o *synset* da Wordnet.Br com o *synset* da Wordnet.Pr. Porém, como mostra a Figura 2.7, podem ocorrer situações em que um *synset* da Wordnet.Br não possui um *synset* na Wordnet.Pr que pode ser relacionado diretamente com uma relação de sinonímia. Por isso, algumas relações foram criadas para que o alinhamento entre as duas wordnets pudesse ser feito mesmo sem uma relação de sinonímia.

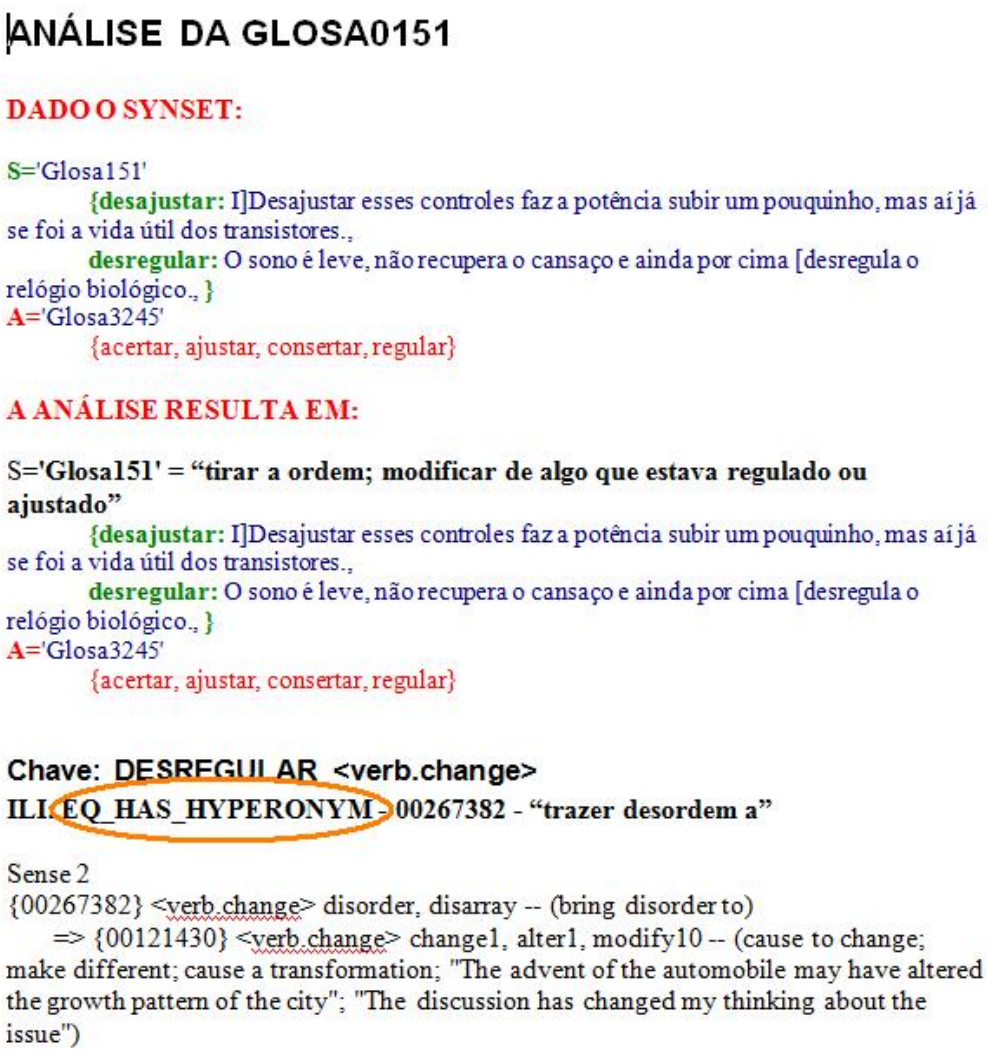

Figura 2.7: Exemplo de uma glosa da Wordnet.Br com a relação **EQ_HAS_HYPERONYM**

No caso da Figura 2.7 a relação que há entre o *synset* da Wordnet.Br desajustar, desregular e o *synset* da Wordnet.Pr {disorder, disarray} não é de sinonímia e sim de hiperonímia, ou seja, o *synset* {disorder, disarray} é hiperônimo do *synset* {desajustar, desregular}. Esta relação foi utilizada pois não existe *synset* na Wordnet.Pr que possua relação de sinonímia com o *synset* {desajustar, desregular}. A Tabela 2.1 apresenta as possíveis relações da Wordnet.Br com a Wordnet.Pr com uma breve descrição delas.

Segundo o criador da Wordnet.Br [Dias-da Silva, 2003], o processo de design e implementação de sistemas de PLN devem compreender três domínios: o domínio da

Tabela 2.1: Lista das possíveis relações entre a Wordnet.Br com a Wordnet.Pr

RELAÇÃO	DESCRIÇÃO
EQ_SYNONYM	Relação direta entre um *synset* da Wordnet.Br com um *synset* da Wordnet.Pr
EQ_NEAR_SYNONYM	Quando mais de um *synset* da Wordnet.Br está relacionado com um *synset* da Wordnet.Pr ou vice-versa
EQ_HAS_HYPERONYM	Quando um *synset* da Wordnet.Pr é hiperônimo de um *synset* da Wordnet.Br
EQ_HAS_HYPONYM	Quando um *synset* da Wordnet.Pr é hipônimo de um *synset* da Wordnet.Br

linguística (é feita a elicitação de informação linguística), o domínio representacional (é feita a identificação da informação linguística que pode ser tratada computacionalmente) e o domínio computacional (as informações linguísticas são tratadas e transformadas em recursos computacionais). No desenvolvimento da Wordnet.Br houve trabalho nos três domínios. Nos domínios representacional e computacional (que é onde nós atuamos) foram desenvolvidos, respectivamente, o projeto e implementação de um Banco de Dados e o desenvolvimento de um Editor para a Wordnet.Br [Dias-da Silva et al., 2008].

Este editor visa auxiliar o linguista na tarefa de alinhamento. Neste programa, o linguista seleciona uma palavra na lista do Editor. Então é realizada uma busca em um dicionário bilíngue on-line (Português do Brasil - Inglês) e a palavra selecionada é relacionada com sua versão em inglês. Por fim, através de links *drop down*, o linguista seleciona o *synset* da Wordnet.Pr que ele quer relacionar com o novo *synset* da Wordnet.Br. Na Figura 2.8 há uma visão geral do editor e na Figura 2.9 há um exemplo da tarefa realiza pelo linguista supondo que ele está trabalhando com o *synset* {atender, satisfazer, corresponder, cumprir, realizar}. As regiões delimitadas por retângulos são a parte que é herdada da Wordnet.Pr.

Porém, o desenvolvimento do editor nunca chegou a uma versão estável capaz de realizar esta tarefa. Além disso, este editor não utiliza um banco de dados relacional (salva os dados em um arquivo binário). Assim sendo, após um estudo detalhado, resolvemos não utilizar o editor para realizar a tarefa de herança automática da relação de hiperonímia. Além disso, decidimos criar uma base de dados em um banco de dados relacional para armazenar a Wordnet.Br e para facilitar a disponibilização deste recurso na web. No Capítulo 3 há mais detalhes sobre a decisão

