PLANO DETALHADO DAS ATIVIDADES DO BOLSISTA[i]

BOLSISTA

Rafael Luis Beraldo – CPF 38554416856 (jul.2011-ago.2012)

OBJETIVOS

O objetivo das atividades propostas para este bolsista, dando continuidade às atividades da previstas no ano anterior, além de ser o estudo de tópicos de semântica lexical, pura e computacional (DIAS-DA-SILVA, 2006), selecionados da Bibliografia Básica [em particular, de CRUSE (1986, 2004), FELLBAUM (1999), MARRAFA (2001), SCOTT (2001), VOSSEN (1998) e EUROWORDNET (2007)] e a aquisição de experiência para desenvolver pesquisa empírica e para elaborar relatórios no nível de IC, é a co-indexação léxico-semântica (isto é, especificação dos synsets das duas bases que são semanticamente equivalentes) entre **500 synsets (identificados de 01000 a 01499)** de substantivos do domínio ARTIFACT ('artefato')[1] da base da WordNet de Princeton (FELLBAUM, 1998; WORDNET, 2007) e os synsets de substantivos conceitualmente equivalentes da base da WordNet.Br (DIAS-DA-SILVA, 2004; DIAS-DA-SILVA et al., 2002, 2006). Por exemplo, estes três synsets de substantivos do inglês que denotam artefatos `{cup, loving cup}`, `{cup1}` e `{cup2}` são, respectivamente, co-indexados a estes três synsets de substantivos do português `{taça, troféu}`, `{xícara}`, `{buraco, caçapa}`. Esse procedimento de co-indexação, que demanda investigação analítica meticulosa, envolve a análise léxico-conceitual dos synsets de substantivos do português e do inglês e inclui, para cada synset do português que for objeto do referido processo de co-indexação, (i) a sua revisão e atualização, (ii) a seleção, no *corpus* de referência do projeto, de uma frase para exemplificar o contexto de uso de cada unidade lexical que o constitui, (iii) a especificação da glosa em português (isto é, a especificação de uma definição intuitiva e informal) para explicitar o conceito lexicalmente por ela representado, e (iv) a identificação dos casos em que não for possível o estabelecimento da co-indexação (devido a lacunas lexicais, por exemplo), para, em uma etapa posterior que não é parte das atividades do bolsista, proceder-se (v) à especificação automática, mas com a supervisão de um linguista, da relação de *antonímia* e das relações hierárquicas de *hiponímia*/*hiperonímia* e de *meronímia*/*holonímia* entre os synsets de substantivos do português que resultaram da análise aqui proposta.

METODOLOGIA

Com os fundamentos fornecidos no projeto do orientador e na Bibliografia Básica, com o auxílio de programas concordanciadores, como o proposto por Scott (2001), e do editor especificamente construído para a montagem da base da WordNet.Br, o trabalho de análise dos synsets consiste na revisão de cada synset do português (isto é, na verificação de que todas as unidades que o compõem lexicalizam o mesmo conceito), na seleção de frases-exemplo extraídas de *corpus*, na proposição das glosas para os synsets do português e na co-indexação entre estes e os synsets da base do inglês. Especificamente, nessa tarefa, os passos, descritos a seguir, devem ser sistematicamente seguidos para cada par de synsets, um de cada língua, que potencialmente podem ser co-indexados. Antes de seguir os 5 passos da análise, iniciam-se, no computador, estes aplicativos (a) o arquivo NOUN.ARTIFACT.doc; (b) o arquivo modelo para posterior indexação (.txt); (c) a WordNet 2.0; (d) o Editor da WordNet.Br; (e) os Dicionários inglês/inglês e inglês/português; (f) o navegador de internet com acesso aos *corpora* e (g) os motores de busca (*Google*, *AltaVista*, *Yahoo*). **O primeiro passo** consiste na seleção do synset do arquivo "NOUN.ARTIFACT.doc" da base de substantivos da WordNet de Princeton. **O segundo passo** consiste na seleção do synset do português que é semanticamente equivalente ao synset selecionado no passo anterior (se não houver um synset equivalente, escolhe-se um novo synset no passo anterior). Para isso, seleciona-se, na base de substantivos da WordNet.Br, o synset que deve ser co-indexado. Para decidir a equivalência semântica entre os synsets, procede-se à delimitação dos conceitos representados lexicalmente no synset do português, analisando-se as frases-exemplo selecionadas no *corpus* de referência do projeto assim formado: 1. *O Corpus do NILC* (Núcleo Interinstitucional de Lingüística Computacional), disponível para consulta on-line; 2. Os textos em português do Brasil disponíveis na Internet que são recuperados pelo motor de busca Google; 3. O conjunto das abonações registradas nos dicionários eletrônicos Michaelis (WEISZFLOG, 1998), Aurélio (FERREIRA, 1999) e Houaiss (HOUAISS e VILLAR, 2001). O resultado dessa análise é registrado em instruções que sugerem a redução, a ampliação, a partição ou a eliminação do synset da base da WordNet.Br. **O terceiro passo** consiste na seleção de uma CHAVE para representá-lo, isto é, na seleção da unidade lexical do próprio synset que se considera mais representativa do synset. **O quarto passo** consiste na especificação da correspondência entre esse synset do português e o synset do inglês. Essa especificação é feita com o auxílio da consulta aos dicionários monolíngues do português e do inglês e aos dicionários bilíngues inglês-português e português-inglês e aos textos em inglês disponíveis na Internet. Esses recursos lexicográficos e textuais, além de auxiliarem a delimitação do significado dos itens lexicais do português, auxiliam também na delimitação dos itens lexicais do inglês, sobretudo aqueles que não constam dos dicionários convencionais. Nesse passo, especificam-se o tipo de alinhamento (por EQ_SYNONYM, por EQ_NEAR_SYNONYM, por EQ_HAS_HYPONYM ou por EQ_HAS_HYPERONYM) para o synset correspondente do português e o tipo semântico ARTIFACT (ou outro tipo relevante), que é uma informação "herdada" do synset do inglês. Esse processo permite a automatização da