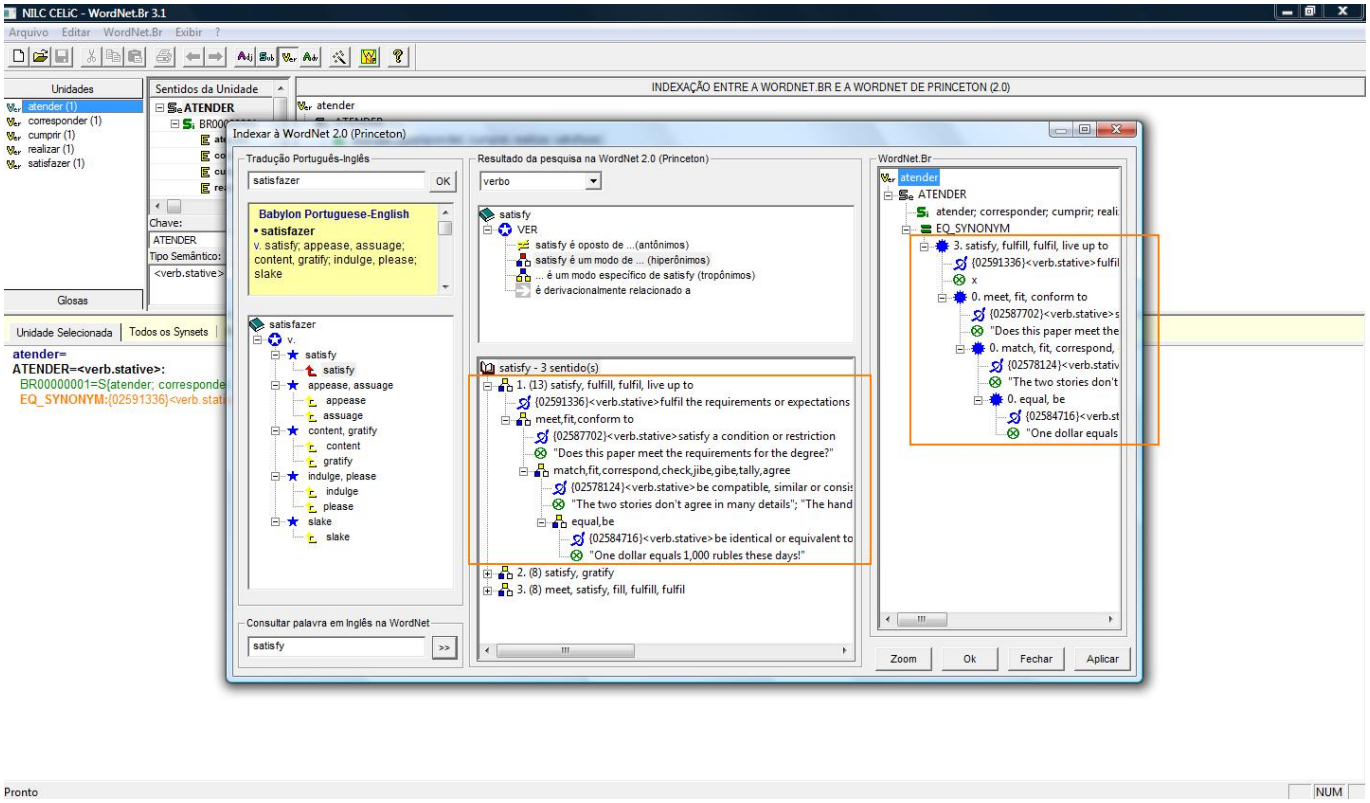

Figura 2.8: Visão geral do editor da Wordnet.Br

de não utilizar o editor e sobre o projeto da base de dados. Já um novo editor capaz de acessar o banco de dados relacional será deixado como trabalho futuro.

2.5 Análise Crítica e Discussão

O trabalho de [Miller et al., 1990] é, sem dúvida, o mais completo e de maior sucesso quando o assunto é wordnet. Tanto isso é verdade, que a construção da maioria das wordnets é feita através da Wordnet.Pr (ou pelo menos, o projeto é baseado neste trabalho). Além disso, Wordnet.Pr está em constante atualização e já está na versão 3.0.

A EuroWordNet [Vossen, 2002] foi a primeira iniciativa de construir wordnets através da Wordnet.Pr e, como mencionado na Seção 2.3.2, é composta de várias wordnets conectadas através do índice inter-lingual (ILI) da Wordnet.Pr. Este trabalho também é de grande importância pois foi a pioneira em utilizar a Wordnet.Pr como base para um novo recurso.

O trabalho de [Bentivogli et al., 2002], a MultiWordNet, provavelmente é o que mais se

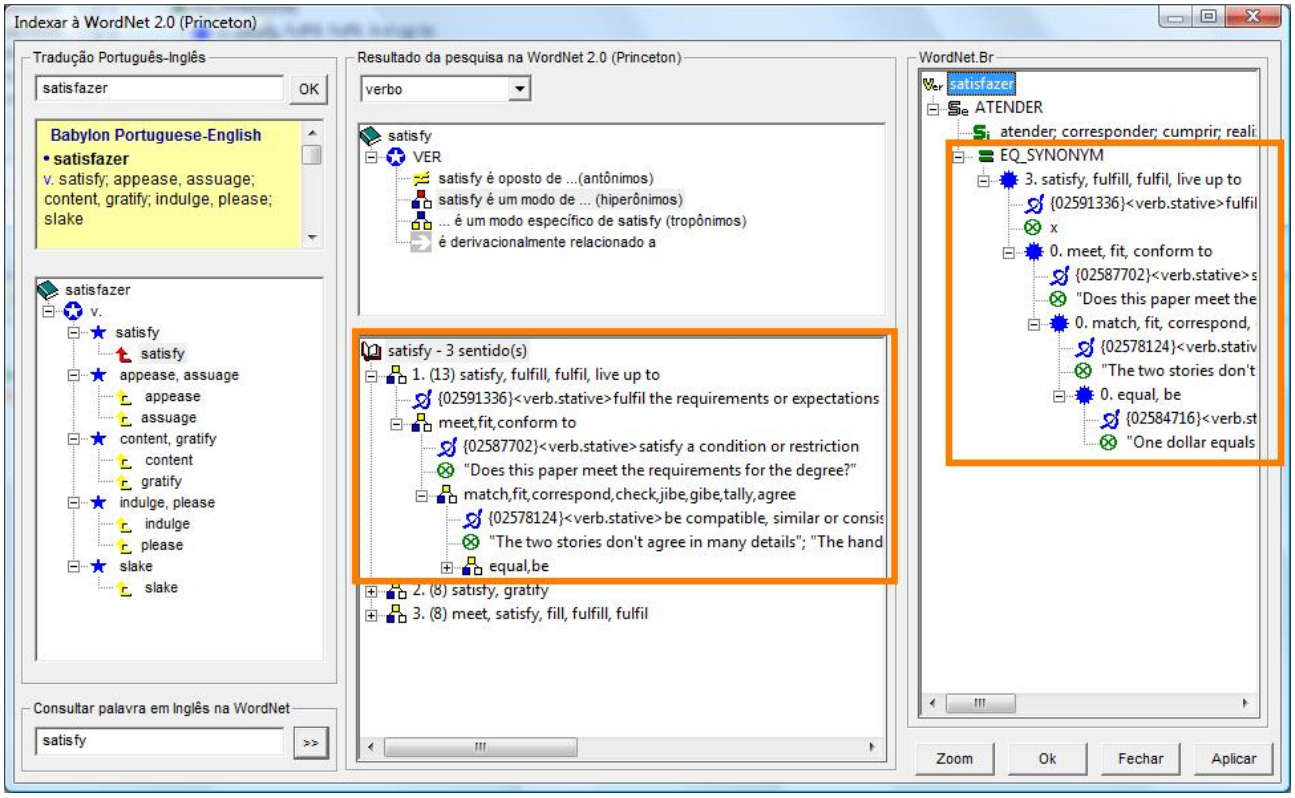

Figura 2.9: Exemplo do trabalho do linguista no editor da Wordnet.Br

aproxima do trabalho realizado para a Wordnet.Br. Afinal, este projeto visa a construção de wordnets independentes mantendo, sempre que possível, as relações semânticas da Wordnet.Pr. Ou seja, assim como na Wordnet.Br a MultiWordNet visa herdar relações da Wordnet.Pr.

A Wordnet.Ru [Balkova et al., 2004] também está diretamente relacionado com o trabalho da Wordnet.Br. Apesar de possui um foco diferente (a construção de uma base multilingual Russo-Inglês), o projeto da Wordnet.Ru compreende o alinhamento desta wordnet com a Wordnet.Pr.

Já o trabalho de [Barbu and Mititelu, 2007] é inovador e, provavelmente, de difícil implementação. Além disso, o uso de heurísticas pode acarretar erros grotescos na base da nova wordnet sendo que alguns provavelmente não possam ser identificados facilmente (dado a grande base de dados que uma wordnet representa). Porém, assumindo que mesmo o alinhamento feito por linguista possui uma margem de erro, trabalhos como este podem ser o futuro das wordnets.

Por fim, como foi apresentado na Seção 2.4, o trabalho já desenvolvido para a Wordnet.Br, na visão computacional, ainda é limitado. Porém, na parte linguística, o professor doutor Bento Carlos Dias da Silva mantém uma equipe que, atualmente, está realizando o alinhamento de substantivos para a Wordnet.Br. O maior envolvimento da área computacional com o projeto da Wordnet.Br é mais do que necessário e pretendemos, com este projeto, reforçar esses laços. Por isso, este projeto é o início de um grande trabalho que pretende disponibilizar a Wordnet.Br via web, além de construir um editor web para que os linguistas possam realizar a tarefa de alinhamento com o auxílio do computador.

2.6 Considerações Finais

Neste capítulo procurou-se fornecer uma visão geral dos principais conceitos deste trabalho, além da explicação de alguns conceitos da área de Linguística necessários para que cientistas da computação compreendam este trabalho. Também foi apresentada uma revisão dos principais trabalhos relacionados com este projeto, buscando explicitar a importância do mesmo e identificar a lacuna ao qual este trabalho pretende preencher.