1 Cf. Distribuição geral dos synsets no cronograma do projeto do orientador.

classificação semântica dos substantivos da base do português. Por fim, **o quinto passo** consiste na tradução da glosa do inglês para o português e no registro das informações resultantes da análise em um arquivo assim nomeado:

<nº_na_Base_WNBr>.<ILI>.<tipo_semântico>.<chave>.void|md|cr.HYPER|HYPO\NEAR|void>.txt

As informações desse arquivo são posteriormente inseridas na base da WordNet.Br por meio de um editor. Ressalta-se que a realização de todos esses passos conta com a supervisão semanal do orientador.

META

Com esses objetivos e essa metodologia, o aluno deverá atingir, ao final dos **doze meses** de trabalho, a meta descrita no Quadro 1.

Ao final de:	**Realizar as atividades:**	**Com a previsão dos seguintes resultados:**
12 meses	Além dos estudos teórico-metodológicos, as atividades incluem, sobretudo, a especificação das frases-exemplos para cada um dos itens lexicais constitutivos de cada synset do português que puder ser indexado a cada um dos **500 synsets** do inglês (ARTIFACT), a análise, a revisão e a glosagem dos synsets de substantivos do português e o seu alinhamento ao respectivo synset do inglês (a indexação).	Domínio das técnicas de análise léxico-semântica, inserção dos synsets indexados na base da WordNet.Br, estabelecimento da correspondência entre estes e os seus equivalentes semânticos na WordNet de Princeton. (**01000 a 01499**)

Quadro 1. Meta para **doze meses** de trabalho.

Destaca-se que as atividades do Quadro 1 contribuem para a realização de mais uma parcela das etapas 1, 2 e 5 do Cronograma do projeto do orientador: "*O Desenvolvimento da Base de Substantivos da WordNet.Br e a sua Co-indexação com a WordNet de Princeton"*.

Nesse período, as Atividades detalhadas a seguir são ordenadas segundo o cronograma do Quadro 2.

ATIVIDADES

a. Estudo dos tópicos de semântica lexical selecionados da Bibliografia Básica voltados para a descrição linguístico e computacional dos sentidos dos substantivos que denotam artefatos;

b. Análise ortográfica (correção das unidades léxicas mal-grafadas) e léxico-semântica (confirmação de que todas elas lexicalizam um mesmo conceito) dos synsets do português extraídos da base da WordNet.Br (em construção), a partir da projeção da análise de mesma natureza de cada um dos 500 synsets de substantivos selecionados no arquivo de base "NOUN.ARTIFACT.doc" do inglês, extraídos da base da WordNet de Princeton; atividade que se realiza pautada na semântica lexical (pura e computacional), na análise de concordâncias e na consulta aos dicionários *Webster's* eletrônico, *RHUD*, *Michaelis*, *Aurélio* e *Houaiss*; importante: caso não haja um synset equivalente na base do português, constrói-se um novo synset que se alinhe conceitualmente ao synset do inglês, ou seja, constrói-se o synset português cujas unidades lexicais sejam traduções das unidades lexicais do synset do inglês;