3 *Desenvolvimento do Trabalho*

3.1 Considerações Iniciais

Neste capítulo é apresentado o projeto deste relatório técnico. Na Seção 3.2 são descritos os objetivos do trabalho e a metodologia utilizada para o desenvolvimento do mesmo. A Seção 3.3 apresenta, detalhadamente, as atividades realizadas durante este projeto. Já na Seção 3.4 os resultados obtidos são comparados com os resultados esperados e com o cronograma proposto. Na Seção 3.5 discutimos as dificuldades e limitações encontradas durante o desenvolvimento do projeto.

3.2 Projeto

O objetivo deste projeto é implementar a herança automática das relações de hiperonímia da Wordnet.Pr para a Wordnet.Br. Esta herança automática é possível pois a Wordnet.Br é alinhada com a Wordnet.Pr, ou seja, para cada glosa presente na Wordnet.Br há um ILI correspondente na Wordnet.Pr. Assim sendo, se o ILI X da Wordnet.Pr é hiperônimo do ILI Y e se o ILI X está alinhado com a glosa A da Wordnet.Br e o ILI Y está alinhado com a glosa B, podemos concluir que a glosa A é hiperônimo da glosa B.

Todos os dados (verbos) da Wordnet.Br estavam armazenados em cerca de 4.000 arquivos .doc (documento da Ferramenta Microsoft Word). Esta ferramenta foi utilizada pois, dada a inexistência de um editor que apoiasse o trabalho de alinhamento da Wordnet.Pr com a Wordnet.Br, os linguistas estavam mais familiarizados a utilizar o Microsoft Word. Dada a necessidade deste editor, iniciou-se um projeto no NILC (Núcleo Interinstitucional de Liguística Com-

putacional) que visava à construção de um editor capaz de auxiliar o linguista na construção dos alinhamentos e também capaz de carregar os arquivos .doc[1] já elaborados pelos linguistas [Dias-da Silva et al., 2008] (como mencionado no Capítulo 2 na Seção 2.4). Este editor é um programa desktop, desenvolvido em C++, utilizando o ambiente Microsoft Visual Studio e que utiliza como base de dados um arquivo binário gerado pelo próprio programa (.wnb). Esta decisão de projeto foi necessária, pois quando do início do projeto da Wordnet.Br, o Departamento de Letras da UNESP de Araraquara, ao qual o professor doutor Bento Carlos Dias da Silva (criador da Wordnet.Br) está vinculado, não possuía uma boa conexão com a internet capaz de suportar um sistema web que acessasse o servidor do NILC. Por isso, a idéia era salvar os dados da Wordnet.Br em vários arquivos binários e depois, através destes arquivos, gerar um código SQL que salvasse os dados em um banco de dados. Porém, este projeto não foi concluído e o editor nunca chegou a uma versão estável.

Inicialmente, pensamos em utilizar o editor da Wordnet.Br. Porém, após vários testes, observamos que este editor é muito instável e ultrapassado. Apesar de possuir uma interface intuitiva e amigável, não trabalha com banco de dados relacional, o arquivo binário corrompe com frequência, não carrega corretamente os arquivos textos (convertidos dos arquivos .doc) e, após certa quantidade de arquivos carregados, o programa do editor para de funcionar.

Analisamos então a possibilidade de corrigir todos os erros deste editor para, posteriormente, conseguirmos herdar automaticamente as relações de hiperonímia. Porém, dado o pouco tempo para desenvolvimento deste projeto e também visando uma nova abordagem na construção da Wordnet.Br, que pretende utilizar a Web como canal de edição e busca na mesma, decidimos abandonar o editor e partir para a construção de uma base de dados relacional com todos os arquivos textos da Wordnet.Br carregados. Justificamos, então, a decisão de abandonar o editor pelo fato dele não trabalhar com uma base dados relacional e não ser uma aplicação Web. Por fim, construímos uma interface de busca simples, somente para prova de conceito.

Portanto, como decidimos construir uma base de dados relacional, tivemos de escolher um banco de dados relacional. O banco de dados escolhido foi o MySQL[2] que trabalha muito bem

[1] Estes arquivos deveriam ser previamente convertidos para texto (.txt).
[2] Para acessar o site do MySQL: http://www.mysql.com/

com aplicações Web e também já foi estudado e utilizado pela aluna em questão em seu projeto de IC. Como linguagem para desenvolvimento escolhemos o Ruby[3] com o framework Rails[4]. A escolha desta tecnologia se justifica pela familiaridade da aluna com a mesma (assim como o MySQL, também foi utilizada no projeto de IC da aluna em questão), pela rapidez e facilidade de desenvolvimento de aplicações Web e porque a linguagem Ruby é muito boa para trabalhar com processamento de textos.

Assim sendo, construímos uma base de dados chamada *wordnet.br* composta das tabelas: *alinhamentos*, *exemplos*, *glosa*, *hiperonimos*, *ilis*, *palavras*, *synsets* e *wnbrs* (Figura 3.1). Já no nível da aplicação, criamos uma classe *controlador* que possui os seguintes métodos: *buscar*, *corrigir*, *hiperonimos*, *inserir* e *encontraHiper*. Os detalhes de implementação do banco de dados e da ferramenta, serão discutidos na Seção 3.3.

3.3 Descrição das Atividades Realizadas

Esta seção descreve detalhadamente as atividades realizadas para a execução deste projeto.

- **Estudo do editor da Wordnet.Br**

  O editor da Wordnet.Br foi a primeira ferramenta que tentou auxiliar o trabalho do linguista na construção da Wordnet do Brasil. Porém, nossa experiência com esta ferramenta nos mostrou que ela é muito instável e ultrapassada. Como já foi mencionado na Seção 3.2, o editor é uma aplicação Desktop, desenvolvida em C++, que salva os dados da Wordnet.Br em um arquivo binário (.wnb).

  Tentamos carregar no editor os 100 primeiros textos da Wordnet.Br. Além de não conseguir carregar todos os textos corretamente (em alguns dos casos as relações de hiperonímia da Wordnet.Pr não eram carregadas), a ferramenta parou de funcionar quando começamos a alterar os dados. Como se não bastasse, quando abríamos novamente o editor o arquivo binário em que os dados estavam salvos havia sido corrompido.

[3]Para acessar o site do Ruby: http://weblog.rubyonrails.org/
[4]Para acessar o site do Rails: http://api.rubyonrails.org/

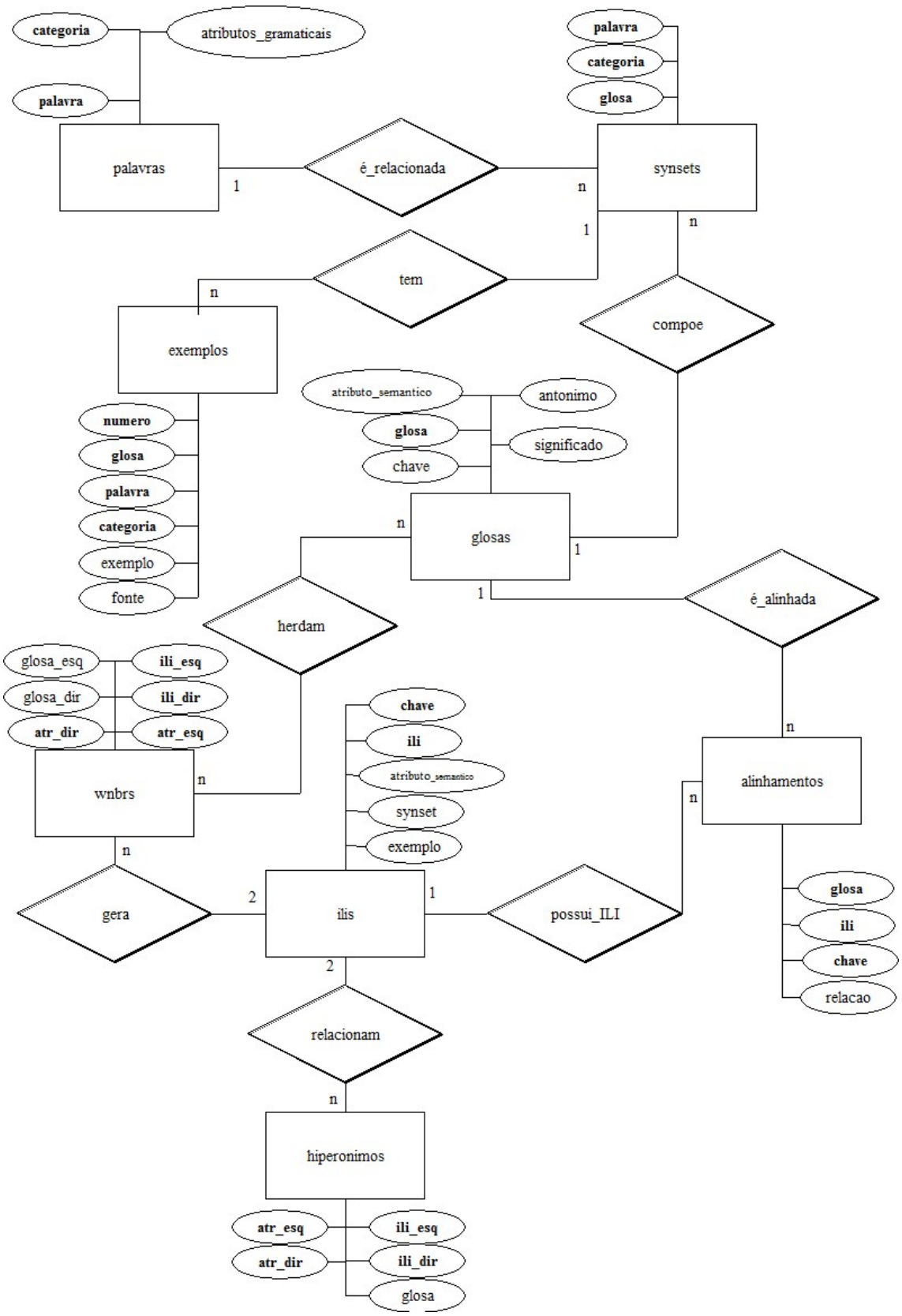

Figura 3.1: Diagrama Entidade-Relacionamento para o projeto da Wordnet.Br

Apesar de possuir uma opção para exportar os dados para a linguagem SQL, este editor só conseguiria fazer esta exportação com sucesso se todos os arquivos estivessem carregados, pois, para cada arquivo binário .wnb, o editor repetia o processo de geração de chaves primárias e, portanto, dava as mesmas chaves que já tinha utilizado para outros arquivos.

Por tudo isso, decidimos abandonar o editor e adotar uma abordagem diferente, visando a criação de uma base da dados relacional já pensando em um recurso via Web para edição e consulta da base da Wordnet.Br.

- **Estudo da literatura da Wordnet.Br**

  Para a familiarização da aluna em questão com a Wordnet.Br a leitura de artigos publicados sobre este recurso foi obrigatória.

  Assim sendo, todas as publicações feitas sobre a estrutura, criação e estado atual da Wordnet.Br foram estudadas pela aluna. Além disso, reuniões com o desenvolvedor do editor, o senhor Ricardo Hasegawa, e com o criador da Wordnet.Br, o professor doutor Bento Carlos Dias da Silva, foram realizadas para introduzir a aluna no projeto da Wordnet.Br.

  Os artigos lidos para estudo foram:

  - *Groundwork for the Development of the Brazilian Portuguese Wordnet* - [Dias-da Silva et al., 2002]
  - *Human language technology research and the development of the Brazilian Portuguese wordnet* - [Dias-da Silva, 2003]
  - *A construção da base da Wordnet.Br: conquistas e desafios* - [Dias-da Silva, 2005]
  - *Towards an Automatic Strategy for Acquiring the WordNert.Br Hierarchical Relations* - [Di Felipo and Dias-da Silva, 2007]
  - *The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database* - [Dias-da Silva et al., 2008]

- **Leitura de artigos científicos e trabalhos relacionados ao tema**

Para o desenvolvimento do projeto foram estudados vários artigos científicos sobre wordnets, construção de wordnets e relações semânticas. Os principais artigos lidos foram:

- *Extração de relações semânticas via análise de correlação de termos em documentos* - [Botero and Ricarte, 2009]
- *Lexical Semantics* - [Cruse, 1986]
- *Introduction to WordNet: an On-line Lexical Database* - [Miller et al., 1990]
- *WordNet: An Electornic Lexical Database* - [Fellbaum, 1998]
- *EuroWordNet Project* - [Vossen, 2002]
- *EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index* - [Vossen, 2004]
- *Coping With Lexical Gaps When Building Aligned Multilingual Wordnets* - [Bentivogli et al., 2000]
- *MultiWordNet: developing an aligned multilingual database* - [Bentivogli et al., 2002]
- *Russian WordNet* - [Balkova et al., 2004]
- *Automatic Building of Wordnets* - [Barbu and Mititelu, 2007]

Todas essas leituras foram fundamentais para o entendimento do tema do trabalho e execução do projeto.

- **Análise e correção dos arquivos da Wordnet.Br**

  Como as glosas da Wordnet.Br foram feitas em arquivos .doc (documentos do Microsoft Word), erros de digitação foram muito frequentes nos 4.128 arquivos. Para os scripts que extraiam os dados e os salvavam no banco de dados alguns desses erros, como por exemplo, a falta de uma vírgula, não afetavam sua execução. Porém, erros como a ausência da chave da glosa, afetavam diretamente a execução do script e precisaram ser corrigidos manualmente.

Por isso, os scripts da ferramenta que alimentou a base de dados da Wordnet.Br exibiam a glosa que estavam trabalhando naquele momento e, por isso, foi possível identificar facilmente onde se localizavam os erros.

- **Elaboração da base de dados da Wordnet.Br**

  A base de dados da Wordnet.Br, cujo modelo entidade-relacionamento aparece na Figura 3.1, foi desenvolvida buscando facilitar ao máximo a herança das relações de hiperonímia da Wordnet.Pr para a Wordnet.Br. Assim sendo, foram criadas 8 tabelas que armazenam dados da Wordnet.Br e dados do alinhamento com a Wordnet.Pr.

  Descrição detalhada de cada tabela

  A tabela *palavras* possui 3 atributos: *palavra*, *categoria* e *atributos_gramaticais*. O primeiro atributo é uma palavra da Wordnet.Br. O segundo é a categoria da palavra (1 - substantivos, 2 - adjetivos, 3 - verbos e 4 - advérbios) sendo que, dado o estado atual da Wordnet.Br, a única categoria utilizada é 3 - verbos. Por fim, o terceiro atributo é deixado para informações adicionais que podem ser associadas a essa palavra, tais como: número (singular, plural) e gênero (masculino, feminino). Como uma palavra pode pertencer a duas categorias, a chave primária desta tabela é uma composição do atributo palavra e do atributo categoria.

  A tabela *glosas* possui 6 atributos: *glosa*, *chave*, *significado*, *atributo_semantico*, *antonimo* e *synset*. O primeiro atributo é a identificação da glosa que é composta de 'Glosa' + 'número' em que este número é pré-definido pelo linguista (este atributo é a chave primária). O segundo é uma palavra que melhor representa a glosa. O terceiro é uma frase que define o significado da glosa. O quarto é o atributo semântico desta glosa (que segue o padrão da Wordnet.Pr e são apresentados na Tabela 3.1). O quinto é a glosa que tem significado antônimo ao da glosa em questão. Por fim, o sexto atributo é uma cadeia de palavras que corresponde ao *synset* da glosa.

  A tabela *exemplos* possui 6 atributos: *numero*, *glosa*, *palavra*, *categoria*, *exemplo*

Tabela 3.1: Lista dos atributos semânticos encontrados na Wordnet.Br

Atributos Semânticos
verb.body
verb.change
verb.cognition
verb.communication
verb.consumption
verb.competition
verb.contact
verb.creation
verb.emotion
verb.motion
verb.perception
verb.possession
verb.social
verb.stative
verb.weather

e *fonte*. O primeiro atributo é um número para controle dos exemplos por palavras, pois uma palavra pode ter mais de um exemplo para um mesmo sentido de uma glosa. O segundo é uma chave estrangeira da tabela *glosas*. O terceiro e o quarto são chaves estrangeiras da tabela *palavras*. O quinto é uma frase exemplo que representa o significado da palavra em relação a uma determinada glosa. Por fim, o sexto atributo é a fonte da qual o exemplo foi retirado (a lista de fontes possíveis é apresentada na Tabela 3.2). A chave primária é composta pelos atributos *numero*, *glosa*, *palavra* e *categoria*.

Tabela 3.2: Lista das possíveis Fontes dos exemplos da Wordnet.Br e seu símbolo

Fonte	Símbolo
Córpus NILC	[
Aurélio	A]
Houaiss	H]
Michaelis	M]
Internet	I]

A tabela *synsets* possui 3 atributos: *glosa*, *palavra* e *categoria*. O primeiro atributo é uma chave estrangeira da tabela *glosas*. O segundo e o terceiro são chave estrangeira

da tabela *palavras*. Esta tabela é utilizada para relacionar cada palavra com uma ou mais glosas.