c. Especificação, para cada synset analisado em (c), de uma glosa, tomando por base a glosa especificada para o synset selecionado na atividade (b), e uma CHAVE, que é a unidade lexical mais representativa do synset (em termos da maior frequência de ocorrência em *corpus*);

d. Registro do resultado da análise em arquivo;

e. Participação de encontros semanais de orientação;

f. Elaboração de relatório científico e de trabalhos para apresentação em eventos de relativos à IC.

CRONOGRAMA

Atividades	**2011**			**2012**		
	Agosto-Setembro	**Outubro-Novembro**	**Dezembro-Janeiro**	**Março-Abril**	**Maio-Junho**	**Julho**
a.						
b.						
c.						
d.						

e.						
f.						

Quadro 2. Cronograma de atividades: de 01 de agosto de 2011 a 31 de julho de 2012.

BIBLIOGRAFIA BÁSICA

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

CRUSE, D. A. *Meaning in language*. Oxford: Oxford University Press, 2004.

DIAS-DA-SILVA, B.C Montagem da Base da Wordnet para o Português do Brasil. *Relatório Técnico da Chamada CNPq 09/2001 – Conteúdos Digitais/Edital SocInfo/ProTeM 01/2001, 01/03/2002 a 31/05/2004.* Araraquara: CELiC, FCL, UNESP, 2004. p.50

DIAS-DA-SILVA, B. C. O estudo linguístico-computacional da linguagem. In: *Letras de Hoje*, v.41, p.103-138, 2006.

DIAS-DA-SILVA, B. C., FELIPPO, A. Di, HASEGAWA, R. Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations In: J. G. CARBONELL; J. SIEKMANN (Eds.) *Lecture Notes on Artificial Intelligence*. Berlin / Heidelberg: Springer, 2006. p.120-130.

EUROWORDNET. *Rede EuroWordNet*. Disponível em: <http://www.illc.uva.nl/EuroWordNet/data/sampleData.html>. Acesso em: 23 mar. 2007.

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

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

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

HOUAISS, A., CARDIM, I. *Webster's Dicionário Inglês-Português.* Rio de Janeiro: Record, 2005.

HOUAISS, A., VILLAR, M.S. *Dicionário Hoauiss da língua portuguesa*. Rio de Janeiro: Objetiva, 2001.

MARRAFA, P. *WordNet do Português – uma base de dados de conhecimento lingüístico*. Lisboa: Instituto Camões, 2001.

SCOTT, M. *WordSmith Tools version 3.0*. Oxford: Oxford University Press, 2001.

TAYLOR, J. L. *Webster's Portuguese-English Dictionary*. Rio de Janeiro: Record, 2003.

VOSSEN, P. Special issue on EuroWordNet. In: *Computers and the Humanities*, v.32, 2-3, p.73-251, 1998, Dordrecht/Boston/London.

WEISZFLOG, W. (Ed.) Mich*aelis português – moderno dicionário da língua portuguesa* - versão 1.1). São Paulo: DTS Software Brasil Ltda, 1998.

WORDNET. *Rede WordNet de Princeton on-line*. Disponível em: <http://www.cogsci.princeton.edu/cgi-bin/webwn>. Acesso em: 23 mar. 2007.

i Cumpre esclarecer que os dois **Planos de Atividades** propostos para cada um dos dois bolsistas no âmbito do edital **PIBIC-CNPq/2011-2012** são de mesma natureza e seguem a mesma metodologia. O que os distingue e justifica a solicitação de renovação das duas bolsas para o mesmo projeto de pesquisa do orientador são os dados diversos que são submetidos para análise em cada um dos planos. Embora os dois bolsistas analisem synsets do mesmo domínio semântico [ARTIFACT ('artefato')], **este bolsista deverá analisar e alinhar 500 synsets de substantivos, identificados com os números de 01000 a 01499**; o outro bolsista procederá a análise e o alinhamento de **500 synsets de substantivos, identificados com os números de 01500 a 01999**.