A tabela *ilis* possui 5 atributos: *ili*, *chave*, *atributo_semantico*, *synset* e *exemplo*. O primeiro atributo é um número que representa uma glosa da Wordnet.Pr. Como descobrimos que o *ili* não é um atributo único (o mesmo número pode aparecer para verbos e para substantivos) inserimos o segundo atributo que define se este ili se refere a verbos ou a substantivos (assim, a chave primária desta tabela é composta do atributo *ili* e o atributo *chave*). O terceiro é o atributo semântico desta glosa. O quarto é o *synset* de uma glosa da Wordnet.Pr. Por fim, o quinto atributo é uma frase exemplo.

A tabela *alinhamentos* possui 5 atributos: *glosa*, *ili*, *chave*, *relacao* e *sentido*. O primeiro atributo é um glosa (chave estrangeira da tabela *glosas*). O segundo e o terceiro representam um ILI da Wordnet.Pr que corresponde a uma glosa (chave estrangeira da tabela *ilis*). O quarto é a relação de alinhamento entre um ILI da Wordnet.Pr e uma glosa da Wordnet.Br (as possíveis relações de alinhamento estão na Tabela 2.1, na Seção 2.4 do Capítulo 2). Por fim, o último atributo é número do sentido na Wordnet.Pr. A chave primária desta tabela é composta pelo atributo *glosa*, *ili* e *chave*.

A tabela *hiperonimos* possui 5 atributos: *ili_esq*, *ili_dir*, *atr_esq*, *atr_dir* e *glosa*. O primeiro e o terceiro atributo representam um ili (chave estrangeira da tabela *ilis*). O segundo e quarto também representam um ili (chave estrangeira da tabela *ilis*), porém este ili é um hiperônimo do atributo ili_esq. Já o quinto atributo é utilizado somente para saber qual glosa criou esta relação de hiperonímia (este atributo foi necessário para identificar problemas com os hiperônimos). A chave primária é composta pelos quatro primeiros atributos.

A tabela *wnbrs* possui 6 atributos: *ili_esq*, *ili_dir*, *glosa_esq*, *glosa_dir atr_esq* e *atr_dir*. O primeiro e o segundo atributo são ilis da tabela *hiperonimos*. Já o terceiro e o quarto atributo são as glosas que correspondem aos ilis. Por fim, o quinto e o sexto atributo são as chaves dos ilis da tabela *hiperonimos*. A chave primária é composta por todos os atributos. Esta tabela é a que contém os hiperônimos da Wordnet.Br.

Detalhes da elaboração da base de dados

Inicialmente pensamos em criar uma tabela que armazenaria para cada glosa sua árvore de hiperônimos. Porém esta abordagem limitaria a Wordnet.Br somente a hiperônimos. Construindo uma tabela com duplas de ILIs da Wordnet.Pr em que o ILI da direita é hiperônimo do ILI da esquerda, e, posteriormente, construindo uma tabela com estes ILIs e suas glosas correspondentes, é possível também extrair os hipônimos para cada palavra. Para buscar por hipônimos, basta fazer o caminho inverso, ao invés de começar a procura a partir dos ILIs da esquerda. Começa-se a busca a partir dos ILIs da direita. Porém esta relação será deixada para trabalho futuro.

Contudo, descobrimos alguns problemas em nossa abordagem que nos fez modificar o projeto de Banco de Dados inicial. Inicialmente, pensamos que o valor do ILI da Wordnet.Pr era único, porém, descobrimos que isso não é verdade. As Figuras 3.2 e 3.3 apresentam as glosas '287' e '3809', respectivamente. Nas áreas destacadas com um retângulo está o ILI '00001740' da Wordnet.Pr que se repete em duas situações distintas. No caso da glosa '287' este ILI possui o atributo semântico *verb.body*, ou seja, este ILI se refere a um glosa de verbos. Já no outro caso, o mesmo ILI possui o atributo semântico *noun.Tops*, ou seja, este ILI se refere a uma glosa de substantivos. Não é possível o alinhamento de um verbo da Wordnet.Br e um substantivo da Wordnet.Pr, portanto o caso da glosa '3809' é um erro. Porém, a partir deste erro, pudemos identificar um problema em nossa abordagem e corrigi-lo antes que o trabalho com substantivos fosse iniciado. Assim sendo, foi necessário incluir como chave primária da tabela *ilis*, não somente o ILI, mas também, uma chave que é *verb* ou *noun*.

- **Alimentação da base de dados da Wordnet.Br**

Para a leitura, processamento e armazenamento no banco de dados dos dados nos arquivos no formato texto da Wordnet.Br, construímos o método *inserir* na classe *controlador* da ferramenta da Wordnet.Br. Este método separa todos os dados e os insere na devida tabela. Este método também constrói a tabela *hiperonimos* através dos níveis de hiperonímia que cada arquivo da Wordnet.Br contém (exemplos de arquivos da Word-

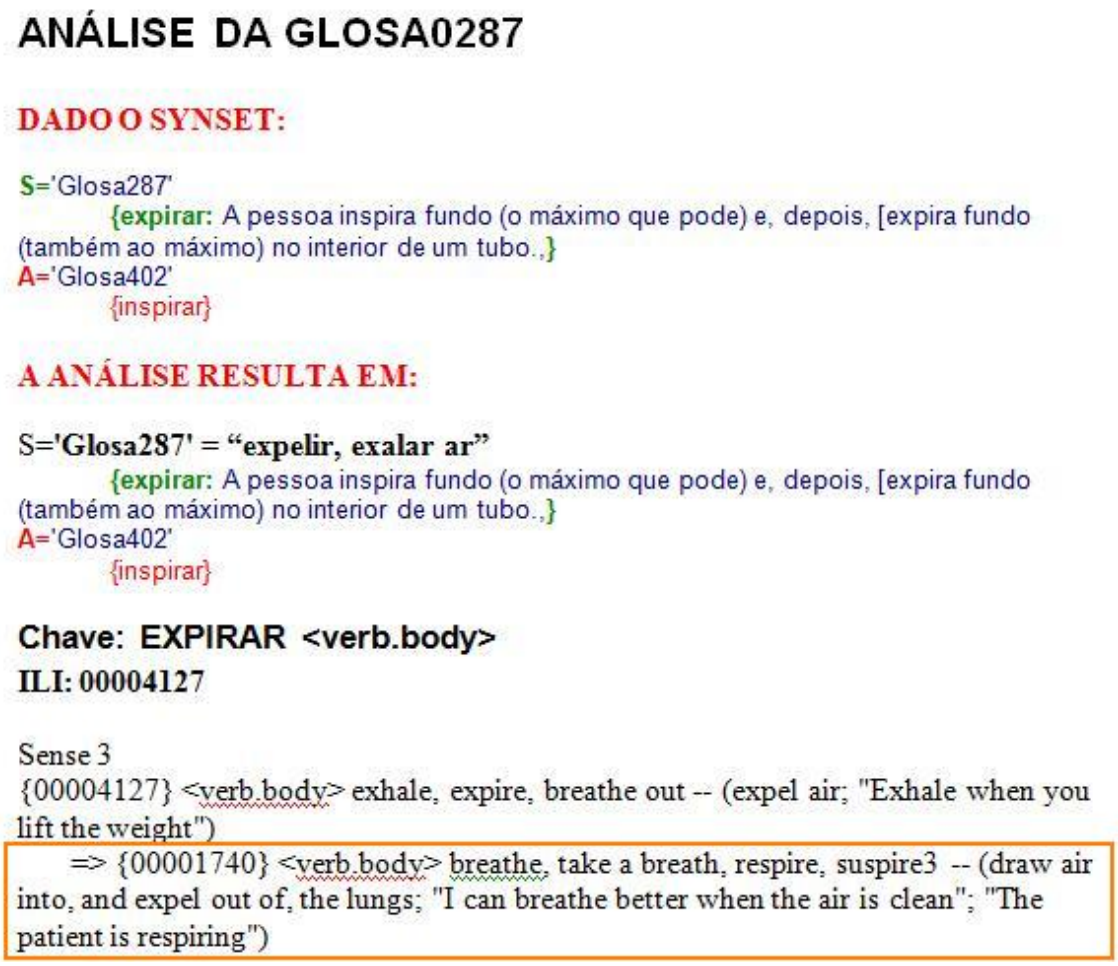

Figura 3.2: Glosa '287'

net.Br estão nas Figuras 2.6 e 2.7 na Seção 2.4 do Capítulo 2).

- **Correção das relações de alinhamento**

  Como mencionado na Seção 2.4 do Capítulo 2, os possíveis alinhamentos de uma glosa da Wordnet.Br com um ILI da Wordnet.Pr são: EQ_SYNONYM, EQ_NEAR_SYNONYM, EQ_HAS_HYPERONYM e EQ_HAS_HYPONYM (Tabela 2.1 da Seção 2.4 do Capítulo 2). Um dos casos em que ocorre a relação EQ_NEAR_SYNONYM é quando uma glosa da Wordnet.Br está relacionada a dois ou mais ILIs da Wordnet.Pr. Outro caso em que esta relação pode ocorrer é quando um ILI da Wordnet.Pr é relacionado com duas ou mais glosas da Wordnet.Br.

  Fizemos um script que buscava o primeiro caso e, para a categoria de verbos, nenhuma ocorrência foi encontrada. Porém, ainda tínhamos o segundo caso. Após a alimentação da

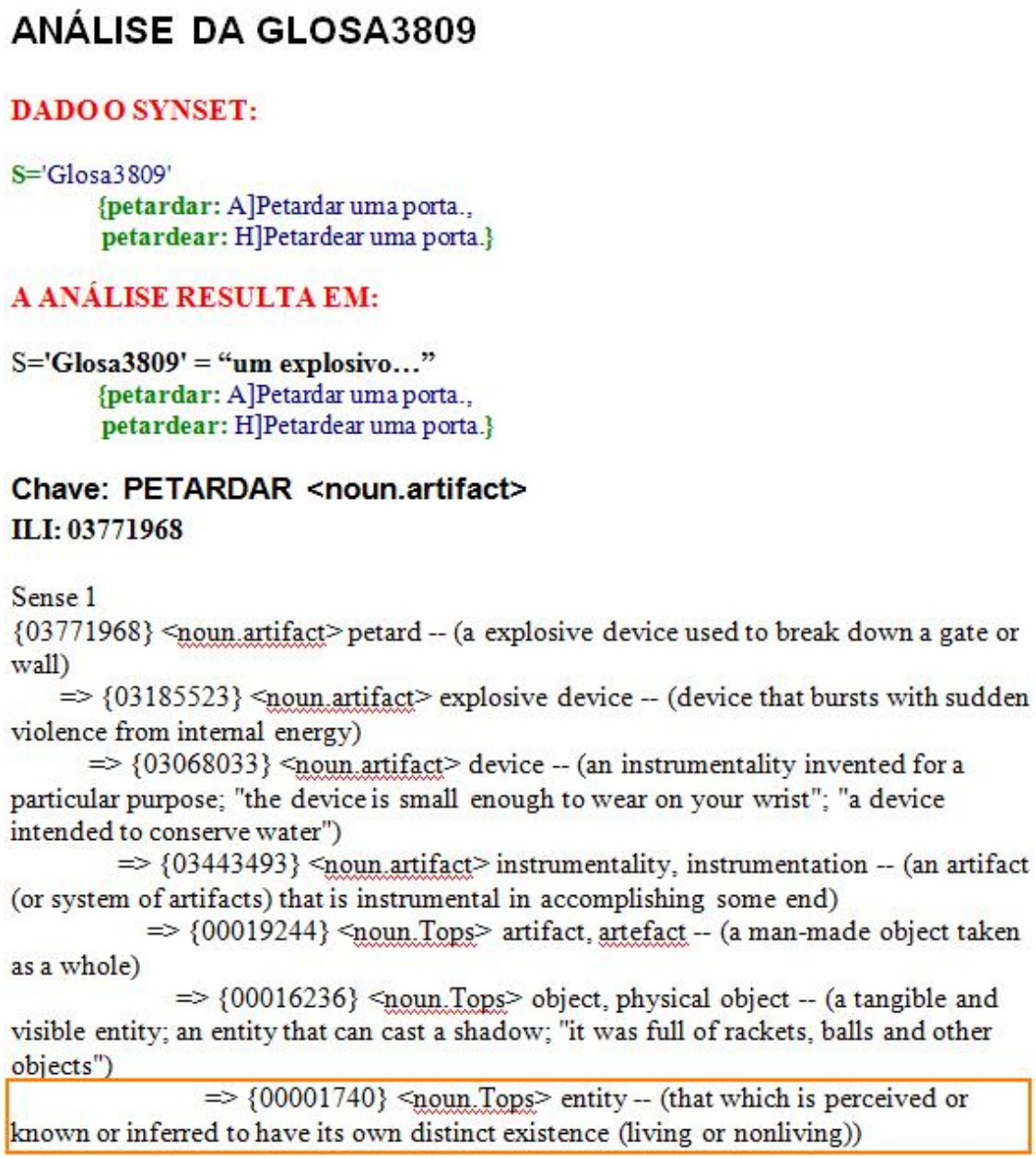

Figura 3.3: Glosa '3809'(erro de alinhamento que nos proporcionou identificar problemas em nosso modelo de Banco de Dados)

base de dados, criamos um método chamado *corrigir* na classe *controlador* da ferramenta da Wordnet.Br, que realiza uma busca da tabela *alinhamentos* e verifica se um ILI está alinhado a mais de uma glosa. Se isto é verdade, a relação de todos os alinhamentos com este ILI que estão como EQ_SYNONYM, passam a ser EQ_NEAR_SYNONYM.

- **Criação da tabela wnbrs com os hiperônimos herdados**

  A criação da tabela *wnbrs* foi feita para facilitar a consulta dos hiperônimos da Word-

net.Br. Sabemos que era possível extrair esta relação através da tabela *hiperonimos*, consultando a tabela *alinhamentos*. Porém, para evitar sobrecarga no processamento e melhorar o desempenho do sistema, decidimos construir esta tabela que já relaciona diretamente os ILIs que possuem relação de hiperonímia com suas respectivas glosas.

Para criar esta tabela, criamos um método chamado *hiperonimos* na classe *controlador* da ferramenta da Wordnet.Br. Este método associa as glosas da Wordnet.Br com os ILIs da tabela *hiperonimos* através da tabela *alinhamentos*. Com isso, construímos a tabela *wnbrs* que contém os ILIs da esquerda e da direita e as glosas da esquerda e da direita correspondentes.

- **Desenvolvimento de uma interface de busca**

Somente para validar nosso trabalho, construímos uma interface simples para buscar uma palavra e retornar as relações da Wordnet.Br[5]. A Figura 3.4[6] é a tela inicial que contém um campo de texto para que uma palavra seja digitada. Ao clicar no botão *Buscar*, chama-se o método *buscar* na classe *controlador* que recupera os dados da palavra na Wordnet.Br. Neste ponto, para auxiliar na recuperação dos hiperônimos, criamos um método recursivo chamado *encontraHiper* também na classe *controlador*. Este método é responsável por percorrer recursivamente a tabela *wnbrs* procurando os hiperônimos de um dado ILI que é associado à glosa a qual a palavra buscada pertence. Um exemplo de saída para a palavra *sonhar* é apresentado na Figura 3.5.

Como alguns alinhamentos ainda não foram realizados, alguns ILIs não possuem glosas correspondentes, então, apresentamos estes ILIs com um 'x' na frente, indicando que não há glosa alinhada para determinado ILI. Atualmente, de 3.247 ILIs, 686 possui a marca 'x', ou seja, não possuem nenhuma glosa alinhada. Futuramente, com o auxílio de um editor, será possível que o linguista alinhe estes ILIs com as glosas pertinentes, se for o caso.

[5]Para acessar a Wordnet.Br: http://caravelas.icmc.usp.br/wordnetbr

[6]Como o projeto de IC da aluna em questão está no escopo do projeto PorSimples (http://caravelas.icmc.usp.br/), utilizamos o *layout* deste projeto como *layout* do protótipo da Wordnet.Br

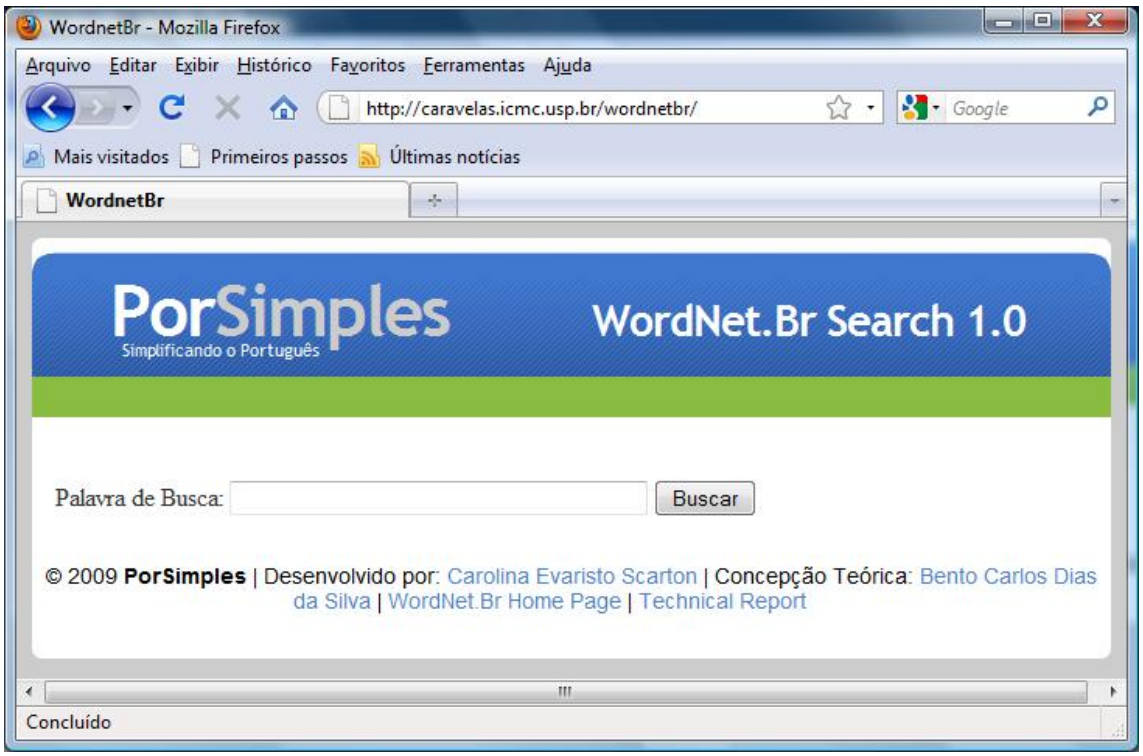

Figura 3.4: Tela de busca da Wordnet.Br (protótipo)

3.4 Resultados Obtidos

Todas as atividades que foram propostas para serem realizadas até o momento da redação deste relatório técnico foram executadas. A única consideração que deve ser feita é que, na proposta inicial, pretendíamos utilizar o editor da Wordnet.Br para a herança das relações de hiperonímia. Porém, como já discutimos na Seção 3.2, tomamos a decisão de abandonar o editor e adotar outra abordagem.

Estavam previstas leituras sobre a Wordnet.Br e a WordNet.Pr. Esta tarefa estava prevista para terminar em agosto, porém se estendeu durante todo o trabalho. Outra tarefa prevista era a realização de reuniões com os criadores da Wordnet.Br. Em agosto, realizamos uma reunião com o desenvolvedor do editor da Wordnet.Br, o senhor Ricardo Hasegawa, e uma reunião com o criador da Wordnet.Br, o professor doutor Bento Carlos Dias da Silva.

Também estava previsto o estudo do editor da Wordnet.Br. Esta tarefa compreendia os meses de agosto e setembro e foi realizada no prazo estipulado. A conversão dos verbos alinha-

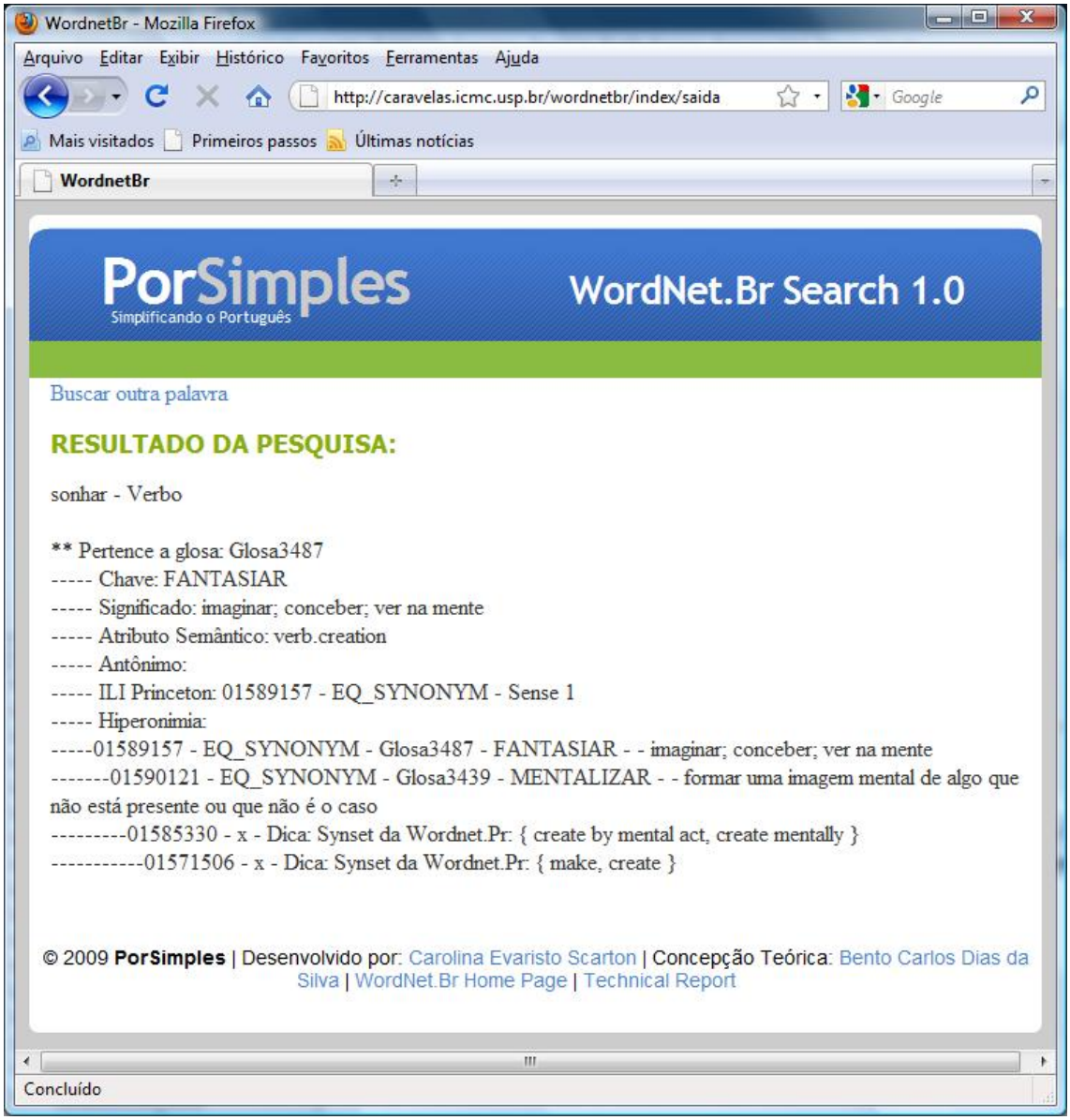

Figura 3.5: Saída da Wordnet.Br para o verbo *sonhar*

dos para a plataforma da Wordnet.Br era outra tarefa. Porém, como decidimos não utilizar o editor, a plataforma da Wordnet.Br passou a ser a base de dados relacional e a conversão dos verbos alinhados passou a ser a alimentação da base da Wordnet.Br. Esta tarefa estava prevista para o mês de setembro. Porém, dado a mudança de planos para o desenvolvimento do trabalho, esta tarefa se estendeu até outubro.

A correção dos erros do alinhamento também era uma tarefa deste trabalho. Como mencionamos na Seção 3.3, realizamos a correção de arquivos com erros de digitação ou com partes faltantes da Wordnet.Br. Também realizamos a correção da relação de EQ_NEAR_SYNONYM.

Esta atividade compreendia os meses de setembro e outubro e foi realizada no prazo.

Para terminar, o desenvolvimento de um sistema que herda as relações de hiperonímia da Wordnet.Pr para a Wordnet.Br era a última tarefa prevista para este projeto. Também conseguimos construir essa ferramenta e herdar as relações de hiperonímia como descrevemos na Seção 3.3. Esta ferramenta e as relações de hiperonímia eram o grande objetivo deste trabalho e, com este resultado, é possível dar continuidade a este trabalho seja criando um editor ou alimentando a base com outras categorias (substantivos, adjetivos ou advérbios).

E, como resultado adicional, temos a criação de uma base de dados relacional para a Wordnet.Br com 7.658 verbos e cerca de 3.713 glosas.

3.5 Dificuldades e Limitações

A principal dificuldade encontrada neste projeto foi o próprio editor da Wordnet.Br. Como inicialmente tínhamos a intenção de utilizá-lo para a realização do projeto, quando decidimos abandoná-lo tivemos que elaborar uma nova ferramenta e uma nova base de dados.

Outra dificuldade foi a inconsistência dos arquivos .doc (documento do Microsoft Word). Esta inconsistência se deu não somente nos erros identificáveis e não identificáveis nos arquivos, mas também na ausência de muitos arquivos. Como o trabalho da criação da Wordnet.Br foi feito por vários alunos, os arquivos não estavam todos juntos e o pesquisador responsável pela Wordnet.Br precisava procurar arquivo por arquivo. Para auxiliar nesta tarefa, desenvolvemos um script capaz de identificar as glosas faltantes.

A falta de um bom conversor de .doc para .txt de livre acesso também se tornou uma dificuldade. As únicas ferramentas encontradas eram Shareware e possuíam uma versão *trial* que funcionava por 30 dias. Porém, a primeira ferramenta utilizada não realizava a conversão corretamente o que interferia diretamente na identificação dos dados. Já a segunda ferramenta, Convert Doc [7], realizou uma boa conversão, porém só permitia que convertidos 250 textos por vez. Como temos 4.128 textos, o trabalho se tornou demorado.

[7] Para acessar a ferramenta Convert Doc: http://www.softinterface.com/Convert-Doc/Convert-Doc.htm

Como limitações para este projeto temos os erros que não pudemos identificar nos documentos da Wordnet.Br e que foram salvos no banco de dados e a ausência de artigos científicos relatando como estão estruturadas as Wordnet.Pr, MultiWordNet ou EuroWordnet (assim não pudemos nos basear em trabalhos anteriores para a criação da base de dados da Wordnet.Br).

3.6 Considerações Finais

Neste capítulo apresentamos todos os detalhes da execução deste projeto, detalhando as tarefas realizadas, apresentando os resultados obtidos (relacionando-o com o proposto) e, por fim, discutindo sobre as dificuldades e limitações encontradas.

Pode-se concluir que este trabalho obteve o êxito esperado e, mais do que isso, abriu novos caminhos para a Wordnet.Br que agora está apta a ser acessada via Web.

4 Conclusão

Neste capítulo são feitas as conclusões sobre o trabalho, além de uma análise crítica sobre o curso. Na Seção 4.1 estão as considerações finais e as contribuições do projeto. Por fim, na Seção 4.2 são descritos os trabalhos futuros previstos para este projeto.

4.1 Contribuições

Com a conclusão deste trabalho, retomamos o trabalho de consolidação da Wordnet.Br, que é uma lacuna muito antiga na área de PLN. Como criamos uma base de dados relacional, é possível criar ferramentas que acessem esses dados via Web e, portanto, disponibilizar este recurso a toda a comunidade acadêmica.

Mesmo sendo um trabalho inicial, já é possível utilizar os dados da Wordnet.Br em trabalhos científicos. O primeiro trabalho que se beneficiará deste recurso será o projeto de IC da aluna em questão que, como mencionado na Seção 1.1 do Capítulo 1, utilizará as relações de hiperonímia para calcular uma métrica de inteligibilidade textual. Outros trabalhos que necessitem de recursos semânticos também poderão se beneficiar deste trabalho.

Além disso, este trabalho abriu caminhos para os outros trabalhos relacionados com Wordnet.Br, que poderão partir deste modelo e utilizar a mesma base de dados, pois, na construção da base de dados, prezamos por mantê-la o mais portável possível.

4.2 Trabalhos Futuros

Um trabalho futuro, talvez o principal, é a disponibilização do recurso criado neste projeto via Web. Uma ferramenta de consulta nos moldes da MultiWordNet [1] é o mínimo que pode ser feito. A disponibilização da base de dados também é possível, porém, é necessário verificar sua disponibilidade com o seu criador.

Outro trabalho futuro, não menos importante, é a criação de um editor capaz de identificar alinhamentos que ainda não existem na base da Wordnet.Br e auxiliar o linguista a relacionar uma glosa com um ILI. Além disso, este editor deve ser capaz de permitir ao linguista criar novas glosas e alterar as antigas.

Por fim, a relação de hiponímia é um trabalho futuro que está mais próximo da realidade. Como mencionamos na Seção 3.3 do Capítulo 3, esta relação está praticamente pronta, pois bastaria seguir o caminho inverso na tabela *wnbrs* da base de dados da Wordnet.Br. Porém, a tarefa de validação desta afirmação fica para o futuro.

[1] http://multiwordnet.itc.it/english/home.php

Referências Bibliográficas

[Balkova et al., 2004] Balkova, V., Sukhonogov, A., and Yablonsky, S. (2004). Russian wordnet. Citado nas páginas 16, 23, e 30.

[Barbu and Mititelu, 2007] Barbu, E. and Mititelu, V. B. (2007). Automatic building of wordnets. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, volume 292 of *Current Issues in Linguistic Theory*, pages 217–226. John Benjamins, Amsterdam & Philadelphia. Citado nas páginas 16, 23, e 30.

[Bentivogli et al., 2002] Bentivogli, L., Pianta, E., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India. Citado nas páginas 7, 15, 22, e 30.

[Bentivogli et al., 2000] Bentivogli, L., Pianta, E., and Pianesi, F. (2000). Coping with lexical gaps when building aligned multilingual wordnets. Citado na página 30.

[Botero and Ricarte, 2009] Botero, S. and Ricarte, I. L. M. (2009). Extração de relações semânticas via análise de correlação de termos em documentos. *STIL 2009*. Citado na página 30.

[Cruse, 1986] Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK. Citado nas páginas , 12, 14, e 30.

[Di Felipo and Dias-da Silva, 2007] Di Felipo, A. and Dias-da Silva, B. C. (2007). Towards an automatic strategy for acquiring the wordnert.br hierarchical relations. *In the Proceedings of the 5th Workshop in Information and Human Language Technology (TIL'2007)*, pages 1717–1720. Citado nas páginas III, 8, e 29.

[Dias-da Silva, 2003] Dias-da Silva, B. C. (2003). Human language technology research and the development of the brazilian portuguese wordnet. *17th International Congress of Linguists*. Citado nas páginas , 7, 8, 20, e 29.

[Dias-da Silva, 2005] Dias-da Silva, B. C. (2005). A construção da base da wordnet.br: conquistas e desafios. *In the Proceedings of the 3rd Workshop in Information and Human Language Technology (TIL'2005)*, pages 2238–2247. Citado nas páginas III, 7, 17, 18, e 29.

[Dias-da Silva et al., 2008] Dias-da Silva, B. C., Di Felippo, A., and Nunes, M. G. V. (2008). The automatic mapping of princeton wordnet lexical-conceptual relations onto the brazilian portuguese wordnet database. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/. Citado nas páginas 7, 8, 17, 21, 26, e 29.

[Dias-da Silva et al., 2002] Dias-da Silva, B. C., Oliveira, M. F. d., and Moraes, H. R. d. (2002). Groundwork for the development of the brazilian portuguese wordnet. In *PorTAL '02: Proceedings of the Third International Conference on Advances in Natural Language Processing*, pages 189–196, London, UK. Springer-Verlag. Citado nas páginas , 7, e 29.

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition. Citado nas páginas 7, 17, e 30.

[Klebanov et al., 2004] Klebanov, B. B., Knight, K., and Marcu, D. (2004). Text simplification. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*, pages 735–747. Springer Verlag. Citado na página 6.

[Leffa, 1996] Leffa, V. J. (1996). Fatores da compreensão na leitura. *Cadernos no IL*, 15:143–159. Citado na página 5.

[Max, 2006] Max, A. (2006). Writing for language-impaired readers. In Gelbukh, A. F., editor, *CICLing*, volume 3878 of *Lecture Notes in Computer Science*, pages 567–570. Springer. Citado na página 6.

[McNamara et al., 2002] McNamara, D. S., Louwerse, M. M., and Graesser, A. C. (2002). Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. *Grant proposal.* Citado na página 5.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: an on-line lexical database*. *Int J Lexicography*, 3(4):235–244. Citado nas páginas 6, 13, 22, e 30.

[Scarton et al., 2009] Scarton, C. E., Almeida, D. M., and Aluísio, S. M. (2009). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *The 7th Brazilian Symposium in Information and Human Language Technology*. Citado nas páginas 5 e 8.

[Siddharthan, 2002] Siddharthan, A. (2002). An architecture for a text simplification system. In *In LEC02: Proceedings of the Language Engineering Conference (LEC02)*, pages 64–71. Citado na página 6.

[Snow et al., 2005] Snow, R. L., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA. Citado na página 11.

[Snow and Patel, 2008] Snow, R. L. and Patel, K. D. (2008). Cs 224n class project automatic hypernym classification. Citado na página 11.

[Vossen, 2002] Vossen, P. (2002). Eurowordnet general document. Technical report, EuroWordNet Project. Citado nas páginas III, 7, 12, 15, 22, e 30.

[Vossen, 2004] Vossen, P. (2004). Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Linguistics, Vol17*. 4 cites: http://scholar.google.com/scholar?num=100&hl=en&lr=&cites=3088026344681176314. Citado nas páginas 15 e 30.